THE SCIENCE WORLD
A Monthly e Magazine
ISSN:2583-2212
July, 2023; 3(07), 1358-1360

Popular Article

Methodology of Whole genome assembly in Prokaryotes

Gunturu Narasimha Tanuj[1] and Chandana M.S.[2]
[1]Division of Veterinary Biotechnology, ICAR- Indian Veterinary Research Institute
[2]Division of Veterinary Microbiology, ICAR- Indian Veterinary Research Institute

https://doi.org/10.5281/zenodo.8150756

Abstract

Genome assembly is done by joining together the short stretches of DNA sequences in a defined fashion to re-create the oiginal form from which it was extracted. A completely defined and resolved genome aids in studying the bacterial, fungal, parasitic and viral disease pathogenesis, identifying and annotating various metabolic pathways in a specific organism, describing the SNPs and mutations leading to specific resistance or virulence condition, deducing the evolutionary pattern and to study the gene expression and aternate splicing events. In this article, I discuss regarding the overview of genome assembly algorithms in prokaryotes and its applications.

Introduction

Recent advances in next generation and third generation sequencing technologies have tremendously advanced the field of whole genome sequencing and assembly. This paved the way for complete genome assembly of many prokaryotes, eukaryotes and viruses. Although traditionally short reads from next generation sequencing platforms were used for assembly generation, long reads from third generation platforms have been replacing the former in the recent times. Nevertheless, short reads are still essential for assembly polishing and reads correction.

Types and algorithms

There are two main types of genome assembly: ***De novo* assembly** and **Reference guided assembly**. In *de novo* assembly, assembly is done from scratch without the help of any reference genome. This method can be used for structural variation detection and identification of novel transcripts. In reference guided assembly, a representative genome is used for easy and quick assembly of the genome. But it faces the limitation of needing a reference genome.

*Published:*14.07.2023

Genome assembly algorithms are of two types: Debrujin graph (DBG) and Overlap Layout Consensus (OLC). In DBG method, the reads are first split into k-mers, then a DBG is generated from the chopped k-mers and a genome sequence is obtained from the graph. K-mer refers to a specific **number of nucleotides** that overlap. Whereas in OLC, overlaps between the reads are identified from which consensus sequences will be built and scaffolded. Contigs are sequences that overlap in a particular fashion such that they provide a contiguous representation of a genomic region. It is the first level in hierarchy of assembly whereas, scaffold or super-contigs come next. Scaffolds are built by joining contigs in the correct order and the gaps are filled with a stretch of Ns.

Assembly evaluation

After the genome assembly is done, it is very important to evaluate its quality. Several parameters are considered to check this, out of which N50 is an important statistical measure. It represents the sum of lengths of all contigs of size N50 or longer contain at least 50% of the total genome sequence. It describes about the completeness of the genome where in, half of the genome sequence is covered by contigs larger than or equal to the N50 size. Other quality check parameters include NG50, N90, Average contig length, Coverage, number of genes and number of gaps. If 90% of the bases have at least 5X read coverage, the genome is considered accurate.

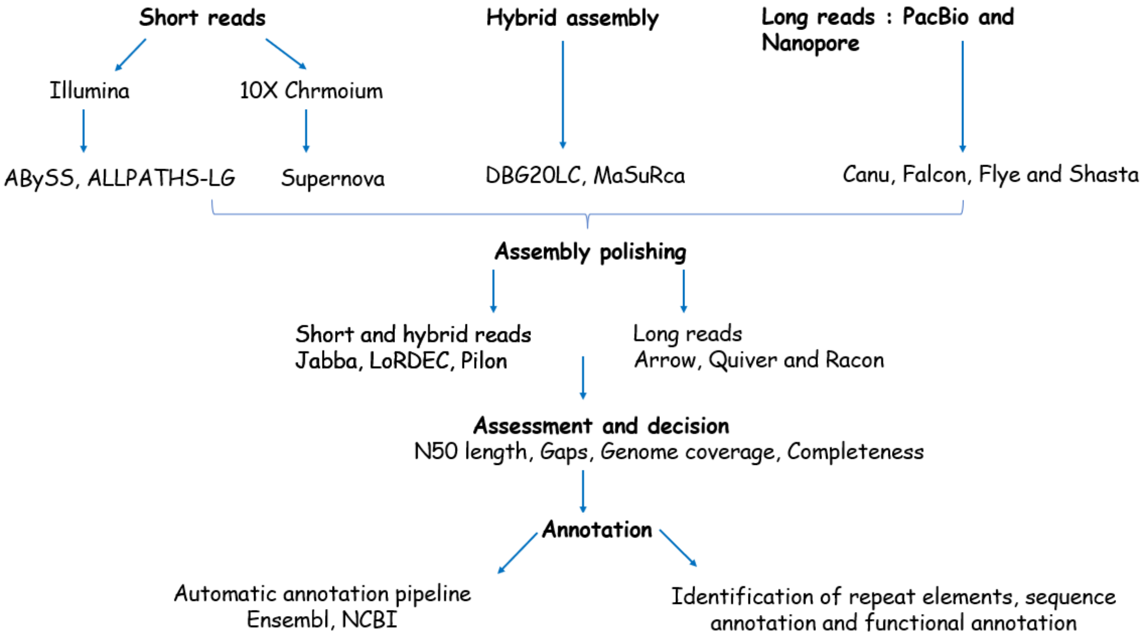

Figure 1: Overview of whole genome assembly in prokaryotes

Applications of whole genome sequencing studies

a. Generate whole genome sequences of different prokaryotes, eukaryotes and viruses
b. Provide high resolution, base-by-base view of the genome
c. Aids in accurate integration of genotypes and phenotypes
d. Helps in advancing the fields of phylogenomic and comparative, functional and population genomics
e. Detailed picture of infectious agents and their taxonomic forms
f. Aids in understanding resistance factors, virulence agents which can improve our knowledge on pathogenesis of infectious diseases
g. Essential in designing diagnostic, therapeutic and prophylactic platforms

Conclusion

Advancements in the sequencing platforms and computational methods have greatly improved the field of genome assembly. Although many prokaryotic genomes have been assembled and annotated in the recent times, there are still lot many pathogenic bacteria that have not been completely assembled and annotated. One of the major reasons for the quick development of vaccines and diagnostic kits during SARS-C0V2 outbreak was the timely completion of its genome assembly. In this regard, genome assembly studies in prokaryotes have a tremendous potential in the field of diagnostics, prophylactics and therapeutics which has to be explored efficiently in the coming days.

References

Khan AR, Pervez MT, Babar ME, Naveed N, Shoaib M. A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evol Bioinform Online*. 2018; 14:1176934318758650.

Wee Y, Bhyan SB, Liu Y, Lu J, Li X, Zhao M. The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Brief Funct Genomics*. 2019;18(1):1-12.

Jung H, Ventura T, Chung JS, et al. Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Comput Biol*. 2020;16(11):e1008325.