Introduction to Bioinformatics

Read mapping and genome assembly project description

Department of Bioinformatics, IBB, University of Tehran

Fall 2023

Presenter: Fereshteh Noroozi

Read Mapping and Genome Assembly: Decoding Genetic Blueprints

1. Fundamental processes in genomics

2. Read mapping:

   Aligning short DNA sequences to a reference genome

   Identifying locations and variations in genomic DNA

3. Deciphering the genetic blueprint

4. Understanding gene relationships

5. Genome assembly

   Piecing together short DNA sequences

   Reconstructing the complete genome

   Critical in genome sequencing projects

6. Comprehending an organism's genetic architecture

7. Insights into potential functions of genes

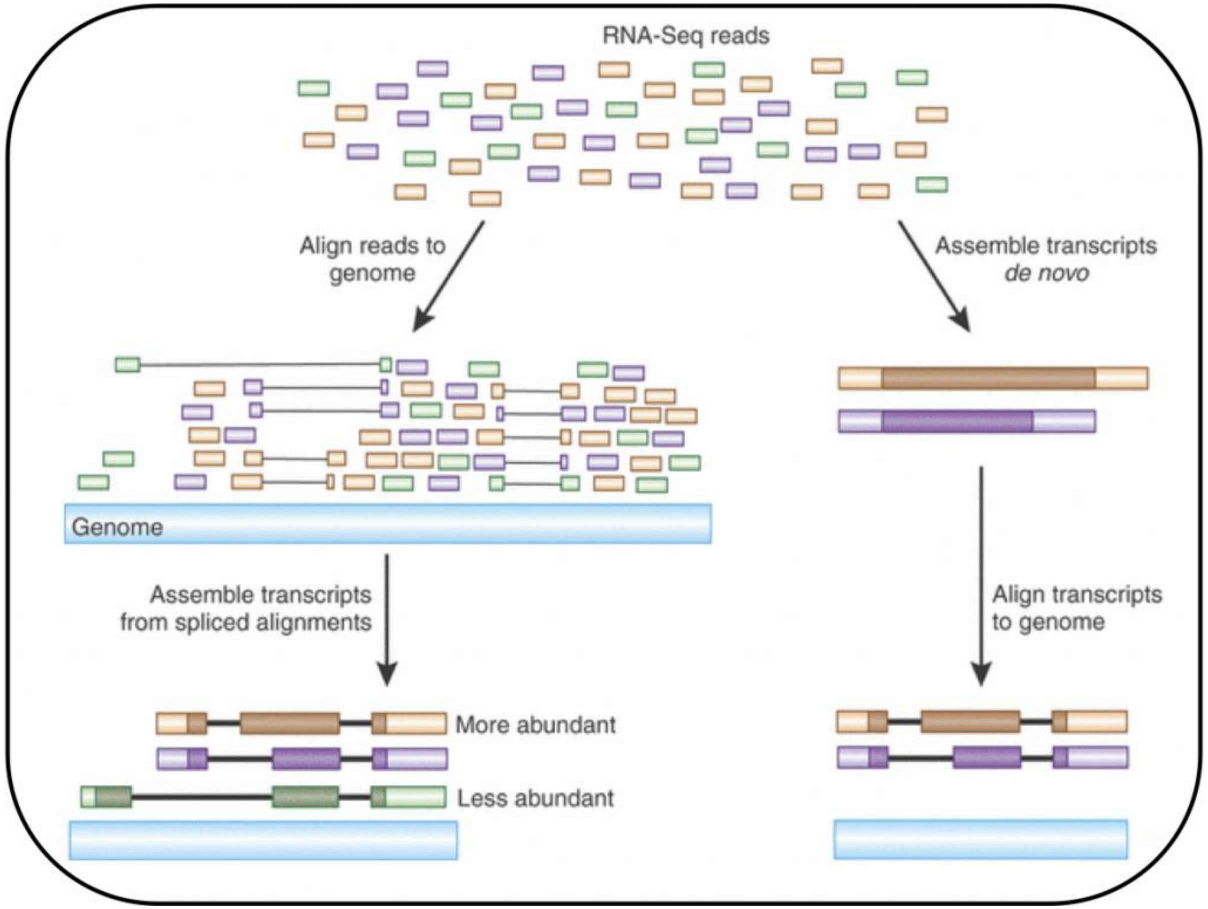
8. Insights into evolutionary processes

9. Pivotal roles in advancing genomic research

10. Exploring genetic variations

11. Studying diseases at the molecular level

12. Unraveling complexities of living organisms


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Next-Generation Sequencing (NGS) Overview

Next-generation sequencing (NGS):

Is a high-throughput DNA sequencing technology, enabling rapid and cost-effective analysis of genetic material for various applications in genomics. It has largely replaced traditional sequencing methods.

