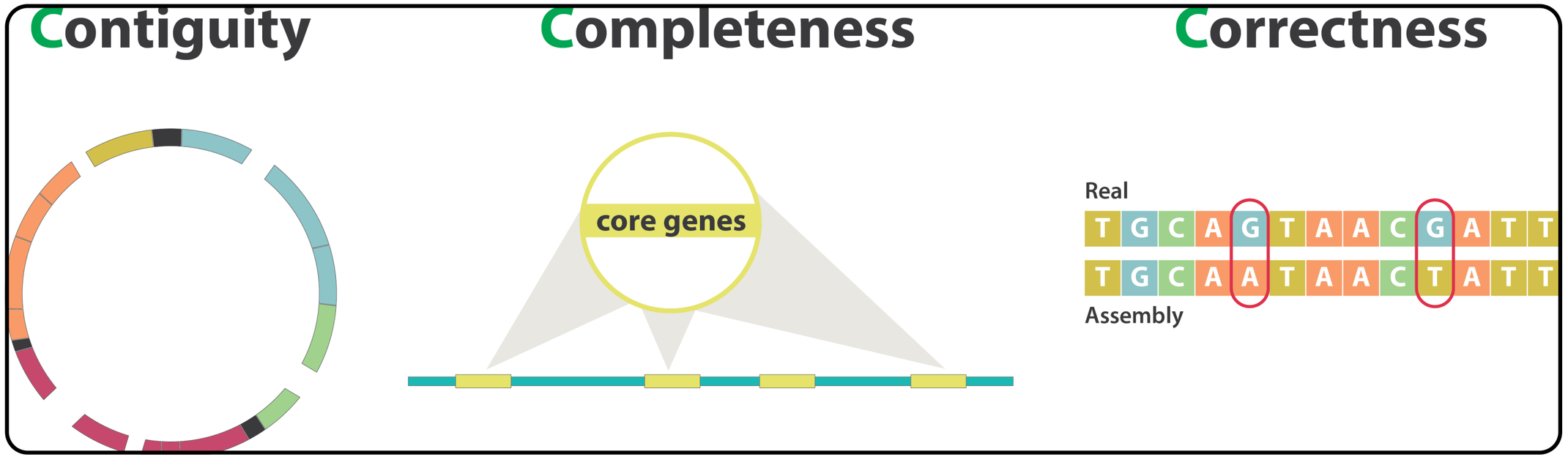
ASSEMBLY QUALITY ASSESSMENT N50

- Total Length
- Number of Contigs/scaffolds
- Mapping

PRESENTATION BY:

1. Khushi Pal
2. Rushikesh Shingole
3. Vrushali Bathe
4. Isha K. Chavan

ASSEMBLY QUALITY ASSESSMENT



Assembly quality assessment of genome is the process of evaluating the quality of a genome assembly, which involves the computation of various metrics to measure the contiguity, correctness, and completeness of the assembly.

1. **Correctness:**

Correctness refers to the accuracy of the assembled sequences in representing the original genome.

Assessment:

- Misassemblies:
- Errors:

2. **Contiguity:**

Contiguity assesses how well the assembled sequences represent the original genome in terms of their length and continuity.

Assessment:

- N50 statistic:
- Longest contig/scaffold

3.Completeness

Completeness is determined by the content of contigs, especially with regard to gene content.

Aim: Is to have the highest percentage of genes identified in your assembly, with a BUSCO complete score above 95% considered good.

BUSCO:Benchmarking Universal Single-Copy Orthologs

Advantages of assembly quality assessment

- Ensuring Accuracy
- Improving Completeness
- Enhancing Contiguity
- Validating Experimental Techniques:

Disadvantages of assembly quality assessment

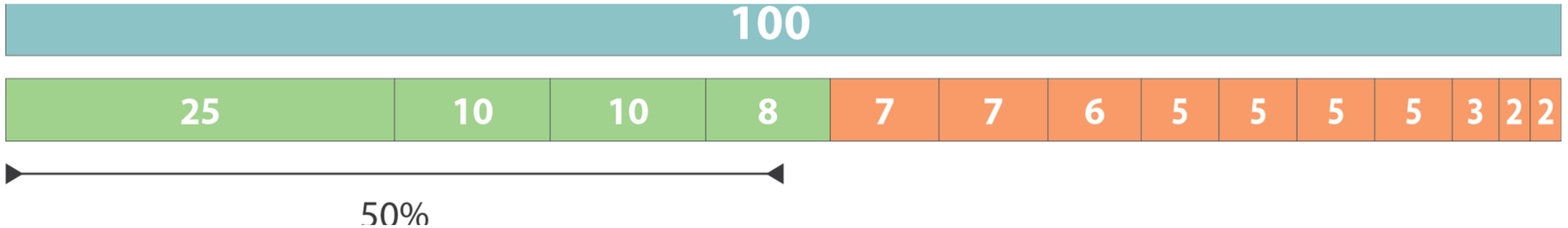
- Difficulty in Correcting Errors
- Challenges with Repetitive Regions
- Resource Intensive
- Difficulty in Assessing Novel Genomes:

