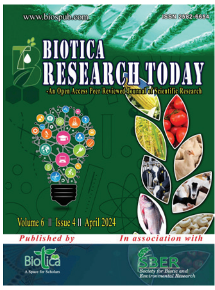

Biotica Research Today

e-ISSN: 2582-6654

April, 2024

Article ID: RT1617

Popular Article

De novo Genome Assembly: Challenges and Solutions

Sindhu D.[1], Satish Hosakoti[2], Bidwan Rath[1*], Sinchana Kashyap G.S.[2] and Basanagouda Gonal[3]

[1]Division of Genetics, ICAR-Indian Agricultural Research Institute (IARI), New Delhi (110 012), India
[2]Dept. of Genetics and Plant Breeding, University of Agricultural Sciences (UAS), Bangalore, Karnataka (560 065), India
[3]CSB-Central Sericultural Research & Training Institute (CSR&TI), Pampore, Jammu & Kashmir (192 121), India

Open Access

Corresponding Author

Bidwan Rath

✉: bidwan2306@gmail.com

Conflict of interests: The author has declared that no conflict of interest exists.

How to cite this article?

Sindhu, D., Satish, H., Bidwan, R., *et al*., 2024. *De novo* Genome Assembly: Challenges and Solutions. *Biotica Research Today* 6(4), 192-194.

Abstract

De novo assembly is a computational process used in genomics to reconstruct genomes from short DNA sequencing reads without a reference genome. Current article outlines the definition, steps, constraints and solutions associated with *de novo* assembly. *De novo* assembly is crucial for studying non-model organisms, identifying genetic variations and understanding evolutionary relationships. A general outline of the steps involved in *de novo* assembly has been provided; however, slight variations may occur based on the approach to assembly employed, whether it is overlap-layout-consensus or de Bruijn graph-based. Constraints such as sequencing errors, repetitive sequences and genome size variations pose challenges to accurate assembly. Solutions to these challenges involve employing advanced algorithms, optimizing sequencing technologies and integrating multiple data sources. Understanding and overcoming these constraints are essential for enhancing the accuracy and completeness of *de novo* assembly, thereby enhancing the output from various genomic studies and applications.

Keywords: De Bruijn graph, *De novo* genome assembly, Fast QC, Overlap layout consensus

Introduction

Potential and orphan crops are of paramount importance in the current era of climate change and rapidly exploding human population. Yet, delving into the genomes of these lesser-studied crops presents a formidable challenge due to unavailability of a robust reference genome. In this context, *de novo* assembly plays as a vital rescue tool. *De novo* genome assembly refers to the process of reconstructing genomic sequences from short sequencing reads without the aid of a reference genome. In other words, it involves piecing together DNA fragments obtained from sequencing technology to create a complete genome assembly, especially when a reference genome is unavailable or when the organism being studied has significant genomic differences from existing reference genomes.

In *de novo* assembly, algorithms are employed to piece together overlapping sequencing reads into longer contiguous sequences, called contigs and further assemble these contigs into larger scaffolds or even complete genomes. This process involves various computational techniques to resolve repetitive regions, correct sequencing errors and bridge gaps between contigs. *De novo* assembly is a crucial step in many genomic and transcriptomic studies, enabling researchers to explore the genetic makeup and functional elements of organisms, identify novel genes, understand evolutionary relationships and investigate genetic variations within and between populations.

De novo Assembly in Bioinformatics

De novo assembly plays a crucial role in bioinformatics for several reasons (Nagarajan and Pop, 2013).

1. Discovery of Novel Genomic Features: *De novo* assembly allows researchers to explore organisms lacking well-characterized genomes, leading to the identification of unique genes, regulatory elements and non-coding RNAs specific to the organism.

2. Population Genomics: By assembling genomes without a reference, researchers can analyze genetic variations like

Article History

RECEIVED on 06th April 2024 RECEIVED in revised form 21st April 2024 ACCEPTED in final form 22nd April 2024

single nucleotide polymorphisms (SNPs) and structural changes within and between populations. This aids in studying evolutionary dynamics, population structures and adaptive processes.

3. Personalized Genomics: *De novo* assembly can analyze individual genomes independently, making it valuable for pinpointing rare genetic variants linked to diseases or drug responses. This supports the development of tailored therapeutic approaches.

