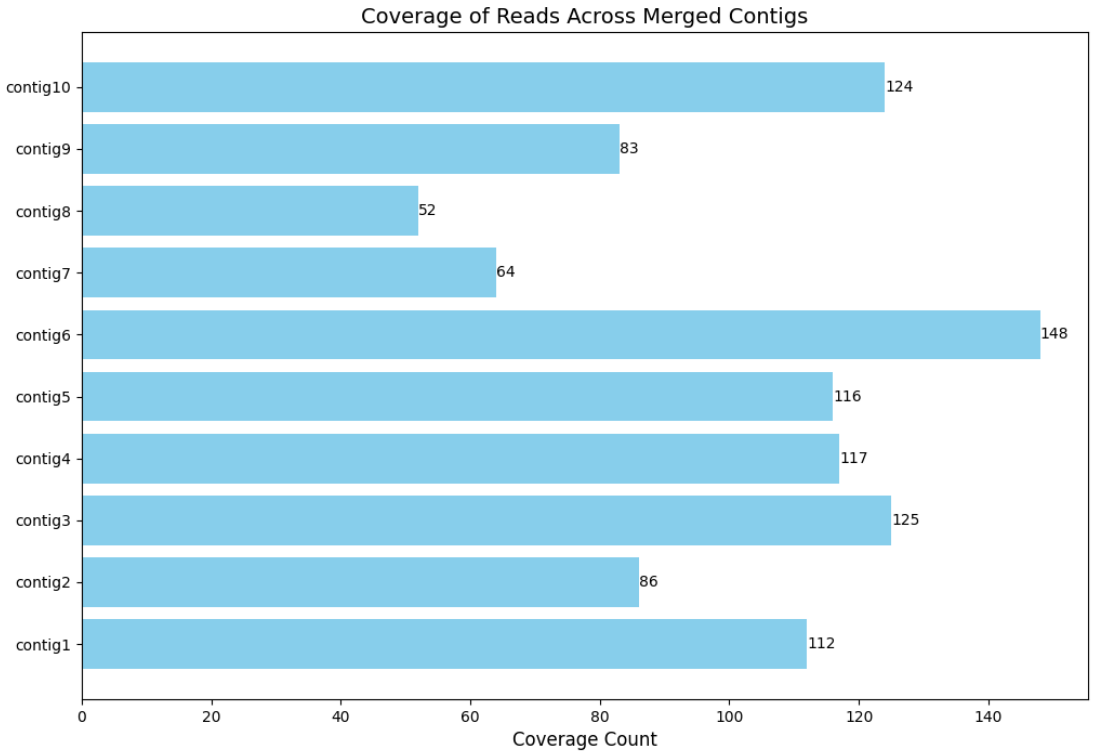
1) **Evaluate the distribution of the reads across each sequence (30 points)**

a. Create a visualization of your choice that shows coverage of reads across each of your sequences (10 points)



I choose horizontal bar graph for visualization because it a good choice for this dataset and suitable for comparing coverage values across multiple contigs.

The read coverage count horizontal bar graph displays the coverage count for each contig, showing the number of reads mapped to each one. In x-axis contigs are labeled as Contig_1 to Contig_10. In y-axis coverage is presented, with numeric labels on the bars, for easy comparison of the read distribution.

b. Based on the visualization, is the sequencing method biased? Explain (10 points)

Sequencing bias is a common issue in DNA sequencing that can negatively impact genome assembly and various downstream applications. This bias arises when reads are unevenly distributed across the genome or when the quality of reads varies depending on the sequence. The consequences of sequencing bias can be significant: under covered regions may lead to missed single nucleotide polymorphisms (SNPs) or result in shorter contigs, and important

loci may be absent from the assembled genomes due to uneven coverage. Additionally, important single nucleotide variants can be lost because of this coverage bias.

To mitigate sequencing bias, several strategies can be employed. Increasing the overall sequencing coverage can help to ensure that low-coverage areas are adequately represented. Alternatively, using different sequencing technologies, such as PacBio or Nanopore, may reduce the biases associated with platforms like Illumina. Moreover, understanding the GC content bias profile of the sequencing process can assist in managing these biases during data analysis.

