

Genome Sequence Analysis of *Bos frontalis* and Genome Wide Association Study (GWAS) between near species

by

Sara Islam

17301069

Rumana Parvin

17301027

Most. Sanjida Afrin

17201038

Liakot Khan Babu

17101324

Oindri Aurunima Sarkar

17101440

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2021

© 2021. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

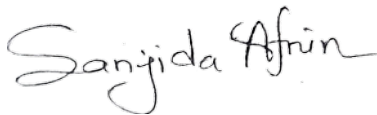
Student's Full Name & Signature:



Sara Islam
17301069



Rumana Parvin
17301027



Most. Sanjida Afrin
17201038



Liakot Khan Babu
17101324

Oindri Aurunima Sarkar

Oindri Aurunima Sarkar
17101440

Approval

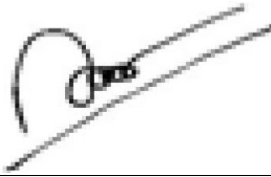
The thesis titled “Genome Sequence analysis of *Bos frontalis* and Genome Wide Association Study (GWAS) between near species” submitted by

1. Sara Islam (17301069)
2. Rumana Parvin (17301027)
3. Most. Sanjida Afrin (17201038)
4. Liakot Khan Babu (17101324)
5. Oindri Aurunima Sarkar (17101440)

Of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 2, 2021.

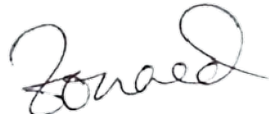
Examining Committee:

Supervisor:
(Member)



Md. Iqbal Hossain
Assistant Professor
Department of Computer Sciences and Engineering
Brac University

Co-Supervisor:
(Member)




AMAM Zonaed Siddiki
Professor
Department of Pathology and Parasitology
Chittagong Veterinary and Animal Sciences University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam
Associate Professor
Department of Computer Sciences and Engineering
Brac University

Head of Department:
(Chair)



Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The Gayal is an enormous estimated imperiled bovine species having a place with the Bovidae family, Bovini tribe, Bovina group, Bos family, and *Bos frontalis* species. It is a semi-domestic animal, kept in gated forest regions rather than in or near villages. It is also known as the mithan or mithun. With $2n=58$ XX chromosome complements and bred in India, Myanmar, China, Bangladesh, and Bhutan's hill forests at an elevation of 1000 to 3000 meters above sea level. The most common animal utilized by Eastern Himalayan groups is the mithun or Gayal (*Bos frontalis*). In Bangladesh, the home plot of Gayal has been found in the Bandarban Slope area. In Bangladesh, the complete number of Gayal was assessed to be between 800-900, with a normal of 5.01 Gayal per rancher. Be that as it may, the beginning of this species is as yet questioned and data on its genomic design is inadequate up until now. Gayal's meat is quite possibly the most favored wellsprings of creature protein among nearby clans and its milk is nutritious. Mithuns are mostly introduced outside of North East India with the intention of cross-breeding with other bovids. By breeding Gayal with different types of cattle we can have hybrids which will be able to produce good quality meat as well as milk. Bangladesh had high reproductive efficiency, with a calving interval of 402 days. As a result, Gayal's reproductive efficiency and rapid growth in the home tract suggest that *Bos frontalis* can be commercially exploited provided a solid breeding plan and competent extension service are implemented. Gayal's genome resequencing will reveal important details on genomic organization, evolutionary divergence, conservation, and total endemic diversity. We will use Genome-wide association studies on *Bos frontalis* and the purpose of our work is identifying some genes-proteins with economic importance and compare them with related species.

Keywords: Bos, GWAS, gene, protein, genotype, phenotype, muscle, growth, leucine.

Dedication

This thesis is dedicated to our parents, who have worked tirelessly to make our lives lovely.

Acknowledgement

Foremost, all praises and thanks to the Almighty Allah for giving us strength and willpower and for whom our thesis has been completed without any major hardships.

Secondly, it's our unfeigned pleasure to convey our humblest and deepest gratitude to our thesis supervisor Dr. Muhammad Iqbal Hossain, and co-supervisor Dr. AMAM Zonaed Siddiki for their incomparable guidance and connivance throughout our entire research. Their kind cooperation along with their vision, knowledge, and passion for their field of work has deeply inspired us throughout the thesis. We feel privileged and honored to have had them as our supervisors, whose continuous motivation and counseling encouraged us to push our limits to think more critically and approach things dynamically so that we could learn more during the whole thesis. We would also like to express our special thanks to our instructor Mr. Anamul Bahar, who has been a constant inspiration during our entire thesis preparation and whose enormous knowledge has motivated us to give our best effort into this research. We are truly grateful for his empathy, enthusiasm, and friendly guidance.

Furthermore, we would like to thank the Department of Computer Science and Engineering for providing a fantastic working environment and remarkable resources that have helped us greatly.

Last but not the least, we are extremely grateful to our parents for their trust in us. Without their invaluable blessings and care it would not have been possible for us to come this far. With their blessings bestowed upon us, we were able to finish the thesis and are now fixing to graduate.

Table of Contents

Declaration	i
Approval	iii
Abstract	v
Dedication	vi
Acknowledgment	vii
Table of Contents	viii
List of Figures	x
List of Tables	xi
Nomenclature	xii
1 Introduction	1
1.1 Motivation	2
1.2 Aims and Objectives	3
1.3 Thesis Organization	3
2 Review of literature	4
2.1 Background	6
2.1.1 Cell	6
2.1.2 DNA	8
2.1.3 Genome size	9
2.1.4 DNA sequence	10
2.1.5 Next generation sequence	14
2.1.6 Fastq reads	14
2.1.7 Fasta reads	14
3 Materials and Methods	15
3.1 Sampling and DNA isolation:	15
3.2 Library preparation:	15
3.2.1 Data availability	15
3.3 Genome assembly:	17
3.3.1 WGS data processing and filtering:	17
3.3.2 Scaffold construction:	17

3.3.3	Assembly visualization:	17
3.3.4	Validation of assembly and Gap resolution:	18
3.3.5	Final evaluation of the assembly:	18
3.4	Gene prediction and functional annotation:	18
3.4.1	Annotation of protein-coding genes:	18
3.5	Genome Wide Association Study (GWAS) to compare specific trait: .	19
3.5.1	Data collection	19
3.5.2	Finding most expressed proteins	19
3.5.3	Extracting the most expressed protein	19
3.5.4	Multiple Sequence Alignment	20
3.5.5	Phylogenetic tree construction	20
3.5.6	Visualizing protein 3D model	20
3.6	Total project pipeline	20
4	Results and Discussion	22
4.1	Data quality	22
4.2	Draft Genome	22
4.3	Genome Annotation:	22
4.4	Genome Wide Association Study:	28
5	Conclusion & Future Work	31
	References	35

List of Figures

2.1	Cell throughout the body.	7
2.2	Classification of cells.	8
2.3	DNA throughout the body.	9
2.4	Molecular Structure of DNA.	10
2.5	Sample collection and processing.	11
2.6	PCR Process.	11
2.7	DNA Sequencing.	12
2.8	Gene and Protein on DNA.	13
3.1	Our dataset on NCBI.	16
3.2	Pipeline of this study.	21
4.1	Per base quality.	23
4.2	Per base sequence quality.	23
4.3	Per sequence GC content.	24
4.4	Per sequence quality.	24
4.5	Per tile quality.	25
4.6	Sequence length distribution.	25
4.7	Per base n content.	26
4.8	Duplication levels.	26
4.9	Adapter content.	27
4.10	Phylogenetic tree of Leucine related protein of <i>Bos frontalis</i> and other Bovine species.	29
4.11	LRRN1_BOVIN on A and its sub-section in B to I.	30

List of Tables

4.1	Overview of data files/data sets	27
4.2	Genome assembly parameters	28
4.3	Matched protein list	28

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

BLAST : Basic Local Alignment Search Tool

bp : Base pair

BUSCO : Benchmarking Universal Single-Copy Orthologs

BWA : Burrows-Wheeler Aligner

BWT : Burrows-Wheeler transform

CVASU : Chattogram Veterinary and Animal Sciences University

DNA : Deoxyribonucleic acid

EDTA : Ethylene_diamine-tetraacetic acid

Gbp : Giga base pair

GO : Gene ontology

IMC : Indian Major Carp

Kb : Kilo base pair

Mb : Mega base pair

mRNA : Messenger RNA

NGS : Next Generation Sequencing

PCR : Polymerase Chain Reaction

qPCR : Quantitative Polymerase Chain Reaction

SRA : Sequence Read Archive

WGS : Whole Genome Sequencing

Chapter 1

Introduction

The major advantage of next-generation sequencing techniques is, it has revolutionized the field of life sciences [8]. Next-generation sequencing generates a significant amount of biological data, which has cleared the road for personalized medicine. As next-generation sequencing (NGS) technology is very fast, large-scale sequencing reads with the exquisite resolution, accuracy, and drastically reduced costs are available, so it has appeared to be very fashionable [9]. For calling variations from the data created by NGS, different types of aligners and variant calling tools and software have been created. For making the tasks easier these tools and software have been combined in various workflow. Generally, an orthodox workflow typically includes an aligner followed by a variant caller where the aligner aligns the sequencing reads of the given reference genome, and the variant caller identifies variant sites then assigns a genotype of the given species's data. These techniques help in many applications such as SNPs detection[11], genomic structural variants[12] [16] in a population, analysis of mRNA expression [10] , personalized medicine [13] cancer genomics [15] and, DNA methylation studies [6] , metagenomics [20], There are some tools and software (e.g. metagenomics) that need de novo sequencing for the given data [18] , while resequencing is needed by other tools and software (e.g., variant detection, cancer genomics). Combinations of these tools and software produced a massive volume of data and this sequencing runs that create a lot of computational difficulties [17].

In recent years, the fast advancement of next-generation sequencing has resulted in the creation of massive amounts of genome sequences (reads) in one run. Roche/454 sequencer (400 bps) can create one million short genomic reads where Illumina/Solexa and ABI/SOLID sequencing (75 bps) can create genomic read about 2.4 billion short genomic reads. Unfortunately, the massive amount of data which is generated by these applications tells us very little information about the DNA [19]. Not having enough analysis tools and software, computational power as well as efficient algorithms are the main reason for this problem. As a result, bioinformaticians have begun to consider more effective methods for handling and analyzing this massive quantity of data.

One of the areas which have pulled in many experts in to work on the predicted genes from the generated sequences of the genome responsible for producing proteins to develop traits with great economic values. For example, there are many proteins

that induce greater muscular growth, better height, and weight. A critical aspect of the pipeline is selecting an efficient protein that has a complete hit with the predicted genome sequence with high precision. For this, going through all the proteins related to the desired trait in a protein data bank such as Uniprot [40] and downloading all the related proteins is the initial step. But it's unlikely to have complete hits of all the proteins with the predicted genome sequences of *Bos frontalis*. So, a filtration process to find proteins with better expression and hits in our generated genome is the second step to finalize a protein to work with. Furthermore, comparing the target protein from our generated sequences with the same protein of other related species like *Bos indicus*, *Bos taurus*, *Bos gaurus* from protein data banks, finding and analyzing the differences in the proteins, an evolutionary difference and link can be highlighted

1.1 Motivation

It is clear that Next-generation sequencing (NGS), also called high-throughput sequencing, can provide high-quality genomic data in a faster and cheaper way. By using this better-quality data, we can get genetic variants that can be used to detect different characteristics and structural details of animals and crops. These variants data can also be used to find out the reasons for these characteristics. So, our target is to find out some genetic variants of our selected species and compare this data with other near species. This type of research result can be used to produce high-quality hybrid animals and crops which can be a solution to the world food crisis problem. Genetics variants data can also be used to detect many fatal diseases like cancer diabetics at an early stage.

In Bangladesh, a scientific study on gayal began in 1985. Bangladesh Livestock Research Institute began an ex situ conservation and enhancement program for this species in 1990. On-station performances and adaptation studies of gayal and gayal x Holstein hybrids were part of the program. However, no useful results have yet been obtained from that program. Except for the research of Furer-research Haimendorf's (1980), there is no scientific assessment on the state and performance of gayal in Bangladesh's home tract. Actually, gayal is an endangered animal genetic resource in Bangladesh, and actions should be taken to protect it through economic methods. According to a recent study, In India, the mithun is also experiencing population loss, owing to an increase in cross-breeding.

Generally, gayal is raised for producing meat which is extremely popular among the tribal peoples of India's north-eastern territory. Nonetheless, some tribal people in Bangladesh's Bandarban Hill area and Chittagong district, as well as north-eastern India, regard Gayal as a ritualistic animal that plays a major part in their social, economic, and cultural lives. Gayal milk is high in protein and fat, just as buffalo milk. Gayal x cattle hybrids are widely employed as milk animals in some parts of north-eastern India and Bhutan. As a result, this species could be considered as a supplement of those region's milk production system. Furthermore, mithun skin has now been reported to produce high-grade leather. Bangladesh has become self-dependent in meat production, headed by poultry, but it still falls short in milk output when compared to demand. Despite a 319 percent growth in milk output

to nearly 100 lakh tonnes in roughly ten years, it still falls short of the current demand of over 152 lakh tonnes. It has already been stated that further research on *Bos Frontalis* can bring rewarding results as hybrids from gayal or *bos frontalis* can produce good quality meat and large quantities of milk.