Short-Read Sequencing:

Definition: Techniques like Illumina sequencing.

Characteristics: Produces brief DNA fragments.

Applications: Well-suited for high-coverage, cost-effective whole-genome sequencing.

usually 50 to 300 bases

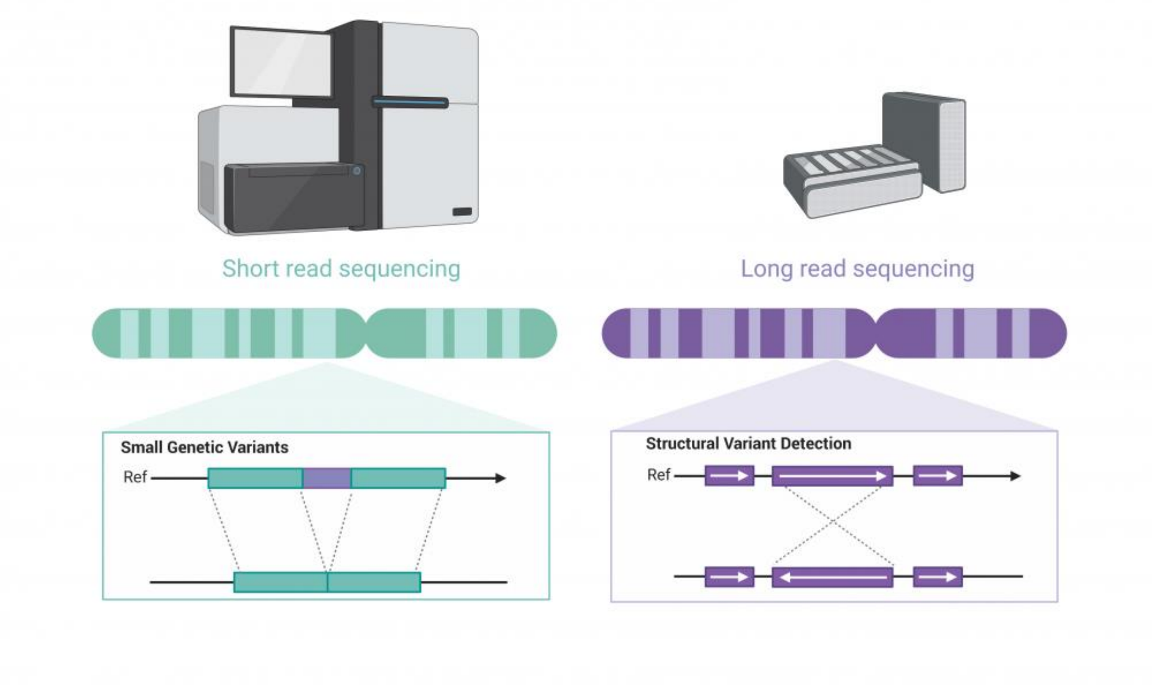
Long-Read Sequencing:

Definition: Technologies such as PacBio and Oxford Nanopore sequencing.

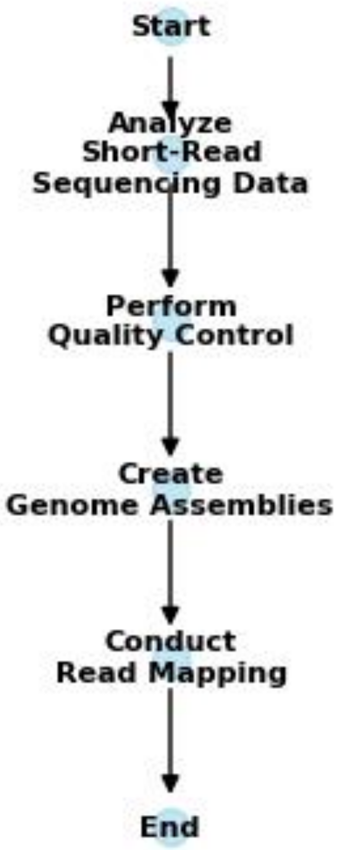
Characteristics: Yields more extended DNA sequences.