In this visualization analysis, some contigs exhibit notably higher coverage than others; for instance, Contig_7 has 148 reads, while Contig_8 only has 52. Ideally, a well-designed sequencing approach would distribute reads more evenly across contigs. A major ambiguity in coverage values may indicate that the sequencing method may favor certain sequences over others. Typically, longer contigs tend to receive more reads than shorter ones, but this is not consistently observed in this data. For example, Contig_3, which is 154 bases long, has 125 reads, whereas Contig_7, at only 82 bases, has an unexpectedly high coverage of 148 reads. This further implies a potential bias in the sequencing process.

Consider the following about your code (20 points)

1) Change the overlap parameter (k) in your code:

a. At what point does the program output change when you decrease k? (5 points)

Reducing the k -value to 4 resulted in 157 contigs, but since the algorithm doesn't check for overlaps after merging, it missed chances to resolve these overlaps. This leads to more fragmented contigs and can affect the overall accuracy of the assembly, as smaller k -mers can create confusion in the graph and make it harder to reconstruct the genome properly.

i. Consider the assumptions you made in your algorithm – what's the issue? (5 points)

Reducing the k -value to 4 in this contig assembly algorithm negatively impacts the merging process and the quality of the resulting contigs. While smaller k -mers save storage space and increase the chances of overlaps, they also create more vertices in the graph, leading to confusion and path ambiguities. This complexity makes it harder to reconstruct the genome, as many vertices might connect to a single k -mer. As a result, there's a higher risk of incorrect merges, which can produce fragmented and inaccurate contigs.

b. At what point does the program output change when you increase k? (5 points)

Using a k -value of 20 in this contig assembly algorithm results 75 contigs. while using a k-value of 10 or greater yields only 10 contigs. This suggests that higher k-values lead to more specific k-mers, which can merge reads more accurately and produce a larger number of contigs. This higher k-value helps to reduce ambiguity by allowing for more specific overlaps between reads, which can improve the quality of the resulting contigs. With fewer, more reliable overlaps, the assembly process is likely to produce more accurate and contiguous sequences. However, the trade-off is that it may require more memory and storage due to the larger k -mer size. Overall, this k-value strikes a balance between reducing fragmentation and maintaining assembly quality.

c. What is the relationship between how high k can go and sequencing coverage? (5 points)

The relationship between the k -value and sequencing coverage is important for genome assembly. When the k -value is high, the k -mers become more specific, making it easier to merge reads accurately and resulting in better quality contigs. However, high k -values also need sufficient sequencing coverage to ensure there are enough overlapping k-mers for accurate assembly. If the coverage is low, important connections may be missed, leading to fragmented contigs. On the other hand, lower k -values increase the chance of overlaps, which can help capture more connections, especially in areas with low coverage. But they also introduce more ambiguity, as different sequences may share the same k-mer, complicating the assembly process. Therefore, finding a balance between the k-value and sequencing coverage is essential for achieving the best results in genome assembly.

Consider sequencing as a whole (30 points)

1) When we look at paired end reads, we gain benefit from increasing L – the distance between the paired ends.

When considering paired end reads, increasing the distance L between the paired ends offers several advantages. The primary benefit is enhanced mapping accuracy, as greater distances between reads can provide more context for aligning them to the reference genome. This is particularly important in regions of the genome that are complex or repetitive, where short paired-end reads may not provide enough information to establish a clear alignment.

With a larger L , paired-end reads can span across structural variations, such as insertions or deletions, allowing for more effective detection and characterization of these variations. Additionally, longer distances help in resolving ambiguities in read alignment, as they can link distant genomic regions, providing insights into the overall structure and organization of the genome.