1.2 Aims and Objectives

The research activities were focused on the following specific objectives:-

1. To develop a draft mitogenome and whole genome of *Bos frontalis* with high sequence coverage (50X) using state-of-the-art Next Generation Sequencing (NGS) approach besides other modern computational biology tools.
2. To determine the genetic characterization & genetic diversity of selected *Bos frontalis*, as well as to detect distinct features of *Bos frontalis*.
3. To determine a specific economically important trait of Gayal and extract the protein related to that trait and compare it with similar Bovine species.
4. To strengthen national and international collaboration towards capacity building and human resource development to promote Gayal conservation

1.3 Thesis Organization

In this thesis starting from DNA sequencing, Genome assembly and annotation, functional gene and protein prediction to Genome Wide Association Study for finding one type of economically important traits have been defined on each step. Using multiple tools we tried to compare our findings and publicly available protein dataset to assess the evolutionary difference between that protein of our studied species and other bovine species.

In chapter 2, we have discussed some of the basic topics of life science and bioinformatics in section 2.1.

In chapter 3 we have discussed the material we have used and the method we have followed for our thesis. In Section 3.1 we have described our data. We discussed the source of our data in the Section 3.2 of chapter 3. The most significant part of our Pipeline flowchart has shown in section 3.3.

Where in section 4.1 Data quality, in section 4.2 Draft genome information, in section 4.3 Genome annotation, in section 4.4 Genome Wide Association Study are shown

In chapter 5 it shows the concluding and future work and chapter 6 is set up with a biography.

Chapter 2

Review of literature

The origin of endangered bovine species is still contested, and data related to their genetic architecture is limited [35]. They believed that making the full genome sequencing data of bovine species and assembly of that data will help to overcome this challenge. At the same time, it will create a lot of information, including mithun's phylogenetic standing. In reality, the primary goal of the research on India's mithun was to create a genetic data and a de novo reference genome for the species mithun to aid future analysis. Their de novo draft assembly approach is the first hybrid technique for Indian mithun's genome assembly. They had obtained genome assembly output which had low fragmentation, good coverage, and better completeness. They believe that, their mithun's genome assembly will create a genomic resources for evolutionary studies in conjunction to other near species, as well as aid in understanding the genetic architecture of many phenotypic and genotype interactions that underpin this distinct cow species from various geographic environments.

Bos frontalis (Gayal) is thought to have been domesticated from its ancestor species gaur some eight thousands years ago, according to the experts. However, researchers have suggested that *Bos frontalis* (Gayal) is either a complete separate species or may be a hybrid descendent of gaur and cow. One identifying feature of *Bos frontalis* that, it has a bony dorsal ridge and whitish stockings on its both forelegs, among other morphological characteristics. It also has a karyotype ($2n = 58$) that distinguishes it from cattle (*Bos gaurus*; $2n = 60$) and gaur (*Bos indicus* and *Bos taurus*; $2n = 56$) [14] [7]. Furthermore, karyotype research of Indian mithun and Indian gaur (*Bos gaurus*) has revealed that, both of the species have equal chromosomes number ($2n = 58$), and it is not same as the Malaysian and Chinese gaur [33]. Depending on the mtDNA cytochrome b gene and regulatory region, Ma [5] and Baig [24] propose that, mithun is a distinct species rather than a direct descendent of gaur. Based on the entire mitochondrial genome, Ren [34] recently found that the Chinese Mithun is genetically more similar to cattle rather than gaur. By using long-range PCR and the primer sets below, the whole mitochondrial genome was amplified into 9.2 and 9.6 kb segments. (BF1F 5-AAT ATG CTC GCC ATC ATT CC-3, BF1R 5-ATT GCA GAG GGA AGT CAT GG-3; BF2F5-TCA CCA GCA TAATTC CCA CA-3, BF2R 5-GGC ATG TCA CCA AGG AGA GT-3). The authors' phylogenetic analysis revealed the existence of two varieties of mithun, first of them was the descendent of gaur and the next one was mixture of mithun bull and cow. All three hypothesized possibilities for the genesis of mithun have been

supported by subsequent studies which stands on the whole-genome sequencing and SNP genotyping. As a result, phylogenetic status of mithun is still debated, as the Mithun from Northeast India is actually closer to gaur, yet two highly diversified mitochondrial genomes have been discovered in China. As a result, there's a chance that domestic mithun could have numerous origins.[36]

Researchers employed the PCR-Direct DNA sequencing technology to find genetic polymorphisms in the growth hormone (GH) gene locus of six separate kind of cattle in this study. The major purpose of that research was to find single nucleotide polymorphisms (SNPs) in exon 5 of the GH gene for comparing it to other livestock species such as the Murrah buffalo, Sahiwal cow, Arunachali Yak, Nagaland Mithun, Malpura Sheep, and Gaddi goats. A total of 36 animals from six different species were used in the experiment. The extraction method of phenol-chloroform was employed to separate genomic DNA for PCR amplification [1]. The primer combination was also utilized to amplify the Growth Hormone gene's exon 5 as well as its flanking region. The DNASTAR software's EDIT Seq tool was used to conceptually translate exon 5's coding DNA sequence to amino acid sequence. This work employed NCBI's Nucleotide BLAST tool [41] to scan public databases for sequence homology. This work employed sequence information gained from direct sequencing of PCR products to find SNPs in exon 5 of the GH gene in Indian cattle species. The sequence analysis result showed two potential replacement mutations in Sahiwal cattle as compared to *Bos taurus*. This study [27]highlighted the polymorphism nature of exon 5 in the GH gene of several species.

Another phylogenetic analysis have revealed that Gayal is a hybrid descendent of male wild gaur and female household cow crossing. Gayal ($2n = 58$) has a different chromosomal number than domesticated cattle (*Bos indicus* and *Bos taurus*, $2n = 60$) and gaur (*Bos gaurus*, $2n = 56$). As a result, he claims that Gayal's origin is unknown. Gayal was thought to be a distinct species by some researchers, while others assumed it was a hybrid descendant of wild gaur and domestic cow. Mei and his colleagues sequenced the entire genome of a Gayal from Yunnan Province, China, as well as the genomes of a Red Angus (RAN) and a Japanese Black cow (JBC), which are both varieties of the worldwide popular species B. Mei and his team used next-generation sequencing technology to complete whole-genome sequencing for the Gayal (*Bos frontalis*) for the first time, and the results were compared to those of two widely used domesticated beef cattle breeds: RAN and JBC. SNPs, InDels, CNVs, and SVs were among the many genetic variations discovered. The two less chromosomes in *Bos frontalis* compared to the other near bovine kinds were thought to be the consequence of chromosomal fusion, based on sequencing coverage and variance distribution. The genetic variety annotations revealed that more than fifty percent of the SNPs or InDels in *Bos frontalis* were novel, so there were alarmingly more variants than in RAN and JBC, and this results says that, *Bos frontalis* represents a very important prospective genetic resource which is still need to be fully utilized. When compared to RAN and JBC, the functional annotation of the breed-specific-SNP/InDel carrying genes in Gayal revealed that, there is a possible link between its phenotypic as well as genotypic characteristics. A study of cattle's population history was also carried out, with the results revealing that all three cow breeds have seen a couple of population expansions as well as population

deduction. As Gayal (*Bos frontalis*) is apparently distinguishable from RAN and JBC (*Bos taurus*), the researchers back up the claim by saying that, *Bos frontalis* is not a *Bos taurus* descendant. Gayal (*Bos frontalis*) was also shown to be phylogenetically distant from both *Bos mutus* and *Bos taurus*, according to phylogenetic result which is based on single ortholog copy genes, but phylogenetic study based on mitochondrial genomes revealed that domesticated cattle (*Bos taurus* and *Bos indicus*) are closely related to Gayal. As a result of these discoveries, it's fair to assume that, *Bos frontalis* is a hybrid descendant of male wild gaur and female household cow [31].

2.1 Background

2.1.1 Cell

The word Cell came from the Latin word *cella* which signifies "little room". In 1665, Robert Hooke found cell. Actually, hypothesis of cell was first evolved by Theodor Schwann and Matthias Jakob Schleiden. Cells are the smallest living unit of every living being. Besides, it may very well be said that every single cell is alive. So it is known as the "building blocks" of living creatures. A cell is the fundamental functional unit that can complete various capacities in its body. There are trillions of cells in the human body that are infinitesimal. Primarily cells consist of cytoplasm which contains numerous constituents like proteins, nucleic acids, and a cell membrane to encompass every one of these constituents. These parts are otherwise called organelles. Every one of these organelles is encircled by a different layer. Among these organelles, the major is the core which does the hereditary data that are fundamental for cell growth and reproduction.

Functions of a cell: There are numerous functions that a cell does. The elements of a cell are as below:

- **Produce Energy:** The fundamental function of cells is to take supplements from food and convert those supplements into energy.
- **Give Structure and Support:** Cells give the fundamental construction of living creatures.
- **Work with Growth Through Mitosis:** Tissues develop by duplication of cells like one tissue into two tissues. These two are shaped into four. By this interaction, a child of a living thing is framed into a developed one. What's more, this happens through the mitosis interaction.
- **Permit Passive and Active Transport:** A cell is encircled by a membrane which is a plasma membrane that shapes a significant piece of a cell and behaves like a primary barrier. It allows supplements to enter and wastage items to leave. Cells assist smaller molecules to get inside the cell membrane through passive transport. What's more, get free from the large molecules by the active transport.
- **Reproduction:** The reproduction system is the imperative part to hold the species. A cell accomplishes this work by the cycles of mitosis and meiosis. In the mitosis interaction, the cells basically partition into additional cells to

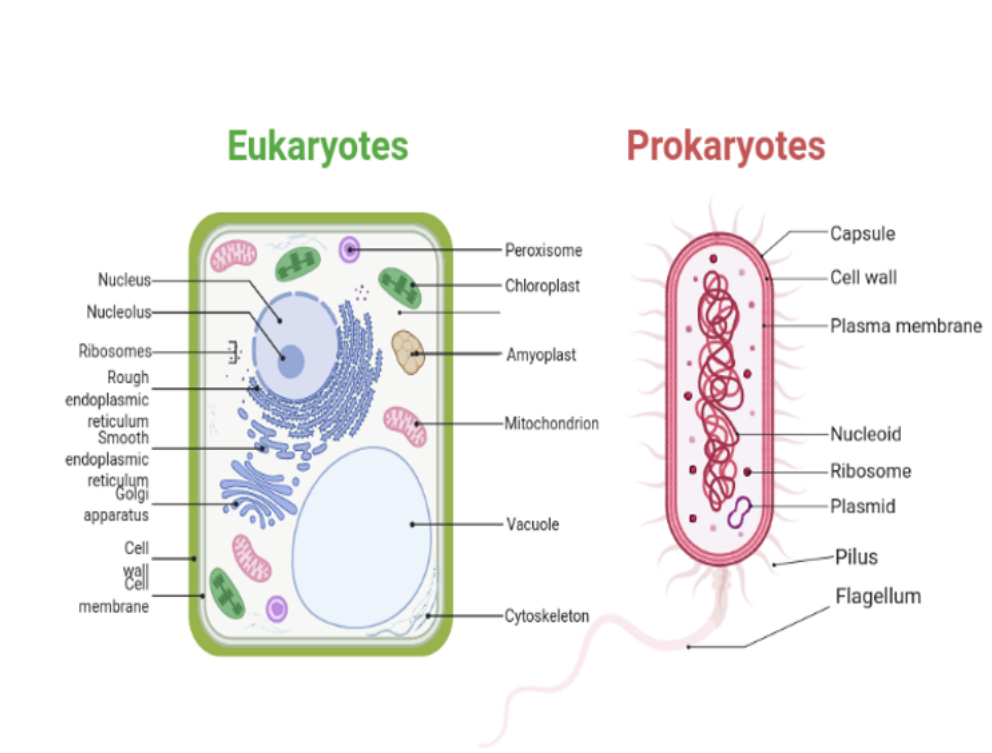


Figure 2.1: Cell throughout the body.

Source: Created with BioRender (<https://app.biorender.com/>)

form new cells. This is named asexual reproduction. Meiosis is the cycle that happens in gametes or reproductive cells where there is a mixing of hereditary data. Which makes it structure a hereditarily unique cell from its parents. Meiosis is a piece of sexual reproduction.

- **Make Metabolic Reactions:** A cell membrane isolates within the cell from the outside. Cells permit life's chemical reactions to occur far faster than they would somehow by moving them into a minuscule region inside a membrane. It assists with keeping alive the cell. It is a chemical reaction. This metabolic reaction has two sorts. One is a catabolic reaction (interaction of energy production) and another is an anabolic reaction (utilizing energy to make a greater substance from the straightforward ones).

Classification of a cell

This cell can be classified in two ways. The classification of cells according to its types is given below:

Prokaryotic Cell: The prokaryotic cell was the first cell on the earth. The prokaryotic cell does not contain any nucleus and cell membrane-bound. This type of cell is usually unicellular (except some cyanobacteria). In this cell's lysosomes, peroxisomes, mitochondria, microtubules are absent. Here multiple proteins fold together and form the prokaryotic cell. Examples of prokaryotic cells are bacteria cells, cyanobacteria, archaea.

