MODULE – 01

Name: Aruna Kumar

BINF 6400 - Genomics in Bioinformatics

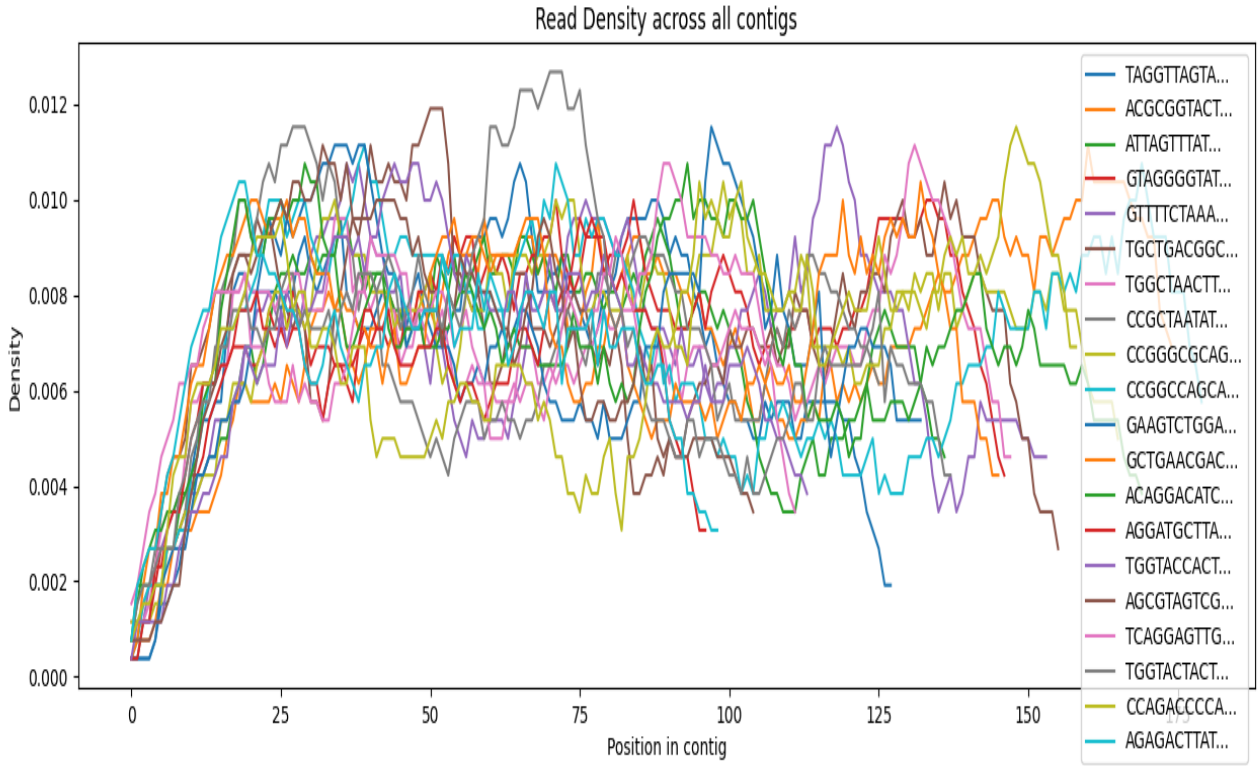
NUID: 0026413844

College of Science, Northeastern University

February 01, 2024

Evaluate the distribution of the reads across each sequence:

Figure 01:



Above plot illustrates the read density across multiple contigs. To know the potential biases in the sequencing method we need to consider many factors, they are:

- **Fluctuation in density:** Normally for unbiased, we might expect a more uniform density across all position in contigs, on the other hand fluctuations are normal to some degree because of the natural variations in the genomic sequence.

  The above plot shows fluctuations in read density across the length of the contigs that is non-uniform density. The fluctuations seen here suggest that some regions are covered by more reads that is peaks while others by fewer (troughs), it might point to a sequencing bias.