Moreover, increasing L improves de novo assembly efforts by offering clearer insights into the relationships between different contigs. When reads from two ends of a fragment are spaced further apart, they can indicate the correct order and orientation of contigs, facilitating the assembly process of complex genomes. Overall, maximizing L in paired end reads enhances genomic analysis and assembly, leading to more accurate and comprehensive insights into the genome being studied.

a. Explain how mate-pair reads enable us to get a higher L than paired end reads (5 points)

Paired-end reads are generated from both ends of the same contiguous DNA molecule, with a user-definable distance between the ends, typically ranging from 100 to 700 bp. In contrast, mate-pair reads refer to pairs that are separated by a greater distance, usually between 2 and 5 kb. This longer separation is achieved by circularizing the DNA, bringing the ends of the template molecules together, and then fragmenting the circularized DNA. While both paired-end and mate-pair libraries yield similar data—paired reads from the same template separated by a known distance—their construction methods and insert sizes differentiate them.

Mate-pair reads allow for capturing longer distances between paired ends compared to standard paired-end reads. In paired-end sequencing, reads typically come from shorter fragments, around 400 bp. Conversely, mate-pair sequencing utilizes larger DNA fragments, often several kilobases long. This process enhances the distance. L between the paired reads, which aids in resolving complex genomic regions and provides better context for assembly and detecting structural variants.

In Illumina mate-paired libraries, constructed from 3- to 5-kb DNA fragments, a biotin molecule enriches fragments at the circularization junction. However, many fragments may lack this junction, resulting in shorter paired end reads. To minimize the risk of reads spanning the junction—an issue that complicates mapping and de novo assembly—Illumina recommends a maximum read length of 36 bases. Specialized software, such as Novoalign, is necessary to effectively handle junction reads. By selecting a library size range of 400–600 bp, which is larger than that of typical paired-end libraries, Illumina aims to reduce the occurrence of problematic junction reads, ultimately enhancing the quality of the assembly.

c. If your average sequencing fragment size is 400bp and your mate-pair selection library size is 3000bp and your read length is 80bp estimate your average distance L for paired end AND mate pair sequencing. Show your work for how you got both numbers (15 points)

For the paired end reads:

Distance L = Fragment size – read length = 400bp – 80bp = 320 bp

For the mate pair reads:

Distance L = Mate-pair library size – read length = 3000bp – 80bp = 2920 bp

2) Why is it better to generate reads from the same sequence (pac-bio) than generating the same amount of reads from the replicate of that sequence? (5 points)

Generating reads from the same sequence using PacBio technology is advantageous because it provides longer, continuous reads that capture more of the genomic context and structural information than multiple shorter reads from replicate sequences. Longer reads help to span repetitive regions and structural variations, allowing for more accurate assembly and variant detection. In contrast, replicating shorter reads may lead to gaps in coverage and an inability to resolve complex genomic regions, resulting in fragmented assemblies and missed variations. Thus, obtaining long reads from the same sequence enhances the quality and completeness of genomic analyses.

3) Choose a type of human genomic variation besides single nucleotide polymorphism(SNP). Explain how long reads help detect that type of variation. (5 points)

One type of human genomic variation besides single nucleotide polymorphisms (SNPs) is structural variation (SV). Genomic structural variations (SVs) are genetic alterations that result in amplifications, losses, inversions, and translocations of segments of DNA greater than 50 bp. SVs are a normal part of genomic variation but can also cause disorders. Standard detection methods include chromosome banding, fluorescent in situ hybridization (FISH), and array comparative genome hybridization (aCGH) that is very useful to detect copy number variations (CNVs) but cannot detect copy-neutral SVs (inversions, balanced translocations). Recent methods include employment of NGS to identify SVs, which are not detectable by cytogenetic methods.

Long reads, such as those produced by PacBio or Oxford Nanopore technologies, are particularly useful for detecting structural variations because they can span larger genomic regions, providing a complete context around the variation. This capability allows long reads to traverse repetitive sequences that often complicate the assembly of shorter reads, enabling the identification of structural variants that might otherwise go undetected. Additionally, long reads can accurately resolve the breakpoints of SVs, leading to more precise characterization of these variations and their potential impact on gene function and disease.