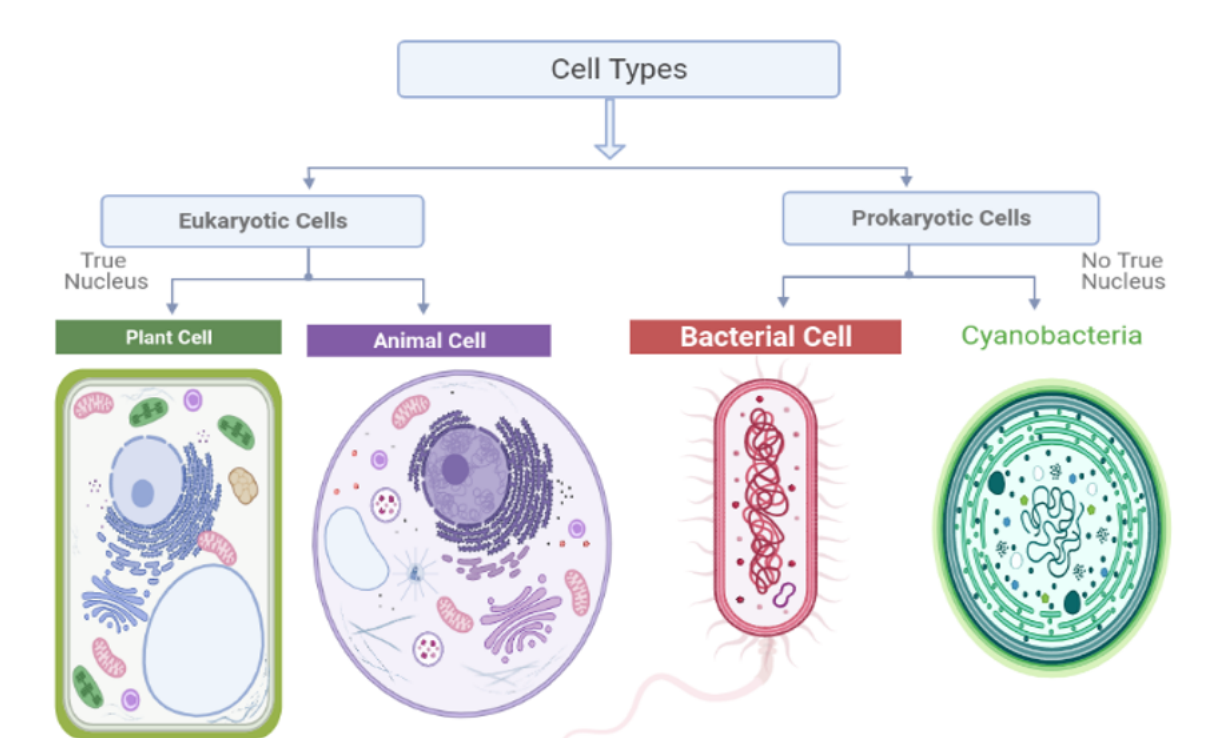


Figure 2.2: Classification of cells.

Source: Created with BioRender (<https://app.biorender.com/>)

Eukaryotic Cell: Eukaryotic cells resulted from the evolution of prokaryotic cells. This type of cell contains a nucleus and cell membrane-bound organelles. Eukaryotic cells can be single-celled or multi-celled. In this cell's lysosomes, peroxisomes, mitochondria, microtubules are present. Eukaryotic wrap their DNA around the proteins. Examples of eukaryotic bacteria are plant cells, animal cells.

2.1.2 DNA

DNA is a sort of form of deoxyribonucleic acid. DNA holds is the sequence of the genetic information in a gene. It is the nucleus of a cell is where DNA is found. However, a small quantity of DNA can occasionally be detected in mitochondria. Nuclear DNA can be found in the nucleus, whereas mitochondrial DNA is found in the mitochondria. DNA remains the same in every cell. The main function of a DNA is to hold the genetic information of a living being and that is done by gathering the information as a code which is made up of the combination of four kinds chemical bases and they are given below:-

1. A (Adenine)
2. C (Cytosine)
3. G (Guanine) and
4. T (Thymine).

In the human body, there are around 3 billion bases, and nearly 99 percent of these bases are found identical. These DNA nucleotides frequently couple up with one another. The lebase A is paired with the base G, and the base C is paired with the base T. They build pairings one by one. These bases also attach with

the sugar molecules and phosphate molecules. Which is called a nucleotide. This nucleotide actually forms a spiral which is called a double helix (twisted double helix). And this helix looks like a ladder. In this ladder, the phosphate molecules and sugar form the vertical-sided pieces where the base pairs form the ladder's rungs.

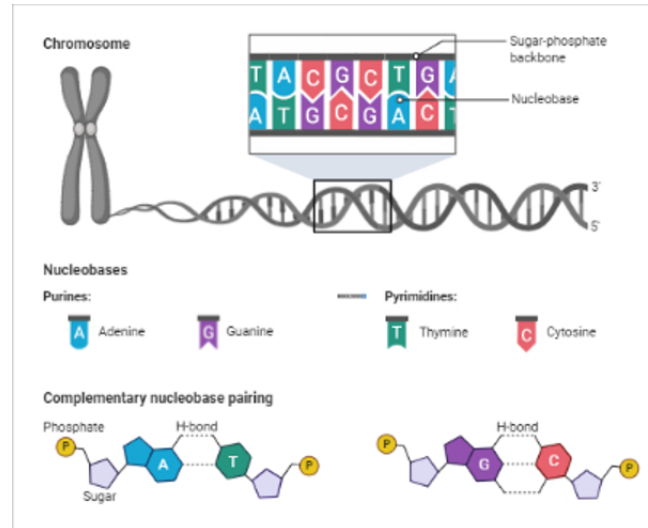


Figure 2.3: DNA throughout the body.

Source: Created with BioRender (<https://app.biorender.com/>)

Functions of DNA:

There are many functions that a DNA does. The functions of a DNA are as below:

- Genetic Information: As it has mentioned above DNA holds the genetic information as well as it also helps to exchange this genetic information from parent to child.
- Transcription: The main aim of DNA is to form a different molecule RNA from DNA. Here the information of DNA is written down on RNA then carries that information to the other parts of the body like a messenger.
- Translations: Ribosomes (cell organelles) transform the messengers' code into protein format or chain amino acids throughout this procedure. Three RNA bases are combined to make one of this amino acid.
- Replications: DNA replicates DNA by increasing the number of chromosomes and cells. Encoding Information.

2.1.3 Genome size

There are 23 pairs of chromosomes in the human body. If we sorted out these 23 pairs of chromosomes, we can get approximately 3.2 billion nucleotides of DNA, which we call this number is genome size. Although we do not analyze the full version of 3.2 Gb, according to finding out the coding, noncoding sequence of the gene we have to sort it out after the whole genome. And this sorted genome is called the whole genome sequence(WGS). To complete the whole genome sequence tasks various types of assembly techniques are followed. Among them, two types of

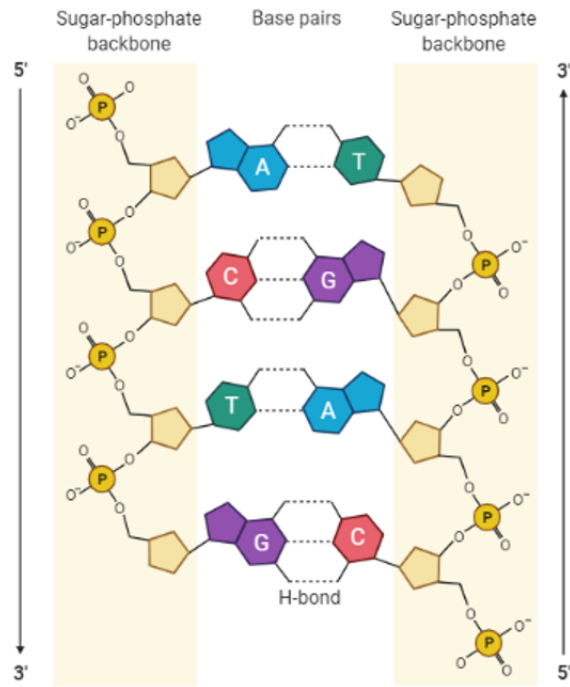


Figure 2.4: Molecular Structure of DNA.

Source: Created with BioRender (<https://app.biorender.com/>)

techniques are very popular, one of which is a copy of the genome of a species, and assembly from it is called reference assembly. And if there is no reference, the De novo technique is followed to sequence the whole genome. Plenty of computational power, skilled analysts are needed for this assembly process.

2.1.4 DNA sequence

DNA sequencing is the action of establishing the sequence of the building blocks, or bases, in a strand of DNA. Before performing DNA sequencing, the DNA must first be sliced up in smaller bits and placed into plasmid DNA before being implanted into bacterial cells. This allows the bacterial cells to make a vast number of copies of it when they grow. After that, DNA of the bacteria is extracted and used for sequencing. After that, The separated DNA is transported to a plate, which will be used for the sequencing procedure. Ingredients are put in a combination. These ingredients contain free DNA bases A,C,G,T DNA polymerase enzymes and DNA primers. These modified DNA bases are labeled with color, fluorescent tags are also added. These bases are known terminator bases. All these things are heated to 96 degrees Celsius for starting the sequencing reaction. As a result, DNA splits into two individual strands. After that, the heat level is decreased to 50 degrees. So the DNA primers bind with the DNA of plasmid. Once again the heat level is increased up to 60 degrees, so the enzyme DNA polymerase connects with the primer DNA. By adding the unlabeled DNA bases with the target DNA, DNA polymerase begins forming a new strand of DNA. It keeps adding DNA bases until it reaches a terminator base. After this, bases can not be inserted into the new DNA strand since the terminator bases have been chemically modified. After adding the terminator base, the DNA polymerase enzyme stops creating DNA and slips away from the strand.