- **Early Uniformity and Later Divergence:**

  This plot shows that all reads start relatively uniform across all contigs but spread later. This pattern suggests that uneven allocation across the contigs, this might mean the method used to read the DNA tends to focus more on the beginning parts and not as much on the rest.
- **Variability Among Contigs**: If some contigs consistently show higher peaks or lower troughs than others, this might suggest a bias toward or against specific parts within the sequencing or assembly process.
- **Technical Factors**: Some technical factors inherent or library preparation method can introduce biases. And some platform struggle with GC-rich or GC-poor regions these factors could give uneven coverage.

Consider the following about your code.

Change the overlap parameter (k) in your code.

a. **At what point does the program output change when you decrease k?**

- When I reduced K to 7 the program output changed.
- This is because by reducing k value we are essentially allowing for shorter overlaps to count as valid when assembling the contigs. This means that more reads may be combined to form a contig, which can potentially lead to longer contigs being formed.
- This is why I see some contigs getting longer when we reduced k to 7.

Consider the assumptions you made in your algorithm – what's the issue?

Assumptions and issues:

- A key assumption in this algorithm is that a certain length of overlap (k) is needed to confidently join two reads.
- If k is too short, you risk incorrectly joining reads due to random chance overlaps, which are more common in smaller k values. This can lead to misassembles and an inaccurate representation of the genome.

b) At what point does the program output change when you increase k?

- The output changed at the point where the increase in k exceeds the length of common overlaps in the reads.
- When k is increased to **14,** some reads that previously overlapped with others by 10-13 bases no longer qualify to be merged.
- As a result, this causes the formation of an additional, separate contig because those reads cannot be joined with any other reads based on the new, stricter criteria.

Contigs increased from 20 to 21 contigs, because a set of reads that previously would have been merged into a single contig at a lower k-value now remain separate.

What is the relationship between how high k can go and sequencing coverage?

- The higher the value of k, the greater the sequencing coverage needed to ensure contig formation.

- Coverage refers to how many times a particular region of the genome is read during the sequencing process. High coverage means many reads overlap each region of the genome, increasing the likelihood of finding longer overlaps.
- If the coverage is insufficient, increasing k too much can result in a fragmented assembly where reads fail to overlap by the required length, leading to a greater number of shorter contigs.

In summary, a higher k-value requires high sequencing coverage to maintain a robust and complete assembly of the genome.

Consider sequencing as a whole

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

a. Explain how mate-pair reads enable us to get a higher L than paired end reads ?

Paired end:

In paired -end sequencing, DNA is fragmented into relatively short pieces, typically a hundred base pair long then Adapters are added to both ends of the fragment and sequenced, resulting in two short reads that originate from either end of each fragment. The distance between these reads, or "L", is limited by the length of the DNA fragment.

Mate-pair:

Mate-pair reads, however, are generated by a different library preparation method that allows for larger fragments of DNA to be sequenced.

In this technique, long strands of DNA are initially bent into a loop(circularized), effectively placing the ends of each strand side by side. The looped DNA is then cut into shorter segments. When these segments are sequenced, the resulting data corresponds to sequences that, in the original DNA strand, are far apart.

However, because of the initial looping, these sequences are now neighbors in the sequencing library. This method allows for the sequencing of DNA regions that are separated by a large distance within the genome, referred to as "L," the span between where the sequences originate in the genomic DNA.

b. 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)

- Average sequencing fragment size for paired-end: 400bp
- Mate-pair selection library size: 3000bp
- Read length for both methods: 80bp

Paired -End Sequencing:

The average distance L between the reads is the length of the fragment minus the lengths of the two reads (since both ends of the fragment are sequenced).

So, for a fragment of 400bp with each read being 80bp, the calculation is:

$L_{\text{paired end}}$ = Fragment Size−2 × Read Length

$L_{\text{paired end.}}$ = 400bp - 2 × 80bp

$L_{\text{paired end}}$ = 400bp – 160bp

$L_{\text{paired end}}$ = 240bp

This means the reads in paired-end sequencing are 240bp apart from each other.

Mate-Pair Sequencing:

For mate-pair sequencing, the distance between the reads is effectively the size of the original long fragment (before circularization and re-fragmentation) minus the lengths of the two sequenced ends.