Applications: Enables comprehensive genomic analysis, structural variant detection.

Up to kilo base pair



Project steps



Start

Analyze Short-Read Sequencing Data

Perform Quality Control

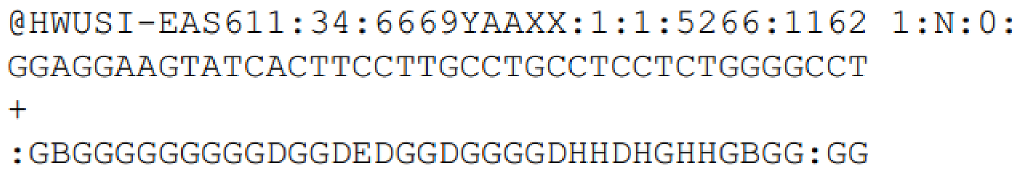
Create Genome Assemblies

Conduct Read Mapping

End

Part A - Downloading E. Coli WGS data, preliminary analyses, and quality controls (Q1)

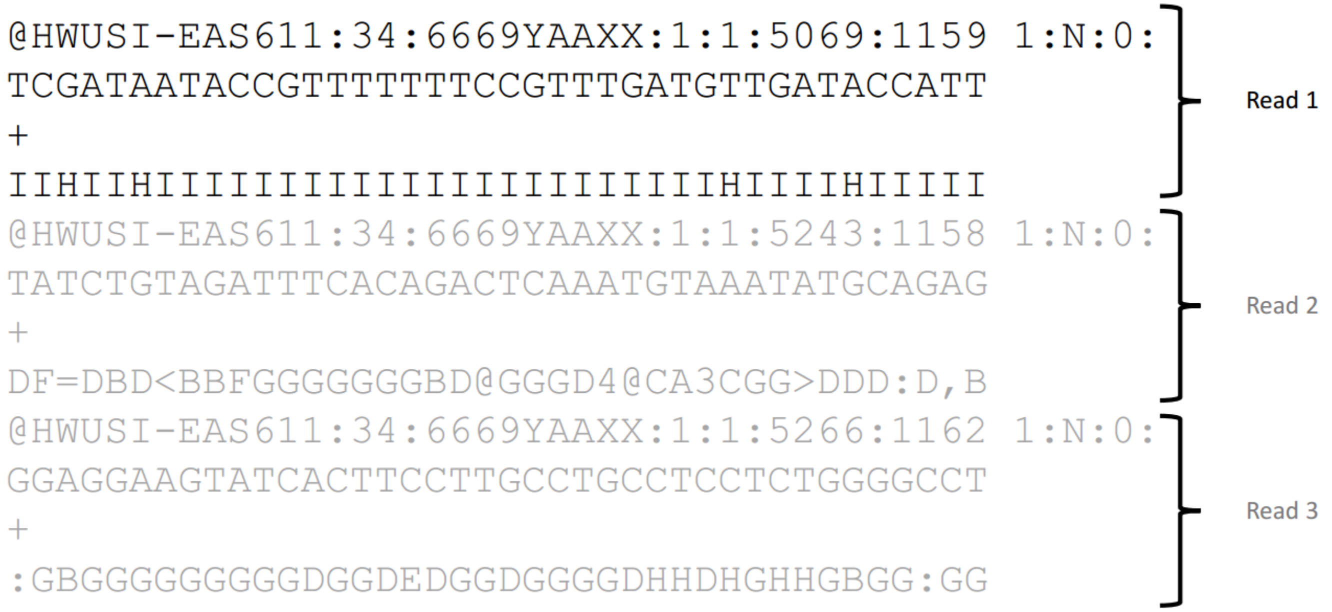
- The Sequence Read Archive (SRA) is a centralized NCBI database storing raw high-throughput sequencing data from diverse genomic studies
- The SRA Toolkit, developed by NCBI, consists of command-line utilities designed for efficient retrieval and manipulation of raw sequencing data stored in the Sequence Read
- **Steps:**
- Install SRA Toolkit in Linux or windows
- Use fastq-dump to download SRR8185316
- Whole Genome Sequencing (WGS) of Escherichia coli (E. coli) refers to the process of determining the complete DNA sequence of the entire genome of the bacterium E. coli.
- Avoid direct downloads from SRA as the fastq file may contain numerous question marks
- **Fastq file format:**



```
@HWUSI-EAS611:34:6669YAAXX:1:1:5266:1162 1:N:0:
GGAGGAAGTATCACTTCCTTGCCTGCCTCCTCTGGGGCCT
+
:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG
```