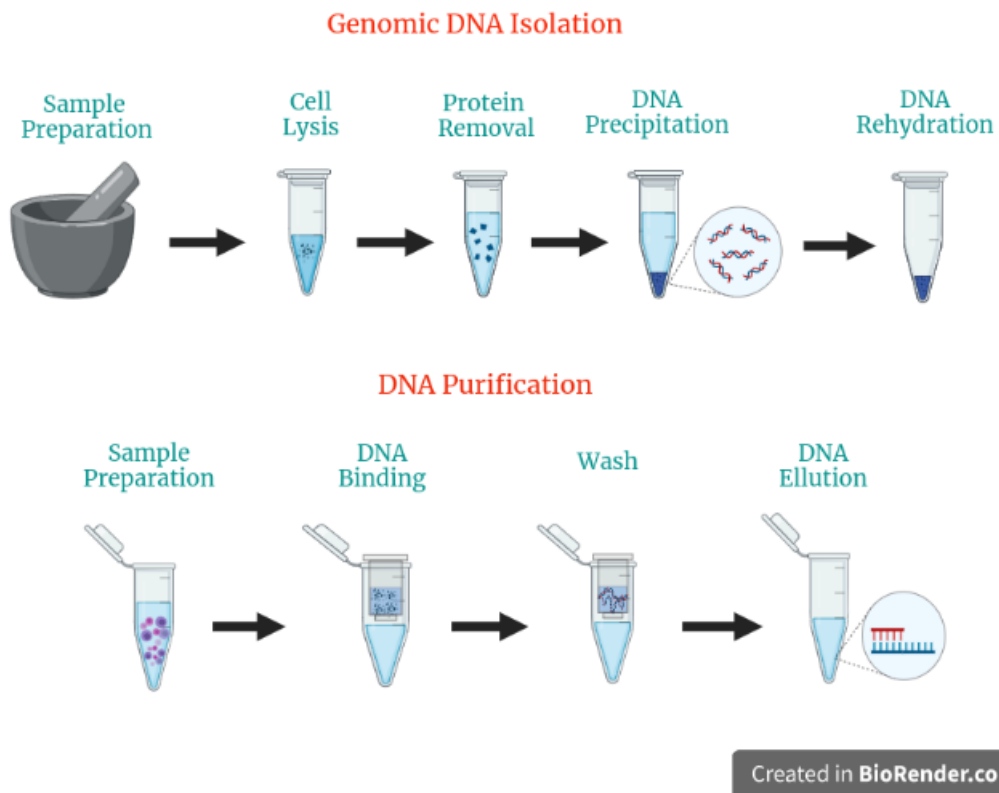


Figure 2.5: Sample collection and processing.
Source: Created with BioRender

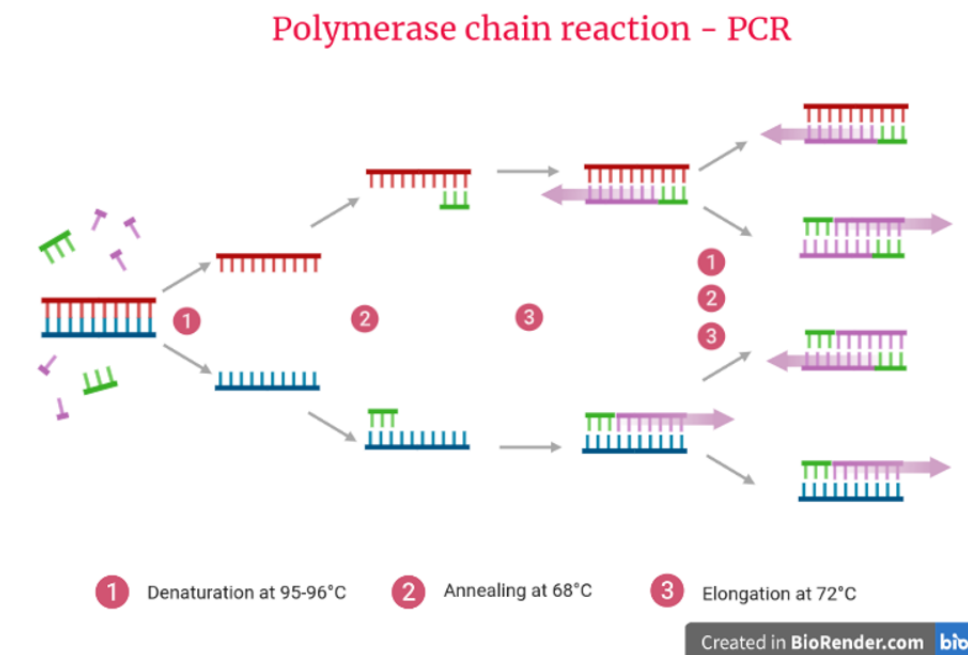
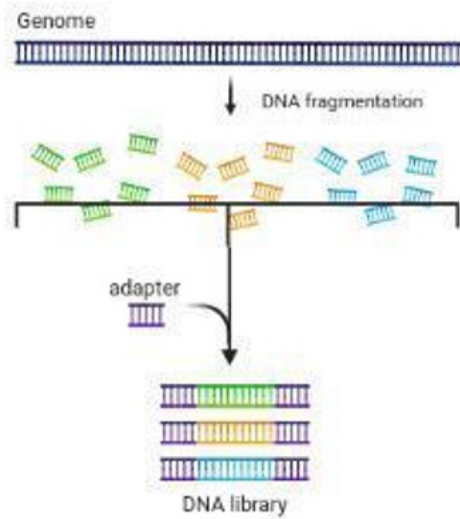
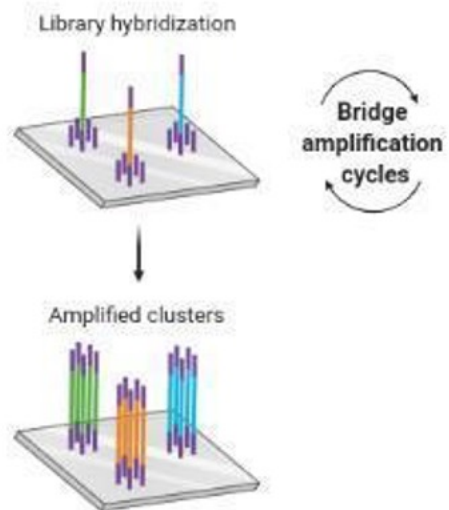


Figure 2.6: PCR Process.
Source: Created with BioRender

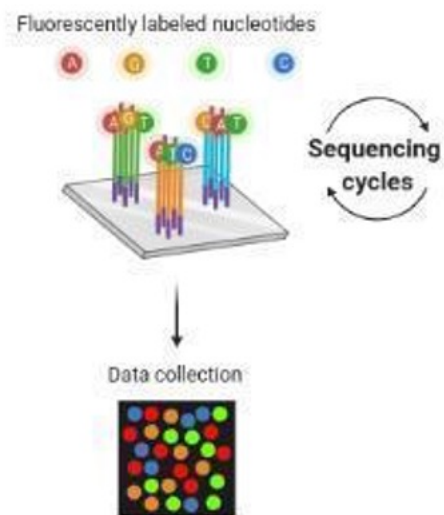
① Library preparation



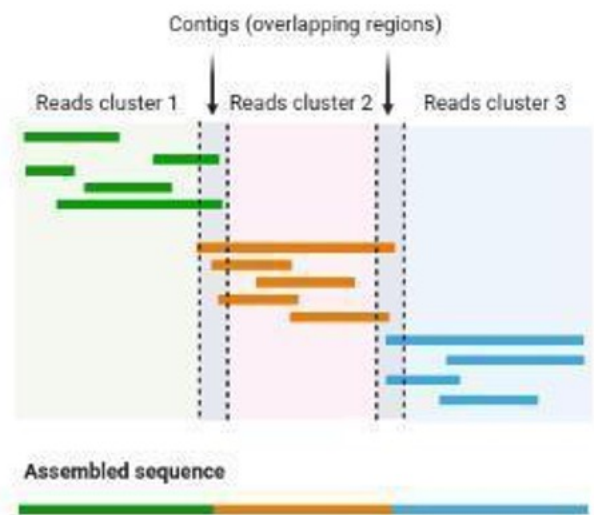
② DNA library bridge amplification



③ DNA library sequencing



④ Alignment and data analysis



Created in BioRender.com

Figure 2.7: DNA Sequencing.
Source: Created with BioRender

All of the thins are is heated upto 96 degrees Celsius once more for extracting the new DNA strand from the old strand. This procedure is done several times to create a large number of DNA fragments of various sizes.

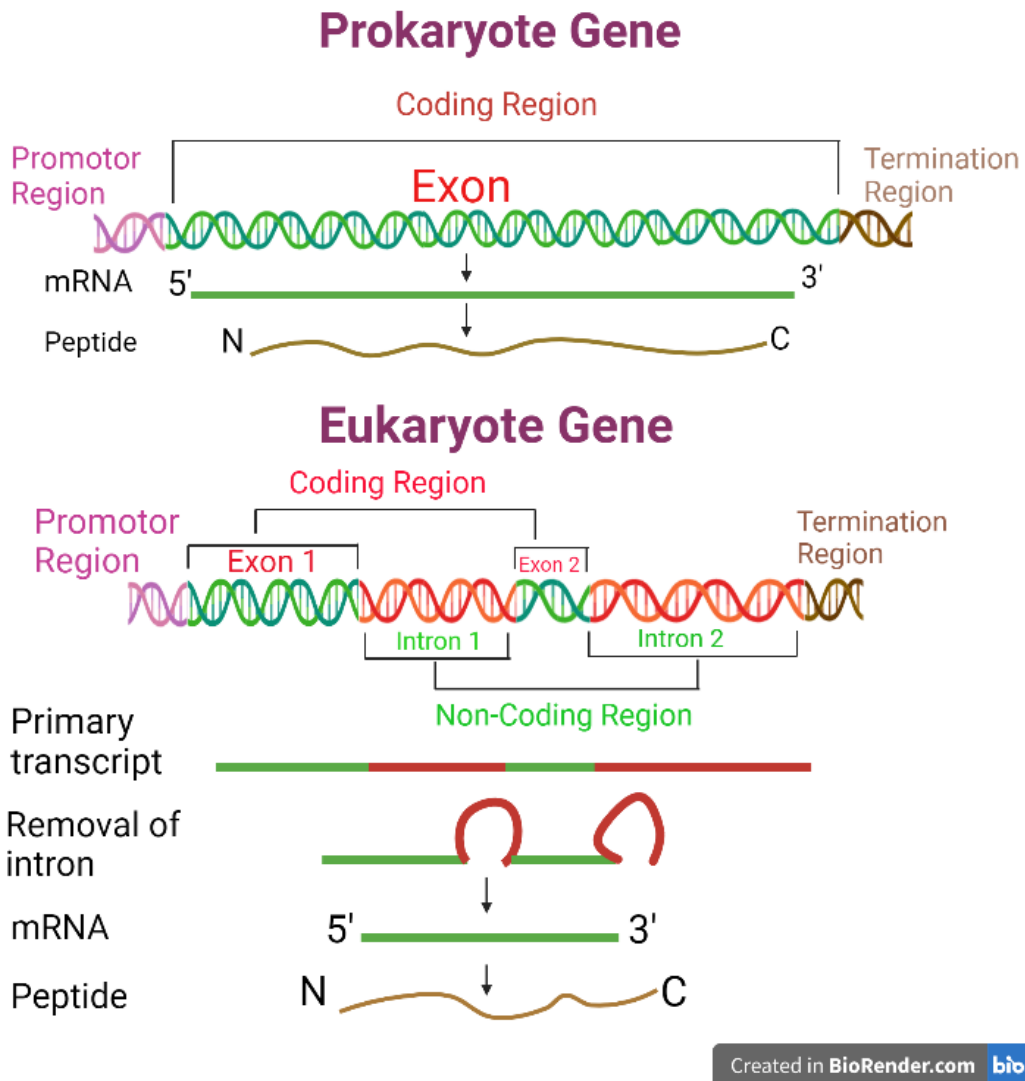


Figure 2.8: Gene and Protein on DNA.
Source: Created with BioRender

The insertion of a terminator base determines the length of each fragment. An electrical charge is applied to each well on the plate through a capillary tube. As a result of this, negatively charged DNA molecules flow through the capillary tube. In each capillary, there is a porous gel. The smaller fragments of DNA travel across the gel more smoothly than the larger DNA fragments. A laser illuminates the terminator bases provided that the DNA fragments reach the capillary's end. A camera detects the color and records it. Each terminator base is labeled with a different color. The

Chapter 3

Materials and Methods

3.1 Sampling and DNA isolation:

A male gayal's fresh blood sample was collected from Rangunia, Chittagong, Bangladesh and was taken to the laboratory to be preserved at 70°C for successive analysis in December 2019. Following that, the blood tissue was anatomized and preserved in 95% pure ethanol. Later, by using AddPrep, Genomic DNA extraction equipment (AddBio, Korea), genomic DNAs with high molecular weight were separated and refined for further evaluation of the DNA's quality and quantity. The purified DNA was then forwarded to a commercial source for library preparation and sequencing. BGI's Illumina HiSeq X Ten platform was used to sequence DNA. As necessary, all techniques were carried out in compliance with the "Regulations for Animal Experiments" at Chittagong Veterinary and Animal Sciences University's special attribute, the Indian and GOB ethical approval.

3.2 Library preparation:

The extracted DNA was cleaned up using commercial kits and sent for subsequent library preparation for performing whole-genome Shotgun sequencing (WGS) at the BGI genomics, China. Using Next-generation sequencing (NGS) technology on an Illumina HiSeq X Ten platform a total of 36 Gigabase pair (Gb) of subread bases with a read length of 150 bp were induced. After sequencing quality of primary sequence reads and trimmed sequencing reads were investigated by using FastQC version 0.11.8. The quality control of the reads was done including removing adaptor sequences, low-quality reads and contamination from the initial reads using BFC. A total of 55,264,452,615 polished reads were comprised in the assembly with 48X coverage.

3.2.1 Data availability

The Illumina raw reads of *Bos frontalis* have been deposited in the SRA [Project ID: PRJNA655910] under the Accession numbers SRR12436427. This WGS project has been uploaded at DDBJ/ENA/GenBank under the accession number SRX8932116 and the assembled genome at GCA017311355.1. In this paper we have used the version JAFDUV000000000.

NCBI Resources How To nextgeninformatics My NCBI Sign Out

BioProject BioProject Search Help

Advanced Browse by Project attributes

Display Settings: - Send to: -

Bos frontalis (gayal) Accession: PRJNA655910 ID: 655910

Whole genome re-sequencing of Gayal (Bos frontalis) and genome annotation to unveil genetic variations to explore the evolution and adaptation at genome level

The The gayal is a large-sized endangered semi domesticated bovine species belonging to the family Bovidae, tribe Bovini, group Bovina, genus Bos, and species Bos frontalis. [More...](#)

[See Genome Information for Bos frontalis](#)

NAVIGATE ACROSS
7 additional projects are related by organism.

Accession	PRJNA655910
Data Type	Genome sequencing and assembly
Scope	Multiisolate
Organism	Bos frontalis [Taxonomy ID: 30520] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Artiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos; Bos frontalis
Submission	Registration date: 9-Mar-2021 Bangladesh Livestock Research Institute; Chittagong Veterinary and Animal Sciences University
Relevance	Agricultural

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (WGS master)	1
SRA Experiments	4
OTHER DATASETS	
BioSample	5
Assembly	1

Assembly details: [Download](#)

Assembly	Level	WGS	BioSample	Taxonomy
GCA_017311355.1	Contig	JAFDUV000000000	SAMN16965583	Bos frontalis

SRA Data Details

Parameter	Value
Data volume, Gbases	147
Data volume, Mbytes	47190

Related information

- Assembly
- BioSample
- Genome
- Nucleotide
- SRA
- Taxonomy
- WGS master

Recent activity [Turn Off](#) [Clear](#)

- SRA Links for BioProject (Select 325061) (92) SRA
- Cetartiodactyla BioProject
- Bos gaurus Genome
- bos gaurus[orgn] (1) Genome
- bos gaurus AND (cluster_public[prop] AND SRA

[See more...](#)

Figure 3.1: Our dataset on NCBI.

3.3 Genome assembly:

For assembling the whole genome several assemblers were used and finally took the best assembly from them. ABySS [32] program was used along with Platanus (version 1.2.4) [26] assembler, MEGAHIT (version 1.2.9) [29], SOAPdenovo2 (version 2.04) [22]. Because, at present there is no de novo assembler guaranteed to exceed others, and because assemblers' overall performance varies depending on the given dataset, these distinct assemblers were employed, and an assembly evolution was then done to discover the best assembler. These assemblers use the classic De Bruijn graph illustration even though the assembly algorithm differs across methods. Finally, to evaluate the standard of the assembly BUSCO (version 5.1.2) [28] was used in respect of gene completeness.

3.3.1 WGS data processing and filtering:

The raw sequence data quality was assessed using FASTX-Toolkit [2], Trimmomatic [25] and Trim Galore [37]. All kind of reads with poor quality, prospective sequencing error were removed. We defined the following types of reads as low-quality reads: 1. Reads with Ns more than 10% of their length. // 2. Reads with a low-quality base more than 50% of their length (With ≤ 5 quality-score); // 3. Reads containing ≤ 10 bp adapter sequences (with mismatches allowed ≤ 2 bp // 4. Small insert size paired-end reads that were overlapped 5. Paired-end reads of read1 and read2 are entirely indistinguishable. These paired-end reads would be considered artifacts of the PCR experiment. We also implemented an error correction procedure on the remaining high-quality reads, with the same method described by Li. All the reads were then checked using FASTQC [38]

3.3.2 Scaffold construction:

The high-quality reads after the above filtering and correction steps were exploited to generate scaffolds using ABySS, MEGAHIT, Platanus, and SOAPdenovo2. All of the reads with good quality were stored inside computer-memory, and the overlap between the reads was represented using de Bruijn graph data structures. The graphs were then cleaned up by deleting erroneous connections and using read route to solve minor duplicates. The high-quality reads were then realigned onto the contigs, yielding aligned PEs (Pair-End sequences). We estimated the PE connections between each pair of contigs and then built the scaffolds one by one, starting with small insert size PEs and working our way up to large insert size PEs. Finally, the gaps in scaffolds were filled, which are most likely caused by repeats, using the high-quality PE reads.