4. Comparative Genomics: *De novo* assembly provides complete or near-complete genome sequences, facilitating comparisons across species. This aids in unraveling genome evolution, identifying gene family dynamics and exploring evolutionary relationships.

5. Improving Reference Genomes: *De novo* assembly contributes to refining existing reference genomes by addressing gaps, resolving repetitive regions and rectifying errors. This results in more accurate reference sequences, enhancing downstream genomic analyses.

Steps Involved in *De novo* Assembly Construction

De novo assembly in bioinformatics involves the following key steps; however, there may be variations depending on the sequencing technology used and the characteristics of the genome being assembled (Nagarajan and Pop, 2013).

1. Pre-Processing

- *Quality Control*: Assess and filter sequencing reads based on quality scores to remove low-quality reads using tools like Fast QC and Trimmomatic.
- *Adapter Trimming*: Trim sequencing adapters and other technical sequences from the reads to ensure cleaner input data.

2. Error Correction

- *Error Correction*: Correct sequencing errors in the reads using error correction algorithms such as Bayes Hammer, Quor UM, or Karect to improve the accuracy of assembly.

3. Assembly

- *Overlap Detection*: Identify overlaps between sequencing reads to determine their relative positions and orientations.
- *Contig Construction*: Assemble the reads into contiguous sequences (contigs) using assembly algorithms like de Bruijn graph-based methods (*e.g.*, SPAdes, Velvet, *etc.* software) or overlap-layout-consensus (OLC) methods (*e.g.*, Canu, Miniasm, *etc.* software).
- *Scaffolding*: Extend contigs and order them into larger scaffolds using paired-end or mate-pair information to bridge gap between contigs (Figure 1).

4. Gap Closing

- *Gap Filling*: Use iterative approaches or specialized algorithms to fill gaps within scaffolds by identifying and resolving regions of ambiguous or missing sequence data.

5. Quality Assessment

- *Evaluation*: Assess the quality of the assembled genome using metrics such as N50 length, genome completeness and alignment statistics.
- *Validation*: Validate the assembly by comparing it to known reference genomes, if available, or using experimental validation techniques such as PCR or optical mapping.

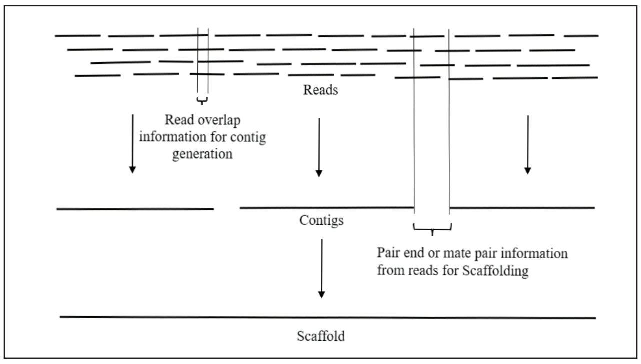
Reads
Read overlap information for contig generation
Contigs
Pair end or mate pair information from reads for Scaffolding
Scaffold

Figure 1: Schematic representation of obtaining contigs from reads and scaffold from contigs in de *novo* assembly

Challenges Associated with *De Novo* Assembly Construction

De novo assembly poses several challenges due to various factors inherent in the sequencing technology and depending on degree of genome complexity (Liao *et al.*, 2019). Few major key challenges associated with *de novo* assembly construction include:

1. Short Read Lengths: Many sequencing platforms produce short reads, which can make it difficult to accurately reconstruct long genomic regions, especially in repetitive or complex regions of the genome.

2. Errors in Sequencing: Sequencing errors, including base substitutions, insertions and deletions, can complicate the assembly process and lead to inaccuracies in the final assembly.

3. Genomic Variability: Genomes can exhibit structural variations, such as duplications, inversions and translocations, which impose challenges for assembly algorithms, particularly when assembling genomes from closely related individuals or populations.

4. Repeat Regions: Repetitive sequences, such as tandem repeats, transposable elements and segmental duplications, can confound assembly algorithms by creating ambiguities in the assembly graph or causing collapses or expansions of repeat regions.

5. Heterozygosity: Heterozygous regions in diploid or polyploid genomes can complicate assembly by creating allelic variations that may be incorrectly merged or separated during the assembly process.

6. Uneven Coverage: Variability in sequencing coverage across the genome can result in regions with low coverage or gaps in the assembly, making it challenging to accurately reconstruct those regions.