Part A - Downloading E. Coli WGS data, preliminary analyses, and quality controls (Q2)

- **I. How many reads are in the fastq file?**
- **II. Print the identifier, quality, and sequence of the first read of the fastq file.**
- In a FASTQ file, a "read" typically refers to a DNA or RNA sequence obtained from a high-throughput sequencing experiment. Each read is represented by four lines in the FASTQ format



```
@HWUSI-EAS611:34:6669YAAXX:1:1:5069:1159 1:N:0:
TCGATAATACCGTTTTTTTCCGTTTGATGTTGATACCATT
+
IIHIIHIIIIIIIIIIIIIIIIIIIIIIIHIIIIHIIIII
@HWUSI-EAS611:34:6669YAAXX:1:1:5243:1158 1:N:0:
TATCTGTAGATTTCACAGACTCAAATGTAAATATGCAGAG
+
DF=DBD<BBFGGGGGGGBD@GGGD4@CA3CGG>DDD:D,B
@HWUSI-EAS611:34:6669YAAXX:1:1:5266:1162 1:N:0:
GGAGGAAGTATCACTTCCTTGCCTGCCTCCTCTGGGGCCT
+
:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG
```

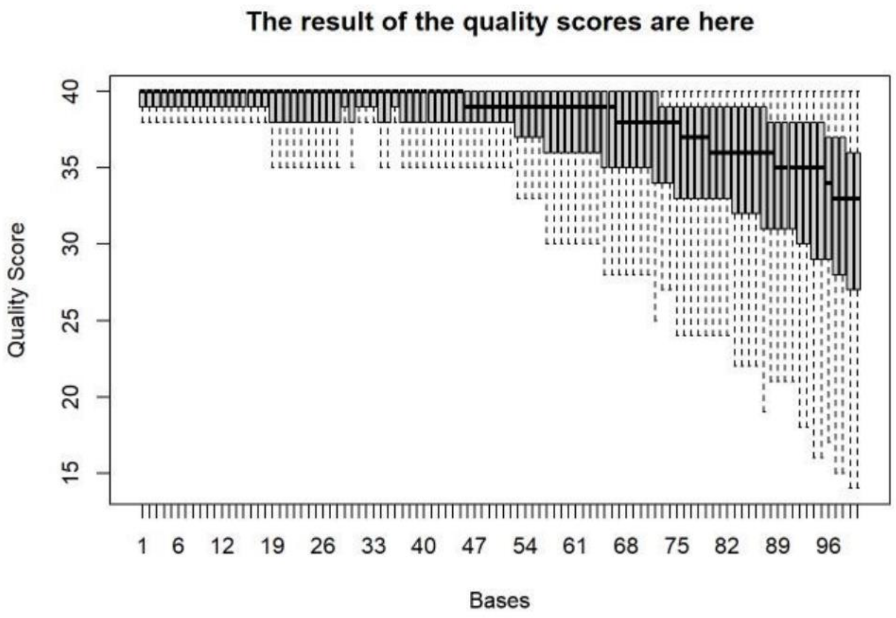
Part A - Downloading E. Coli WGS data, preliminary analyses, and quality controls (Q2)

- **III. How many times does the TTAAATGGAA subsequence appear in the file?**
- **IV. Extract the first 1000 sequences of the fastq files (4000 lines).**

*** A subsequence is a portion of a longer sequence that retains the order of the nucleotides but allows for gaps or additional nucleotides between them. For example, in the DNA sequence "ATGCTA," the subsequences "ATG" and "CTA" are valid. The subsequence "AGT" is not valid because the order of nucleotides is not maintained.**