3.3.3 Assembly visualization:

A number of web based open-source software are available for WGS analysis. Bandage [30], Cortex [21] were used for visualizing de novo assembly graphs. QUAST [39] were also used for genome assembly evaluation and comparison.

3.3.4 Validation of assembly and Gap resolution:

Once the assembly was achieved, further verification was required. Sequence gap locations were determined. We will utilize simple alignment techniques like BLAST to find identical gap locations on assemblies. A trans-scaffold break was defined as a location where pieces corresponded to two different scaffolds or chromosomes. The gap is deemed open if one or both pieces surrounding it do not align, or if there presents indefinite bases (two or more) between aligned fragments. SAMtools depth version 1.3 confirms the gaps to evaluate Illumina WGS read alignments from the sequenced animal to the gap area. The gap was considered unresolved if bases (1 or more) in the field region had a read depth of <5 . GMcloser and GapFiller [23] were also used.

3.3.5 Final evaluation of the assembly:

The N50 metric and the base quality of the generated contigs were used for additional analysis of assembly. An actual standard for genome assemblies is base quality, obtained by BWA and based on the PHRED quality scale [3]. The N50 metric, which is the same size of the smallest contig that has to be appraised to cover minimum 50% of a sequence. Assemblies with higher N50 scores are thought to be more effective. MUMmer v.4 [4] can also be used for assembly evaluation.

3.4 Gene prediction and functional annotation:

In every genomic sequencing model gene structure prediction is the first step in genome annotation. MAKER ver. 3.01.03, which establishes probability distributions using hidden markov model for the different parts of the genomic sequence, was used to predict genes. Both specified and default parameters were used to do ab initio gene prediction. Functional annotation was obtained by InterProScan ver. 5.46-81.0. The functional annotation report has been stored in Figshare database.

3.4.1 Annotation of protein-coding genes:

Information from many approaches was combined to predict protein-coding genes, including ab initio prediction, homology-based annotation, and. Augustus and MAKER tools were used to predict genes from the start.

Homologous genes which are shorter than the reference genes by two or more exons will also be filtered out. The homology gene set derived from sheep reference proteins and cattle reference proteins were then merged. The record with the longest coding sections and/or the highest gene-wise score was kept for each gene locus. Poor-quality genes (containing gaps in more than 10% of the coding region) and short genes (CDSlength < 150 bp) were filtered. Finally, InterProScan was used to annotate all protein-coding gene models.

3.5 Genome Wide Association Study (GWAS) to compare specific trait:

When the genome and all genes are ready for further analysis we discuss which economically important traits or gene-protein we will work on. We've decided we will focus on muscle development or growth-related genes because our data sample was Male so we couldn't practically work on milk production with this dataset. Then we search and find all the genes and proteins of Bovine species related to growth and muscle development. We've statistically analyzed the downloaded protein and our dataset and sorted out that Leucine and Leucine type proteins are more expressed than other growth-related proteins. Then we finally analyzed and visualized the difference between our predicted proteins and similar Bovine proteins with nest steps.

3.5.1 Data collection

We've collected 1083 growth and muscle development-related proteins from Uniprot [40] and manually curated them. There are numerous unreviewed proteins on Uniprot and other public databases. We've collected only reviewed datasets that are biologically tested and confirmed. We've simplified the header names of the downloaded proteins for better understanding the further analysis.

3.5.2 Finding most expressed proteins

Two command line software tools Transcriptome to Nucleotide BLAST (tBLASTn) and Protein BLASTp used to search the downloaded proteins on our genome and predicted proteins. The genome contains nucleotides (A,C,G,T) and protein contains amino acids (20 amino acids translated from nucleotides). So to search protein query on nucleotide subject needs to be done with tBLASTn and BLASTn used to protein-protein search like searching protein file on predicted protein. After checking the search results we've sorted out all the matches less than 90 percent and E-value of 0.001. After that, we've chosen a statistically significant protein and found Leucine is most expressed in our genome. Then we've again collected all Leucine-related proteins and searched again which specific Leucine-related protein is most expressed and we've found 17 Leucine-related proteins and LRRN1 is the most expressed one.

3.5.3 Extracting the most expressed protein

We've used SAMtools and Bedtools to extract the most expressed proteins from our search results. The search result only contains the protein position on the genome or predicted gene. So we have to manually make a bed file (tab-separated text file with .bed extension) of proteins' start position and end position and use that file to extract the specific proteins from the genome. SAMtools used to make a fasta index file which will help Bedtools to extract the final proteins.

3.5.4 Multiple Sequence Alignment

A Multiple Sequence Alignment also called as MSA, was done by using extracted proteins and collected Leucine-related Bovine proteins with MAFFT software [42]- a command-line tool to create Multiple Sequence Alignment. Another program called MUSCLE can be used to do the same work. But we've used MAFFT for our research. Finally, we converted the FASTA output of MAFFT software to Phylip format so we can use that for creating a Phylogenetic tree in the next step.

3.5.5 Phylogenetic tree construction

PhyML (web-based tool for constructing phylogenetic tree with Phylip formatted alignment file) was used to construct the Phylogenetic tree of our predicted Leucine related protein and other Bovine Leucine related protein. Another program called RAxML (a command line and web-based tool to construct phylogenetic trees from FASTA and Phylip formatted alignment files) can be used to do the same work. We've downloaded the phylogenetic tree as Newick format and uploaded it to iTOL (web-based tool to manipulate phylogenetic tree) to visualize the tree easier to understand. We changed the tree view to circular form and color the branch according to Bootstrap value and created another one with just the Bootstrap number. And finally downloaded the tree in PNG format.

3.5.6 Visualizing protein 3D model

To visualize the most expressed Leucine protein we've predicted, we've used Swiss-Model (a web-based 3D modeling tool for protein) to visualize the protein in 3D. A pdb formatted file exported for future use.

3.6 Total project pipeline

Here is the visual representation of the total Pipeline of our project.

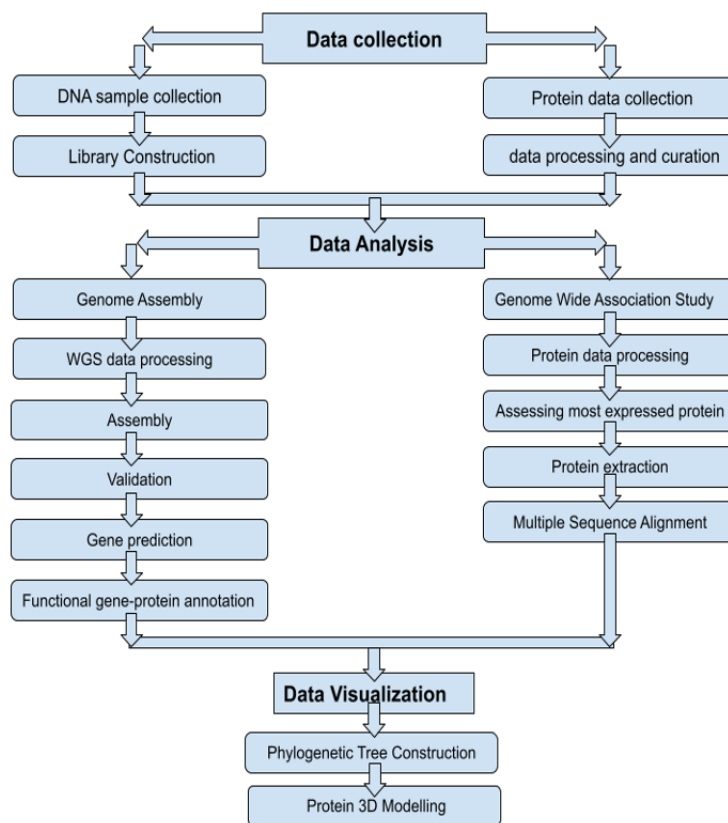


Figure 3.2: Pipeline of this study.

Chapter 4

Results and Discussion

4.1 Data quality

We get quite a good Fastqc report of our data that's why we have not performed any trimming. Per base sequence quality of read seems to be a very good category for each and every forward and reverse base.

The actual distribution should fit pretty close to the theoretical distribution of % GC per 50 bp sequence. From Figures 4.5 4.6 we can see that the actual distribution fits pretty close to the theoretical distribution of % GC per 50 bp sequence.

Per sequence quality score is a very important parameter to determine the quality of data. We get very good results per sequence quality score for all sequences. Where the read average quality was above 35.

4.2 Draft Genome

A total of 2.7 Gbp reads and 1,262,134 scaffolds were generated in ABySS. We have listed the details of standard assembly metrics in Table:4.2. Among the other assemblers, we got best assembly from ABySS because it generating fewer and longer sequence with higher N50 scaffold level. BUSCO analysis suggested that ABySS is the best assembler in terms of our genome completeness. In NCBI GeneBank under the Accession numbers JAFDUV000000000 our genome sequence data have been submitted.

4.3 Genome Annotation:

Generic 21610 predicted genes and 18638 confirmed unique GOs (Gene Ontology on Public Database) found on our Genome from Gene prediction with AUGUSTUS and MAKER ab-initio gene prediction tools. All of these genes are not economically important. Most of them are just housekeeping genes and genes that are helping the cell to regulate all biological needs. We've found 56.5% gene completeness on BUSCO benchmark with Eukaryotic Ortholog Database. There are 21.13% repeat content found with the help of RepeatMasker on our Genome which is normal.

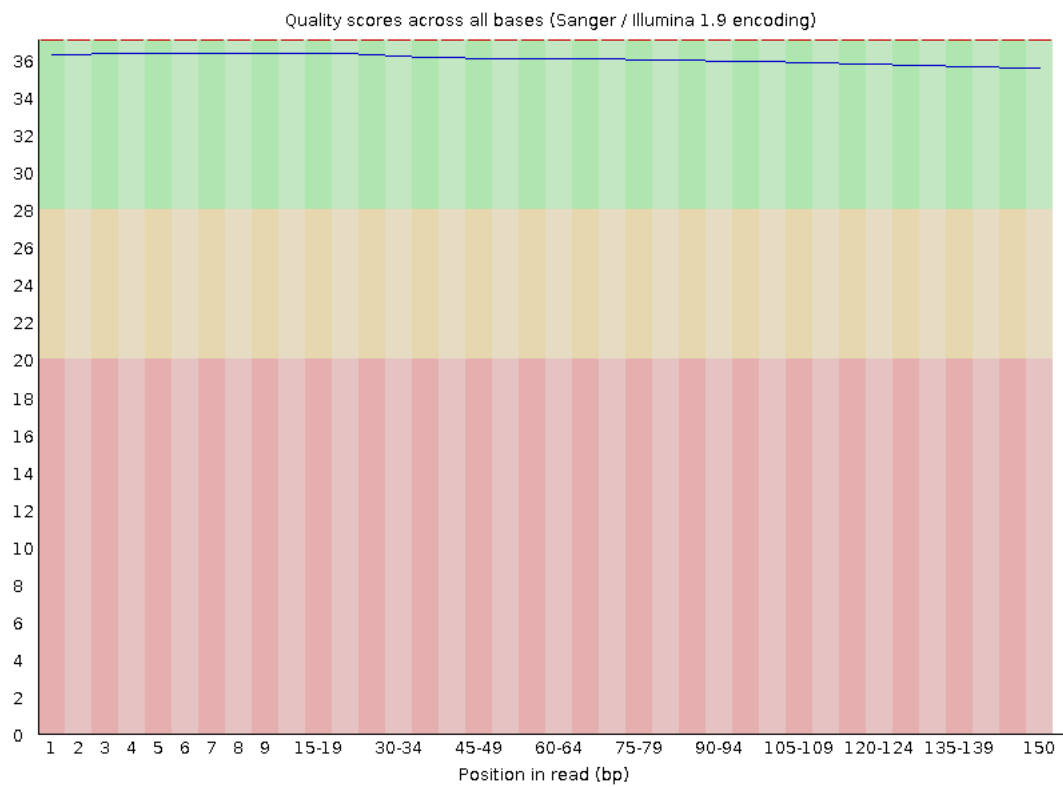


Figure 4.1: Per base quality.