7. Computational Complexity: *De novo* assembly algorithms often demands significant and intensive computational resources. Moreover, when genome under *de novo* assembly is large and complex, it further makes the task difficult.

8. Assembly Validation: Assessing the quality and accuracy of *de novo* assemblies can be challenging and misassemblies or assembly errors may not always be easily detectable without additional validation techniques.

To address these challenges, a combination of experimental optimization, algorithm development and computational approaches are required to improve the accuracy as well as completeness during *de novo* assembly.

Solution for the Challenges Associated with *De Novo* Assembly Construction

1. Sequence Errors

a) Error Correction Algorithms: Pre-processing raw sequencing data using error correction algorithms such as Bayes Hammer, Quor UM, or Karect can help identify and correct sequencing errors before assembly.

b) Quality Trimming: Trimming low-quality bases from sequencing reads using tools like Trimmomatic or Cutadapt can improve the accuracy of assembly by removing unreliable data.

2. Repetitive Regions

a) Long-Read Sequencing: Long-read sequencing technologies like PacBio or Oxford Nanopore can span repetitive regions more effectively than short-read sequencing, facilitating the accurate reconstruction of these regions.

b) Optical Mapping: Optical mapping techniques, such as Bionano Genomics, can provide complementary long-range genomic information to aid in resolving repetitive regions during assembly.

3. Genome Size

a) Parallelization and Distributed Computing: Leveraging parallelization and distributed computing resources can accelerate the assembly process for large genomes by distributing the computational tasks among multiple processors or nodes.

b) Subsampling: Subsampling or partitioning the genome into smaller, manageable chunks for assembly can mitigate the computational challenges associated with assembling large genomes.

4. Uneven Coverage

a) Read Filtering: Filtering sequencing reads based on coverage depth or quality thresholds can help mitigate the effects of uneven coverage by removing low-quality or excessively redundant data.

b) Depth Normalization: Normalizing sequencing coverage across the genome using tools like BB Norm or khmer can reduce biases and improve the evenness of coverage, leading to more uniform assembly quality.

5. Chimeric Contigs

a) Mate-Pair Libraries: Utilizing mate-pair or long-insert libraries during sequencing will help to get additional information on relative positions of genomic fragments, helping to identify and resolve chimeric contigs.

b) Quality Assessment Tools: Employing quality assessment tools such as QUAST or BUSCO to evaluate assembly completeness and accuracy can aid in detecting and mitigating chimeric contigs through iterative refinement of the assembly process (Gurevich *et al*., 2013).

Conclusion

It is pivotal in understanding novel organisms or those with complex genomes, shedding light on genetic variations, evolution and functional genomics. The significance of *de novo* assembly is immense in situations where a reference genome is lacking. Moreover, gene mapping techniques such as MutMap-Gap have incorporated *de novo* assembly in its pipeline (Takagi *et al*., 2013). *De novo* assembly faces challenges such as computational complexity, sequencing errors, repetitive sequences, heterozygosity, *etc*. To overcome these constraints, innovative methodologies such as graph-based algorithms, hybrid approaches and error correction techniques have been developed. Additionally, advancements in sequencing technologies, coupled with computational tools, continue to refine *de novo* assembly, making it an indispensable tool in genomic research and beyond.

References

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29(8), 1072-1075. DOI: https://doi.org/10.1093/bioinformatics/btt086.

Liao, X., Li, M., Zou, Y., Wu, F.X., Wang, J., 2019. Current challenges and solutions of *de novo* assembly. *Quantitative Biology* 7(2), 90-109. DOI: https://doi.org/10.1007/s40484-019-0166-9.

Nagarajan, N., Pop, M., 2013. Sequence assembly demystified. *Nature Reviews Genetics* 14, 157-167. DOI: https://doi.org/10.1038/nrg3367.

Takagi, H., Uemura, A., Yaegashi, H., Tamiru, M., Abe, A., Mitsuoka, C., Utsushi, H., Natsume, S., Kanzaki, H., Matsumura, H., Saitoh, H., Yoshida, H., Cano, L.M., Kamoun, S., Terauchi, R., 2013. MutMap-Gap: Whole-genome resequencing of mutant F2 progeny bulk combined with *de novo* assembly of gap regions identifies the rice blast resistance gene *Pii*. *New Phytologist* 200(1), 276-283. DOI: https://doi.org/10.1111/nph.12369.