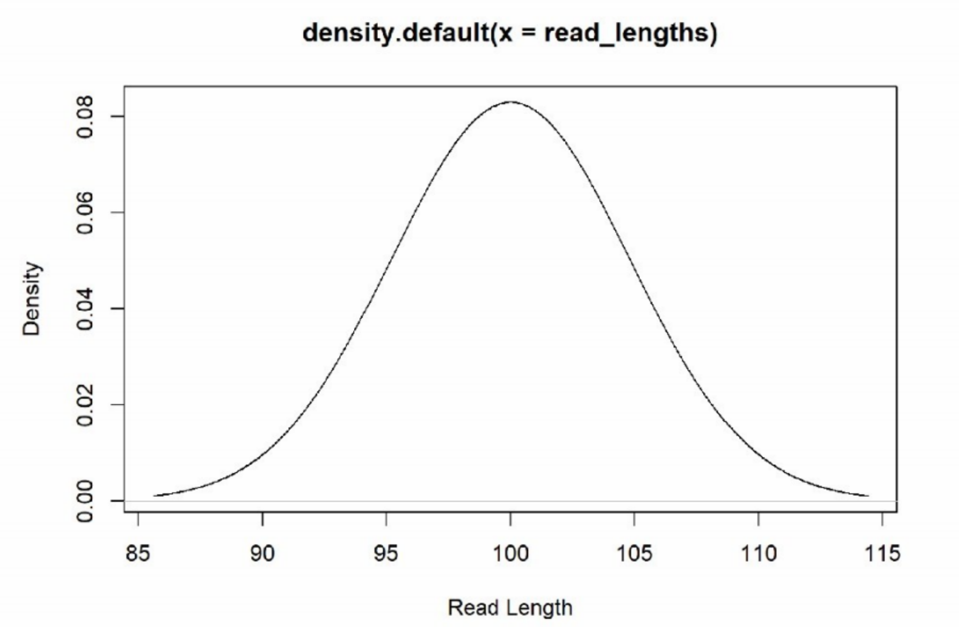
Part A - Downloading E. Coli WGS data, preliminary analyses, and quality controls (Q2)

- **V. Plot the quality of the reads in the fastq file using a box plot.**
- **In the context of DNA sequencing, the term "score" typically refers to the quality score or Phred score. Quality scores represent the confidence or accuracy of each base call in a DNA sequence. The Phred score is a logarithmic scale used to quantify the probability of an incorrect base call.**
- Processes are limited to the first 100,000 sequences to avoid potential errors that could arise from handling the entire dataset



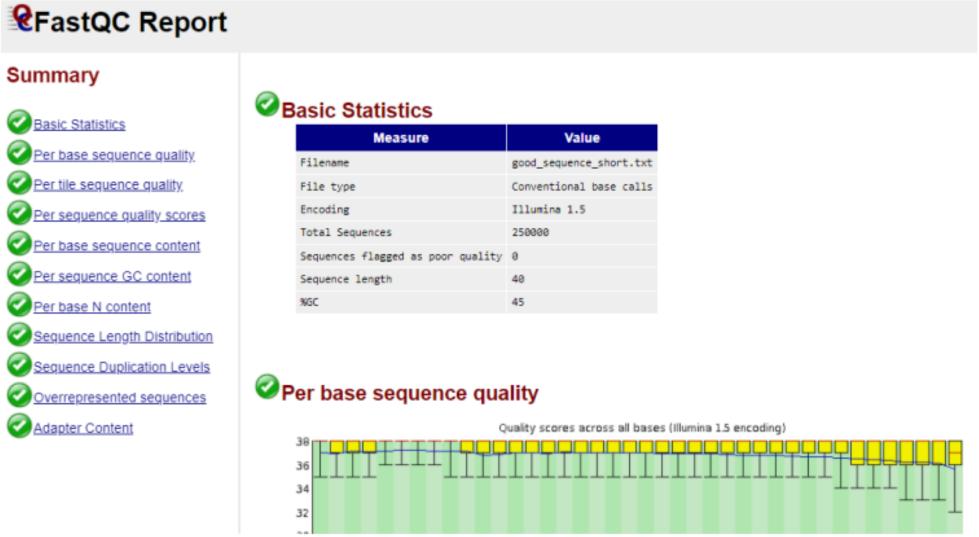
Part A - Downloading E. Coli WGS data, preliminary analyses, and quality controls (Q2)

- **VI. Show the distribution of read lengths using a density plot.**
- A density plot generated from a FASTQ file would illustrate the probability distribution of read lengths. Each read in a FASTQ file has an associated length, representing the number of nucleotides in that sequence. The density plot allows you to visualize the variation and concentration of read lengths within the sequencing data. Areas of higher density indicate regions where specific read lengths are more prevalent, providing valuable insights into the composition and characteristics of the sequence library. This visualization is useful for assessing the quality of sequencing data and understanding the distribution of sequence lengths in the dataset.