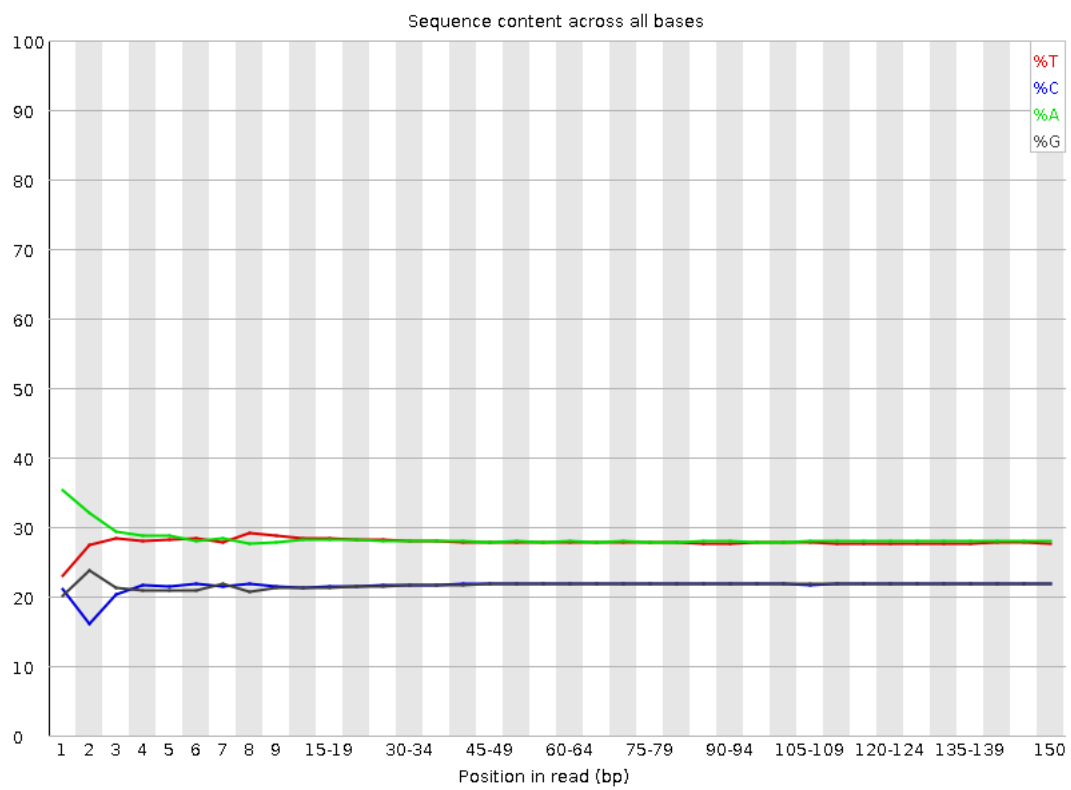


Figure 4.2: Per base sequence quality.

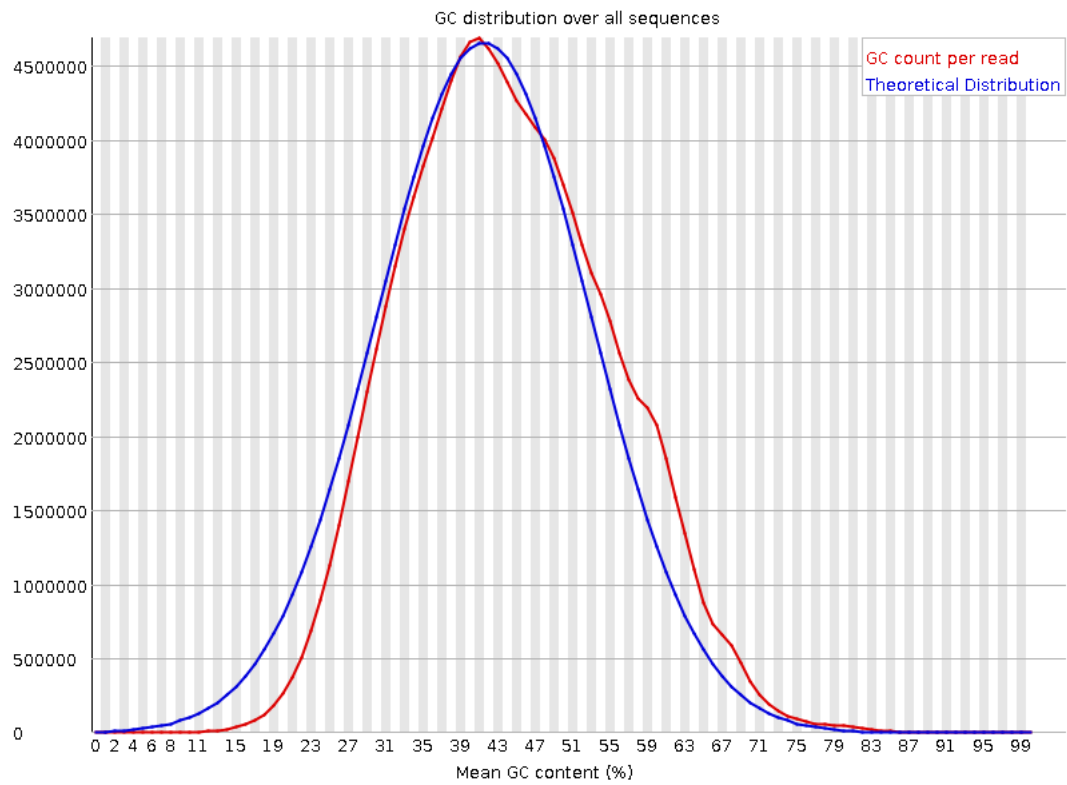


Figure 4.3: Per sequence GC content.

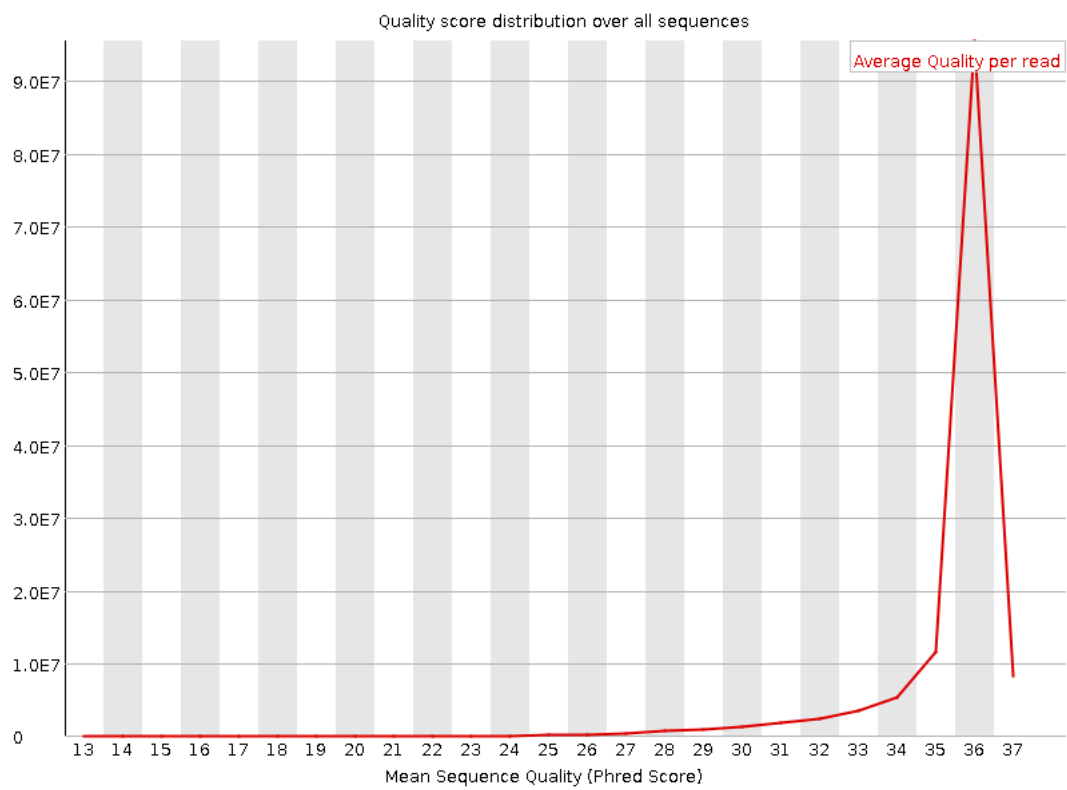


Figure 4.4: Per sequence quality.

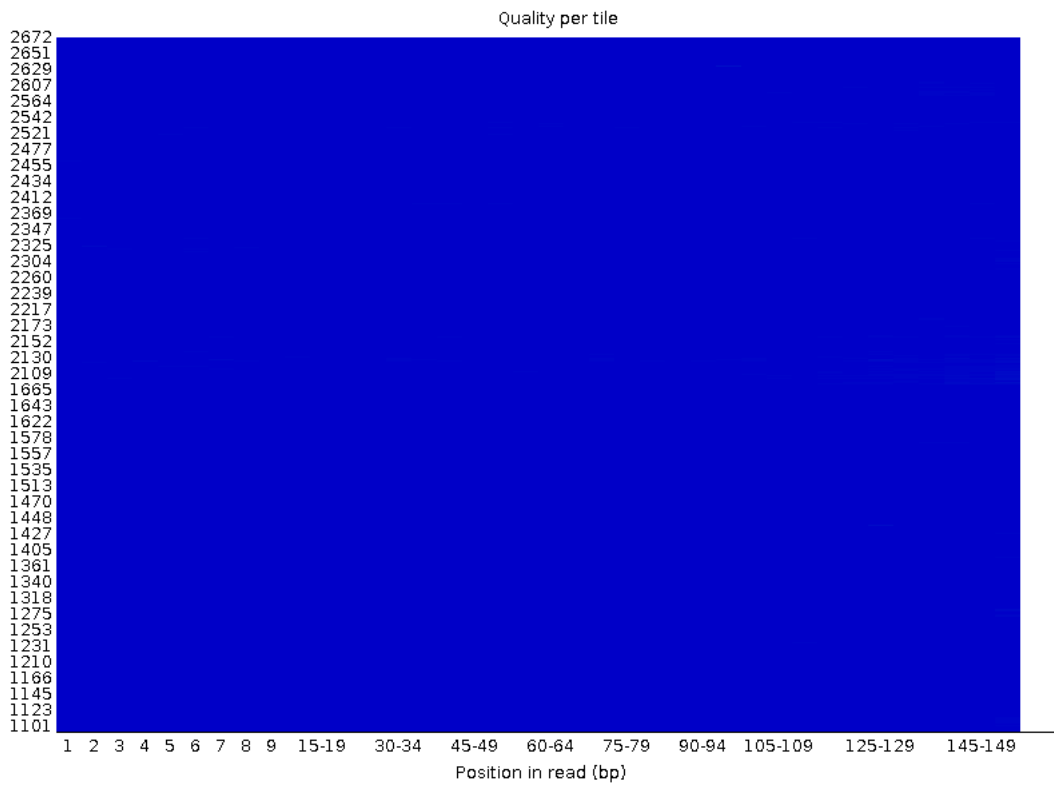


Figure 4.5: Per tile quality.

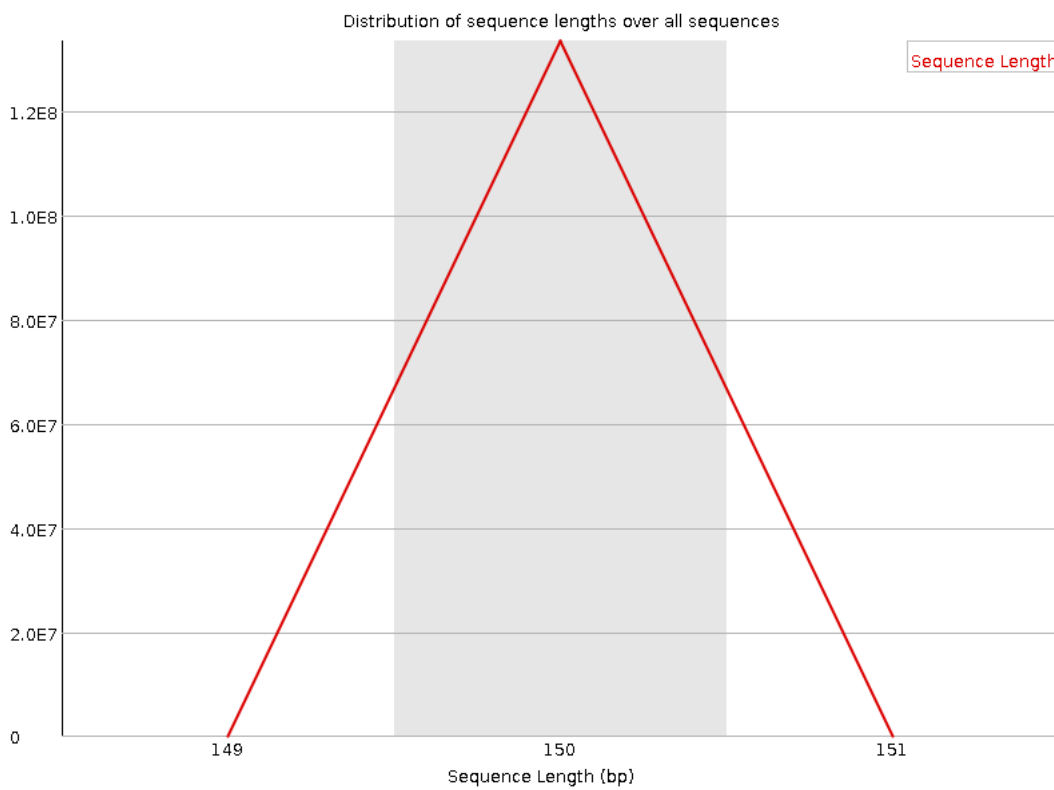


Figure 4.6: Sequence length distribution.