So, for a mate-pair library size of 3000bp with each read being 80bp, the calculation is:

$L_{\text{mate-pair}} = \text{Library Size} - 2 \times \text{Read Length}$

$L_{\text{mate-pair}} = 3000\text{bp} - 2 \times 80\text{bp}$

$L_{\text{mate-pair}} = 3000\text{bp} - 160\text{bp}$

$L_{\text{mate-pair}} = 2840\text{bp}$

This means the mate-pair reads come from regions in the genome that are 2840bp apart from each other on average.

So, in summary:

- The average distance L for paired-end sequencing is 240bp.
- The average distance L for mate-pair sequencing is 2840bp

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?**

Generating reads from the same sequence using a long-read technology like PacBio (Pacific Biosciences) instead of generating the same amount of reads from replicates of that sequence (typically with short-read technologies) has several advantages:

1. **Resolution of Complex Regions**: Long reads can span entire genomic regions, including repetitive sequences and structural variants, which are often challenging for short-read technologies to resolve. This comprehensive view helps in reconstructing these complex regions accurately.
2. **Simplified Assembly**: Long-read sequencing simplifies the genome assembly process. Short reads must be pieced together through a computationally intensive process that can result in errors or gaps, especially in repetitive regions. Long reads, due to their length, provide more contiguous information, reducing the complexity of the assembly.
3. **Reduction of Ambiguity**: When assembling short reads, especially in areas with high sequence similarity, there can be considerable ambiguity in determining the correct order and orientation of sequences. Long reads greatly reduce this ambiguity because they cover these regions in a single, continuous read.
4. **Phasing of Haplotypes**: Long reads can keep track of variants that occur together on the same chromosome, known as phasing. This is beneficial for understanding the genomic structure and for studies involving compound genetic variations, which short reads often cannot achieve due to their limited length.

5. **Detection of Structural Variants**: Long reads are better at detecting structural variants like insertions, deletions, and inversions because they can cover the full length of these variants. Short reads might only capture parts of these variants, making it difficult to identify and characterize them fully.
6. **Time and Cost-Effectiveness**: Although long-read sequencing might have a higher cost per base, the reduced complexity of data analysis can make it more time and cost-effective overall, especially for certain applications like de novo assembly or structural variant analysis.
7. **Error Profile**: Long-read sequencing technologies have a more random error distribution, which can be corrected with higher coverage. Short-read sequencing technologies might have systematic biases or errors that are harder to correct.

In conclusion, while short-read sequencing technologies are useful and cost-effective for many applications, long-read sequencing technologies like PacBio offer advantages that can be critical for comprehensive genome analysis, particularly in complex or poorly understood regions.

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

let's consider **structural variations (SVs)**, which include deletions, insertions, duplications, inversions, and translocations.

Long reads are beneficial for detecting structural variations for several reasons:

1. **Spanning Variations**: Long reads can span the entire length of many SVs, providing clear evidence of the variation. This is unlike short reads, which may only partially cover the variation and require complex computational efforts to infer the SV.
2. **Unambiguous Mapping**: Long reads can often be uniquely mapped to the genome, even in repetitive regions, allowing for precise identification of SV breakpoints.
3. **Complex Rearrangements**: Long reads can resolve complex genomic rearrangements that are difficult to reconstruct with short reads, which might break the rearrangements into many small, confusing pieces.
4. **Phasing**: Long reads can help in phasing variants, which means determining which variants co-occur on the same segment of DNA. This is particularly useful for understanding the compound effects of multiple SVs.

In summary, long reads provide a more direct and less ambiguous means of detecting structural variations, leading to a better understanding of the genome's structure and its variations.

References

O'Connell, J., Schulz-Trieglaff, O., Carlson, E., Hims, M. M., Gormley, N. A., & Cox, A. J. (2015). NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics (Oxford, England)*, *31*(12), 2035–2037. https://doi.org/10.1093/bioinformatics/btv057.

Fujimoto, A., Wong, J. H., Yoshii, Y., Akiyama, S., Tanaka, A., Yagi, H., Shigemizu, D., Nakagawa, H., Mizokami, M., & Shimada, M. (2021). Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Medicine*, *13*(1), 65–15. https://doi.org/10.1186/s13073-021-00883-1.