Part A - Downloading E. Coli WGS data, preliminary analyses, and quality controls (Q3)

- **3. Perform quality control of the reads using FastQC and interpret the results.**
- Install SRA FastQC in Linux
- Both the command-line function as well as the FastQC graphical interface had html outputs



FastQC Report

Summary

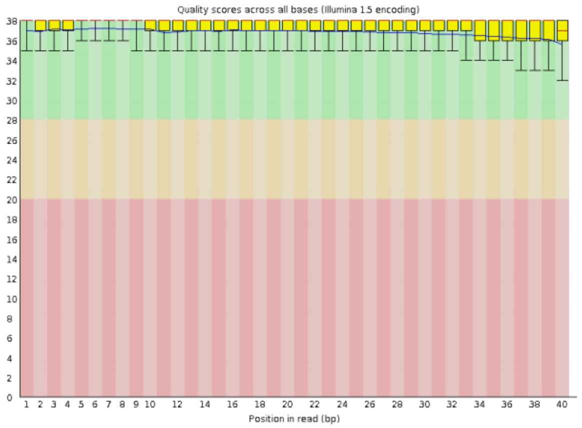
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

✓ Basic Statistics

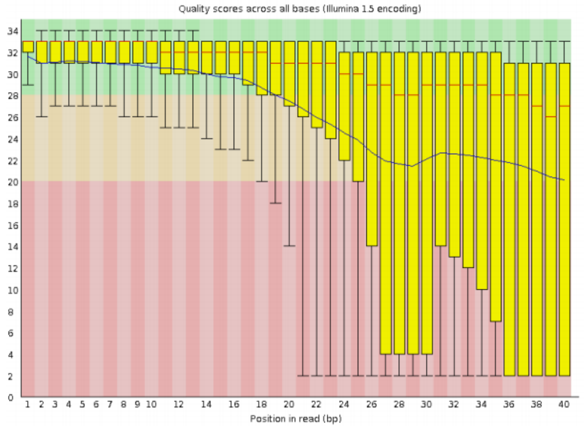
Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

✓ Per base sequence quality

Good Illumina Data



Bad Illumina Data



Part B - De novo genome assembly (Q1)

- **Run SPAdes to generate draft genome assemblies from short reads.**
- Install SPAdes in linux
- The aim of using SPAdes or similar genome assembly tools is to create draft genome assemblies from short reads, assembling them into contigs and scaffolds. This process yields insights into genomic structure and function.
- **List of Key Output Files from Genome Assembly Analysis:**
- Contigs and scaffolds.
- Assembly statistics.
- Assembly graphs.
- Annotation files.

Part B - De novo genome assembly (Q2)

- **Assess the quality of the draft genome assembly using Quast and compare it to the reference genome**
- Install Quast in linux
- Download reference genome from : https://www.ncbi.nlm.nih.gov/nuccore/NZ_MT158477.1?report=fasta
- The goal is to evaluate the quality of a draft genome assembly using Quast, assessing metrics like contig length and coverage, and comparing the results to a reference genome for accuracy validation. This step ensures the reliability of the assembled genome for subsequent genomic analyses
- **List of Key Output Files from Genome Assembly Analysis:**
- Quast report.
- Contig metrics file.
- Aligned sequences file.
- Conserved genes file.
- Misassemblies file.
- Gaps file.

Part C - Read mapping (Q1)

Map the Illumina short-read data to the reference genome using BWA.

Install BWA in Linux

The goal is to create a mapping or alignment file that can be used for downstream analyses, such as variant calling, identifying genomic variations, and understanding the genetic landscape of the sequenced sample in comparison to the reference genome.

List of Key Output Files from Genome Assembly Analysis:

The resulting SAM file can be further processed, converted to BAM format, and sorted using tools such as samtools.