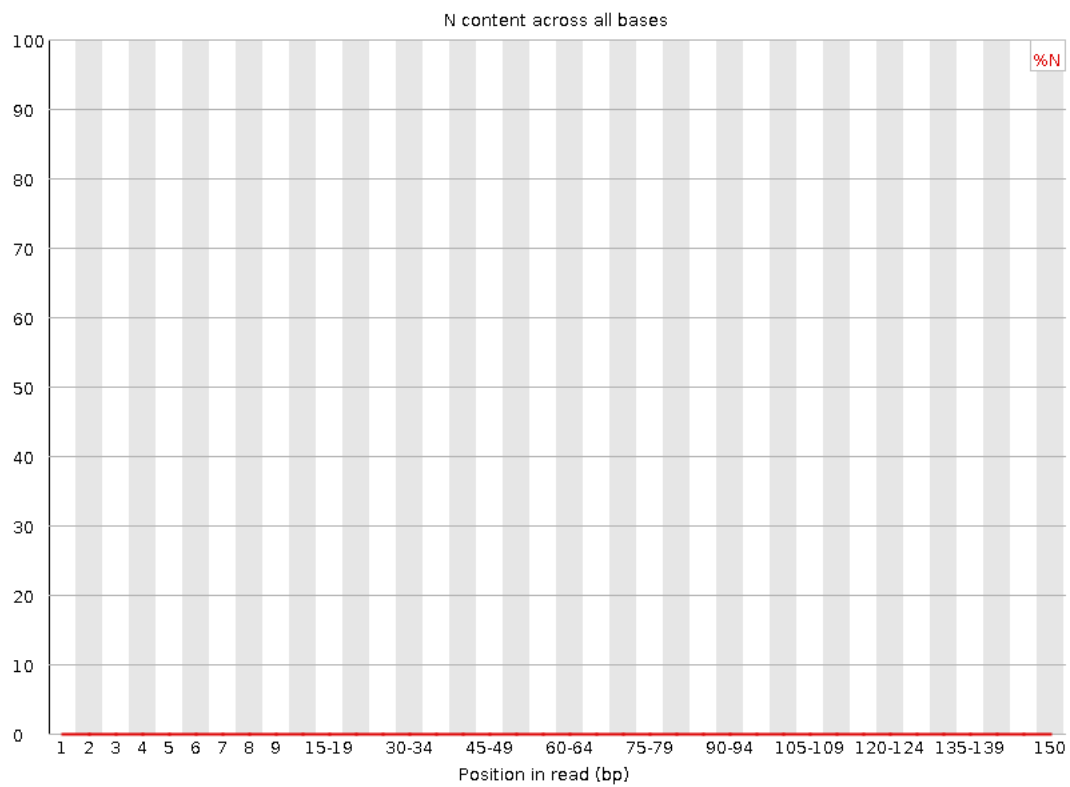


Figure 4.7: Per base n content.

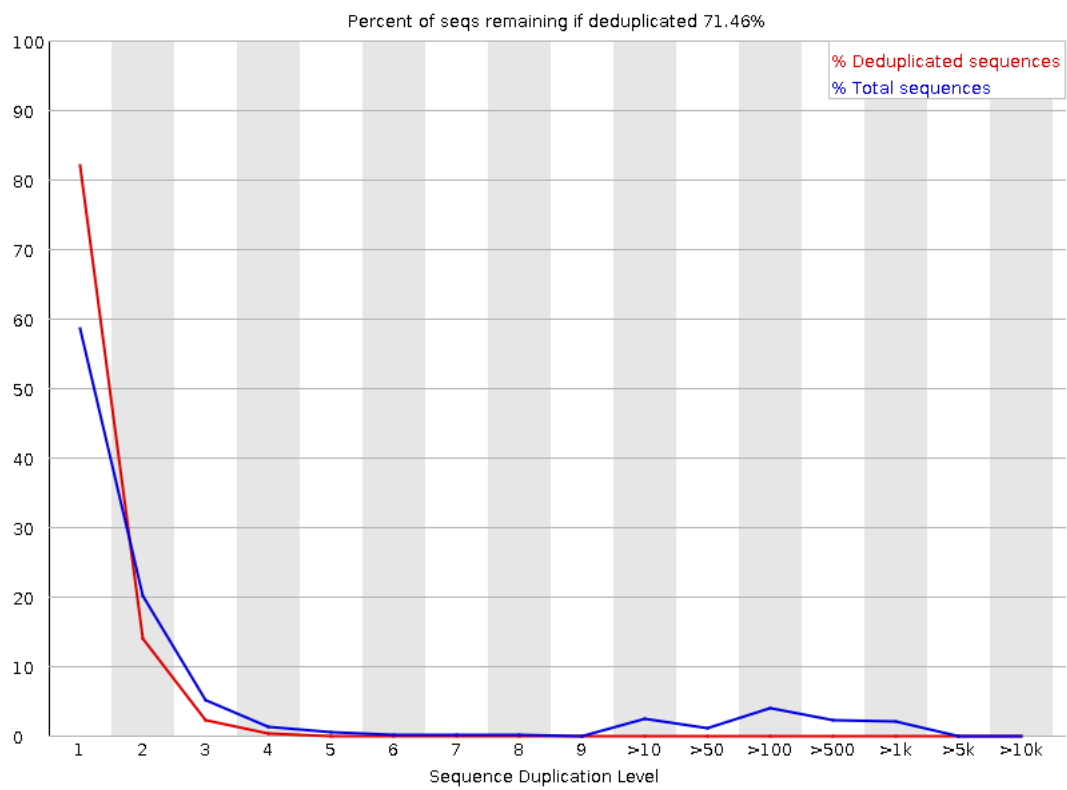


Figure 4.8: Duplication levels.

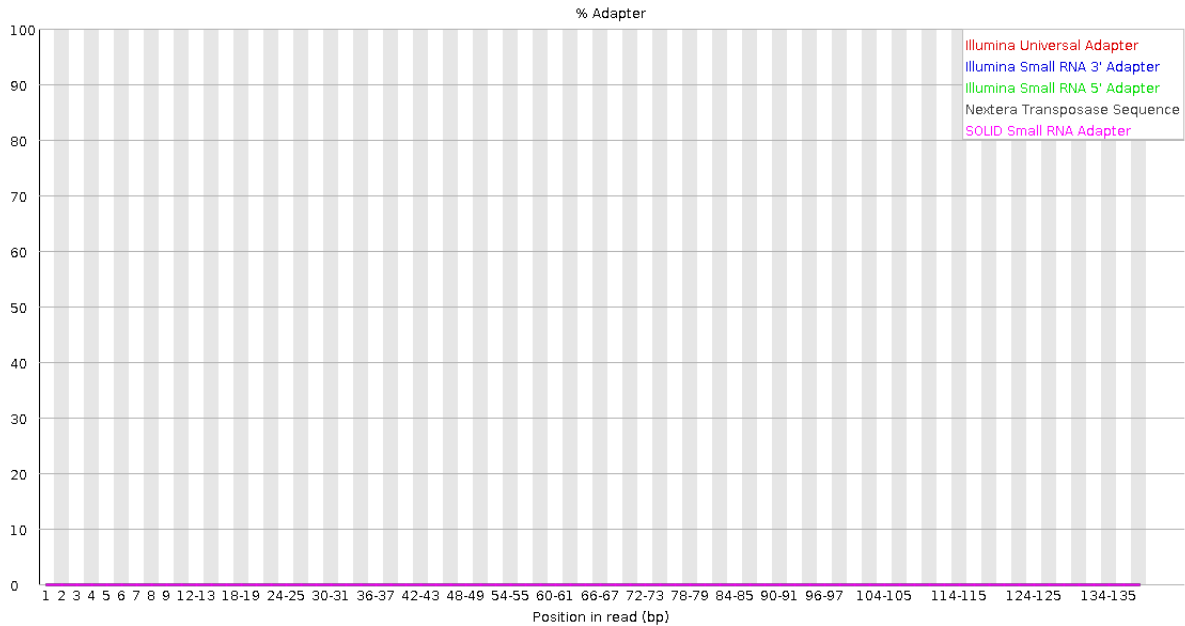


Figure 4.9: Adapter content.

Serial No	Dataset	File types [file Type]	Accession Number]
Data file1	BioProject	Website link	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA655910
Data file2	SRA Raw Data	FastQ File on Website link	https://www.ncbi.nlm.nih.gov/sra/SRR12436427
Data file3	Whole genome assembly data	FASTA	NCBI GeneBank Assembly [Accession number: GCA_017311355.1][https://www.ncbi.nlm.nih.gov/nuccore/JAFDUV000000000]
Data file4	Whole genome sequence	FASTA	NCBI GeneBank [Accession numbers : https://www.ncbi.nlm.nih.gov/nuccore/JAFDUV000000000.1/]

Table 4.1: Overview of data files/data sets

Total sequence length	2,795,315,450
Total Ungapped length	2,795,315,450
Gaps between scaffolds	0
Number of Scaffolds	1,262,134
Scaffold N50	9,646
Scaffold L50	76,340
No. of contigs	1,262,134
Contigs N50	9,646
Contigs L50	76,340
Total number of chromosomes and plasmids	-
Number of component sequences [WGS or clone]	1,262,134

Table 4.2: Genome assembly parameters

4.4 Genome Wide Association Study:

We've found 17 Leucine-related proteins which have a significant role on growth and muscle development on our predicted protein and genome and LRRN1_BOVIN had the highest match.

Serial	Protein IDt	Matched base	E-value	Score
1	LG11_BOVIN	277	0	578
2	LGR5_BOVIN	71	2.15E-41	147
3	LRC10_BOVIN	278	0	561
4	LRC14_BOVIN	493	0	989
5	LRC28_BOVIN	58	3.51E-33	117
6	LRC28_BOVIN	54	6.65E-31	120
7	LRC3B_BOVIN	227	8.38E-173	476
8	LRC43_BOVIN	157	2.67E-113	332
9	LRC51_BOVIN	143	1.00E-104	297
10	LRC59_BOVIN	139	7.59E-95	277
11	LRRC3_BOVIN	257	0	530
12	LRRC7_BOVIN	138	1.48E-87	283
13	LRRF2_BOVIN	57	1.12E-28	115
14	LRRN1_BOVIN	716	0	1481
15	LRRT1_BOVIN	491	0	1021
16	LSME1_BOVIN	42	1.49E-22	83.2
17	SHOC2_BOVIN	195	1.19E-134	390

Table 4.3: Matched protein list

We've created a Multiple Sequence Alignment from the matched protein and curated Leucine related protein. And finally a Phylogenetic from that MSA.

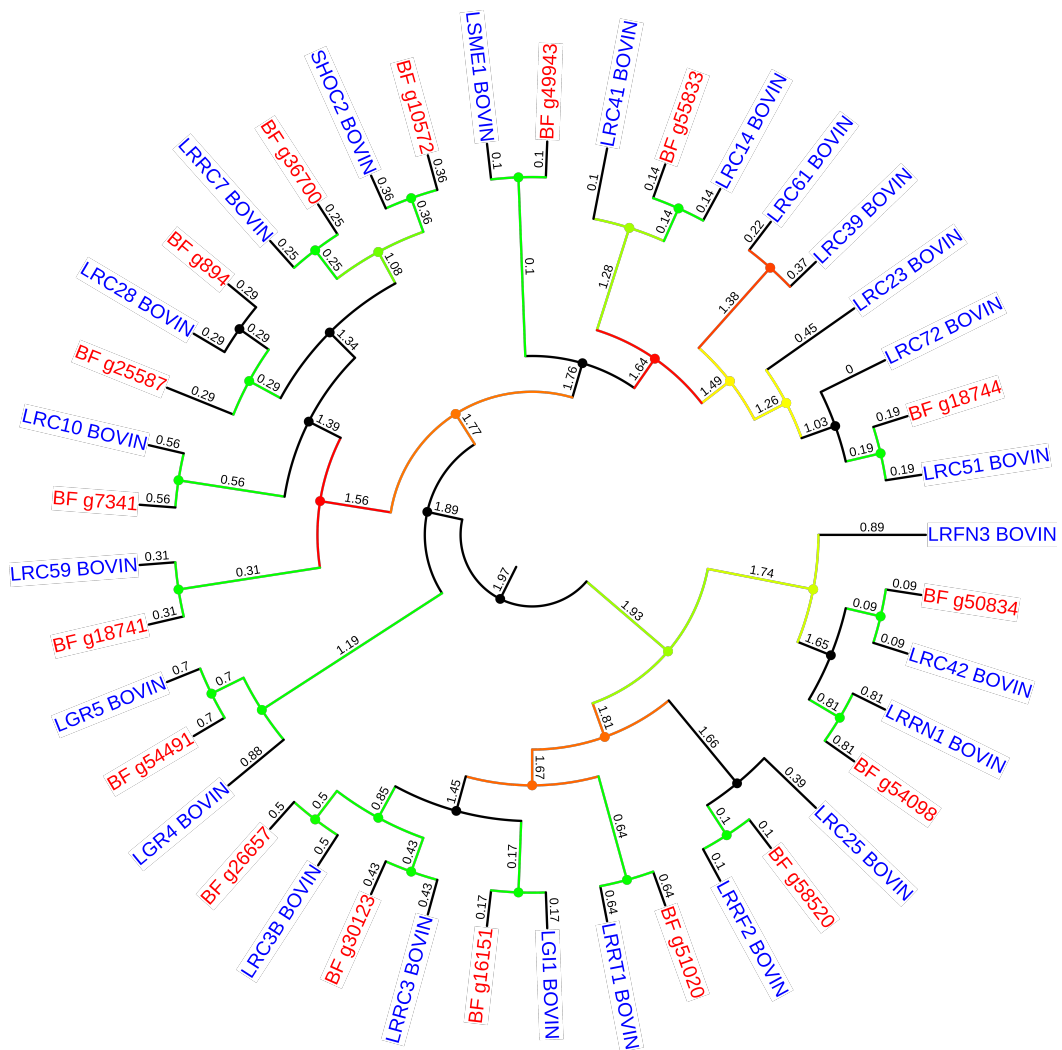


Figure 4.10: Phylogenetic tree of Leucine related protein of *Bos frontalis* and other Bovine species.