Part C - Read mapping (Q2)

- **Print the head of the obtained SAM file from previous question. Explain what you see for the first hit (you can do this step either in Linux or R).**

SAM files are commonly used to store the results of sequence alignment performed by bioinformatics tools such as Bowtie, BWA, or STAR. The header section provides important metadata about the reference sequences and the alignment process, which can be used for downstream analysis and interpretation of the alignment results.

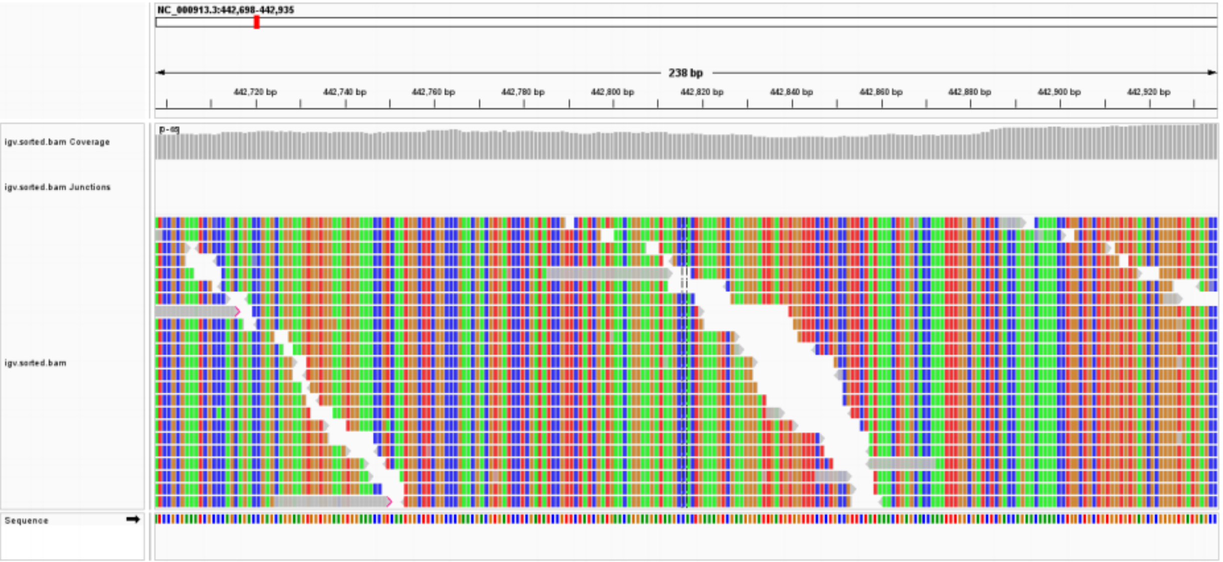
Part C - Read mapping (Q3)

- **Convert the SAM file to an indexed BAM file. Hint: use samtools view, samtools sort, samtools index.**
- Install BWA in samtools
- Use samtools view to convert a SAM file to a binary BAM file.
- Use samtools sort to sort the BAM file.
- Use samtools index to create an index file for the sorted BAM file.
- The resulting sorted and indexed BAM file is ready for downstream analysis and visualization of sequence alignments.

Part C - Read mapping (Q4)

- **Use the Integrative Genomics Viewer (IGV) to visualize the mapped reads in a 200-b genomic region of your choice. Select the reference genome fasta and GTF file (GTF is optional).**

- **Example:**



Part C - Read mapping (Q5-Q7)

- **Determine the percentage of short reads that are mapped to the reference genome. Hint: use samtools flagstat.**
- **Get the read depth for the sorted BAM file at all positions of the reference genome and report the mean of all reads. Hint: use samtools depth.**

The term "CIGAR" (Compact Idiosyncratic Gapped Alignment Report) is not applicable to FASTQ files; instead, it is a component of the SAM (Sequence Alignment/Map) format, which is commonly used to represent the results of sequence alignment in genomics.

In the SAM format, the CIGAR string is a compact representation of the alignment details between a read and the reference genome. It describes how each base in the read aligns to the reference, including information about matches, mismatches, insertions, and deletions. The CIGAR string consists of a series of operators and lengths, providing a concise summary of the alignment pattern.

For example, a CIGAR string like "50M2I30M" indicates that the first 50 bases of the read match the reference, followed by an insertion of 2 bases, and then another 30 bases that match the reference.