Here the visualization depicts every single Leucine-related protein of *Bos frontalis* that has been going through some mutation over the last millions of years. The green branches are more statistically correct than the yellow and red ones. And the branch length shows how long they are going through a mutation. Some mutations are good for a species, and some are not. If this mutation helps *Bos frontalis* for being better at growth and muscle development than other bovine species, then this information can be used for future reference and those who will want to work on cloning and hybridization.

We've also 3D modeled one special Leucine-related protein LRRN1_BOVIN with SwissModel and we found 8 sub-sections on the protein which are active and help to complete the task on the cell.

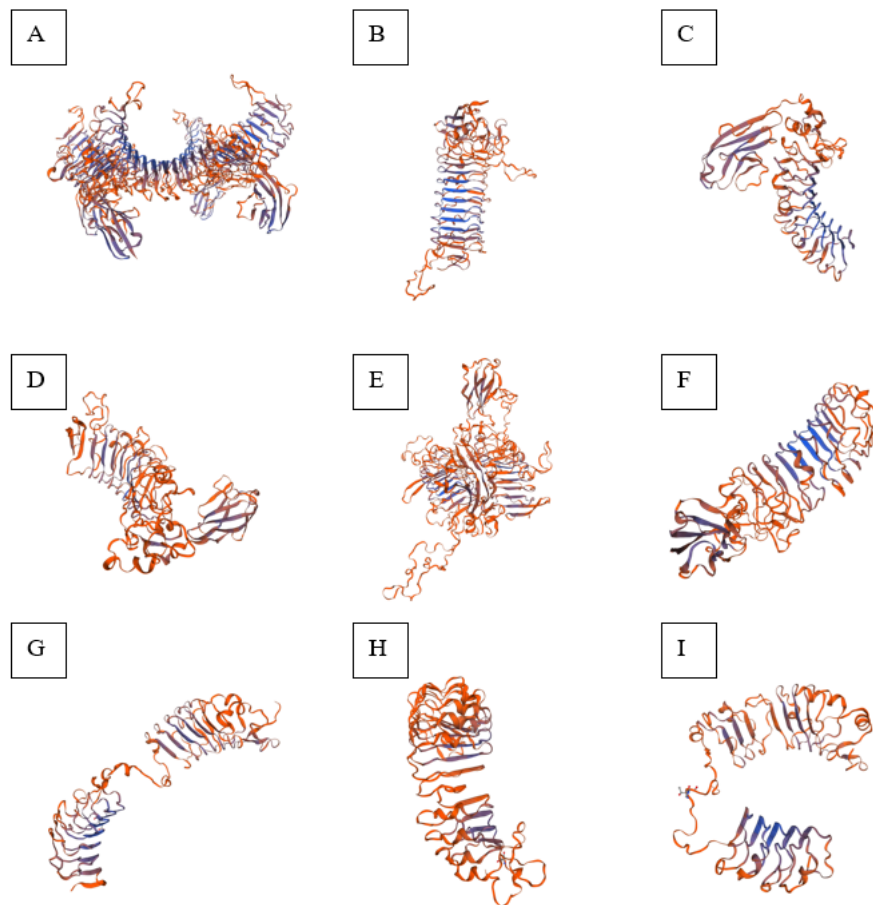


Figure 4.11: LRRN1_BOVIN on A and its sub-section in B to I.

Chapter 5

Conclusion & Future Work

In this thesis, after doing all the data analysis, filtering and simulation we have got 17 proteins who are related to Leucine and responsible for muscle growth in our sample of *Bos frontalis*. Our phylogenetic tree shows when and how the present leucine related protein had been mutated through the time. This tree also shows a comparison between leucine protein of *Bos frontalis* and other leucine proteins available in other nearby Bovine species.

The output of our thesis indicate that *Bos frontalis*'s genome can be used to produce faster growing hybrid species which will be able to solve protein deficiency problem of our country. We hope that, the output of our research will encourage live science researcher to produce gayal's hybrid which will have an economic impact in our society.

To conclude our thesis, we want to mention that our thesis is only focused on the proteins who are responsible for growth but our selected species have more unique characteristics which are economically important and narrowed down our analysis to single type of growth related protein to make a evolutionary comparison. For example, it has a broad variety of adaptations to hostile environments and it has low milk yield but the quality of milk is better than other nearby species. So, in the future we can do more research to find some more proteins related to these characteristics which can ultimately lead future researchers who will try to work on biological cloning and hybridization to improve these Bovine species.

Bibliography

- [1] G. A. Reid, *Molecular cloning: A laboratory manual, 2nd edn: By j. sambrook, ef fritsch and t. maniatitis, cold spring harbor laboratory press, 1989. (3 vols; 1659 pages) isbn 0 87969 309 6*, 1991.
- [2] W. R. Pearson, T. Wood, Z. Zhang, and W. Miller, "Comparison of dna sequences with protein sequences," *Genomics*, vol. 46, no. 1, pp. 24–36, 1997.
- [3] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. ii. error probabilities," *Genome research*, vol. 8, no. 3, pp. 186–194, 1998.
- [4] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, "Alignment of whole genomes," *Nucleic acids research*, vol. 27, no. 11, pp. 2369–2376, 1999.
- [5] G. Ma, H. Chang, S. Li, H. Chen, D. Ji, R. Geng, C. Chang, and Y. Li, "Phylogenetic relationships and status quo of colonies for gayal based on analysis of cytochrome b gene partial sequences," *Journal of genetics and genomics*, vol. 34, no. 5, pp. 413–419, 2007.
- [6] K. H. Taylor, R. S. Kramer, J. W. Davis, J. Guo, D. J. Duff, D. Xu, C. W. Caldwell, and H. Shi, "Ultradeep bisulfite sequencing analysis of dna methylation patterns in multiple gene promoters by 454 sequencing," *Cancer research*, vol. 67, no. 18, pp. 8511–8518, 2007.
- [7] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen, "Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning," *Nature*, vol. 452, no. 7184, pp. 215–219, 2008.
- [8] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature methods*, vol. 5, no. 1, pp. 16–18, 2008.
- [9] J. Shendure and H. Ji, "Next-generation dna sequencing," *Nature biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.
- [10] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, *et al.*, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome," *Science*, vol. 321, no. 5891, pp. 956–960, 2008.
- [11] C. P. Van Tassell, T. P. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard, "Snp discovery and allele frequency estimation by deep sequencing of reduced representation libraries," *Nature methods*, vol. 5, no. 3, pp. 247–252, 2008.

- [12] C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, *et al.*, “Personalized copy number and segmental duplication maps using next-generation sequencing,” *Nature genetics*, vol. 41, no. 10, p. 1061, 2009.
- [13] C. Auffray, Z. Chen, and L. Hood, “Systems medicine: The future of medical genomics and healthcare,” *Genome medicine*, vol. 1, no. 1, pp. 1–11, 2009.
- [14] P. Flicek and E. Birney, “Sense from sequence reads: Methods for alignment and assembly,” *Nature methods*, vol. 6, no. 11, S6–S12, 2009.
- [15] A. Guffanti, M. Iacono, P. Pelucchi, N. Kim, G. Soldà, L. J. Croft, R. J. Taft, E. Rizzi, M. Askarian-Amiri, R. J. Bonnal, *et al.*, “A transcriptional sketch of a primary human breast cancer by 454 deep sequencing,” *BMC genomics*, vol. 10, no. 1, pp. 1–17, 2009.
- [16] P. Medvedev, M. Stanciu, and M. Brudno, “Computational methods for discovering structural variation with next-generation sequencing,” *Nature methods*, vol. 6, no. 11s, S13, 2009.
- [17] D. S. Horner, G. Pavesi, T. Castrignano, P. D. De Meo, S. Liuni, M. Sammeth, E. Picardi, and G. Pesole, “Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing,” *Briefings in bioinformatics*, vol. 11, no. 2, pp. 181–197, 2010.
- [18] J. R. Miller, S. Koren, and G. Sutton, “Assembly algorithms for next-generation sequencing data,” *Genomics*, vol. 95, no. 6, pp. 315–327, 2010.
- [19] J. M. Otero, W. Vongsangnak, M. A. Asadollahi, R. Olivares-Hernandes, J. Maury, L. Farinelli, L. Barlocher, M. Østerås, M. Schalk, A. Clark, *et al.*, “Whole genome sequencing of *saccharomyces cerevisiae*: From genotype to phenotype for improved metabolic engineering applications,” *BMC genomics*, vol. 11, no. 1, pp. 1–17, 2010.
- [20] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, *et al.*, “A human gut microbial gene catalogue established by metagenomic sequencing,” *nature*, vol. 464, no. 7285, pp. 59–65, 2010.
- [21] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean, “De novo assembly and genotyping of variants using colored de bruijn graphs,” *Nature genetics*, vol. 44, no. 2, pp. 226–232, 2012.
- [22] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, *et al.*, “Soapdenovo2: An empirically improved memory-efficient short-read de novo assembler,” *Gigascience*, vol. 1, no. 1, pp. 2047–217X, 2012.
- [23] F. Nadalin, F. Vezzi, and A. Policriti, “Gapfiller: A de novo assembly approach to fill the gap within paired reads,” *BMC bioinformatics*, vol. 13, no. 14, pp. 1–16, 2012.
- [24] M. Baig, B. Mitra, K. Qu, M.-s. Peng, I. Ahmed, Y.-W. Miao, L.-s. Zan, and Y.-p. Zhang, “Mitochondrial dna diversity and origin of *bos frontalis*,” *Current Science*, pp. 115–120, 2013.
- [25] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: A flexible trimmer for illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

- [26] R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, *et al.*, “Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads,” *Genome research*, vol. 24, no. 8, pp. 1384–1395, 2014.
- [27] L. V. Singh, A. Sharma, N. Kumari, N. Kaur, S. Jayakumar, S. Dixit, N. Gupta, and S. Gupta, “Comparative sequence analysis in the exon 5 of growth hormone gene in the various livestock species of india,” *Animal biotechnology*, vol. 25, no. 1, pp. 69–72, 2014.
- [28] F. A. R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, “Busco: Assessing genome assembly and annotation completeness with single-copy orthologs,” *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, 2015.
- [29] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph,” *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, Jan. 2015, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv033. eprint: <https://academic.oup.com/bioinformatics/article-pdf/31/10/1674/17085710/btv033.pdf>. [Online]. Available: <https://doi.org/10.1093/bioinformatics/btv033>.
- [30] R. R. Wick, M. B. Schultz, J. Zobel, and K. E. Holt, “Bandage: Interactive visualization of de novo genome assemblies,” *Bioinformatics*, vol. 31, no. 20, pp. 3350–3352, 2015.
- [31] C. Mei, H. Wang, W. Zhu, H. Wang, G. Cheng, K. Qu, X. Guang, A. Li, C. Zhao, W. Yang, *et al.*, “Whole-genome sequencing of the endangered bovine species gayal (*bos frontalis*) provides new insights into its genetic features,” *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.
- [32] S. D. Jackman, B. P. Vandervalk, H. Mohamadi, J. Chu, S. Yeo, S. A. Hammond, G. Jahesh, H. Khan, L. Coombe, R. L. Warren, *et al.*, “Abyss 2.0: Resource-efficient assembly of large genomes using a bloom filter,” *Genome research*, vol. 27, no. 5, pp. 768–777, 2017.
- [33] M.-S. Wang, Y. Zeng, X. Wang, W.-H. Nie, J.-H. Wang, W.-T. Su, N. O. Otecko, Z.-J. Xiong, S. Wang, K.-X. Qu, *et al.*, “Draft genome of the gayal, *bos frontalis*,” *GigaScience*, vol. 6, no. 11, gix094, 2017.
- [34] Q. Ren, Y. Liu, X. Xie, B. Yan, K. Zhang, Y. Yang, and Q. Qiu, “Complete mitochondrial genome of bovine species gayal (*bos frontalis*),” *Conservation Genetics Resources*, vol. 10, no. 4, pp. 889–891, 2018.
- [35] S. Mukherjee, Z. Cai, A. Mukherjee, I. Longkumer, M. Mech, K. Vupru, K. Khate, C. Rajkhowa, A. Mitra, B. Guldbrandtsen, *et al.*, “Whole genome sequence and de novo assembly revealed genomic architecture of indian mithun (*bos frontalis*),” *BMC genomics*, vol. 20, no. 1, pp. 1–12, 2019.
- [36] V. R. Prabhu, M. S. Arjun, K. Bhavana, R. Kamalakkannan, and M. Nagarajan, “Complete mitochondrial genome of indian mithun, *bos frontalis* and its phylogenetic implications,” *Molecular biology reports*, vol. 46, no. 2, pp. 2561–2566, 2019.
- [37] [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

- [38] [Online]. Available: <https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/fastqc.html>.
- [39] [Online]. Available: <http://quast.sourceforge.net/>.
- [40] U. C. B. I. I. R. S. I. of Bioinformatics, *Uniprot consortium*. [Online]. Available: <https://www.uniprot.org/>.
- [41] *Blast: Basic local alignment search tool*. [Online]. Available: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.
- [42] *Multiple alignment program for amino acid or nucleotide sequences*. [Online]. Available: <https://mafft.cbrc.jp/alignment/software/>.