Metrics used in Quality assessment tool

N50

- N50 is a metric widely used to assess the contiguity of an assembly, which is defined by the length of the shortest contig for which longer and equal length contigs cover at least 50 % of the assembly.

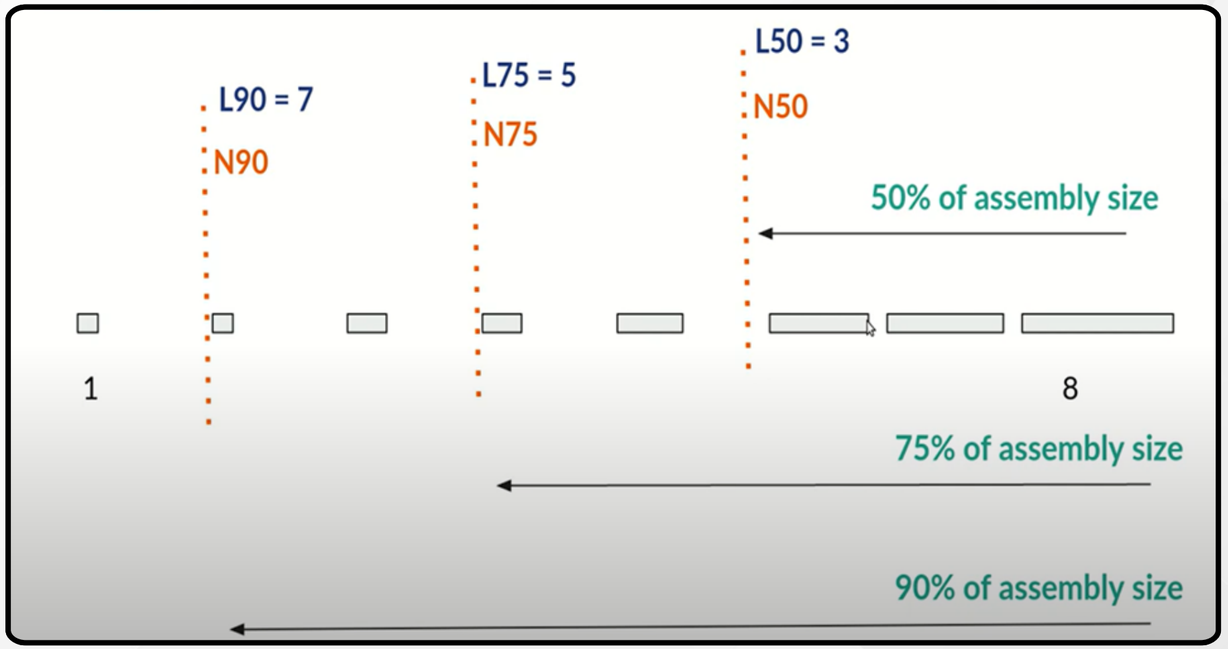
- N50 represents the point of half of the mass of the distribution, with half of the entire assembly contained in contigs or scaffolds that are larger than or equal to the N50 value.



N50 is defined as the sequence length of the shortest contig at 50% of the total assembly length.

L50 count is defined as the smallest number of contigs whose length sum makes up half of assembly size.

1 8



Example of N50

Example

Contig-1 : 5
Contig-2 : 4
Contig-3 : 2
Contig-4 : 1
Assembly size = 5+4+2+1 = 12
50% assembly size = 6

Half of the genome length is covered by the two largest contigs, including the 4kb contig.
N50=4kb is the minimum contig length required to cover 50 percent of the assembled genome sequence.
L50 = 2, Min. 2 contigs required.

N50 or NG50?

Note that *N50* is calculated in the context of the assembly size rather than the genome size. Therefore, comparisons of N50 values derived from assemblies of significantly different lengths are usually not informative, even if for the same genome. To address this, the authors of the Assemblathon competition came up with a new measure called *NG50*. The **NG50 statistic** is the same as *N50* except that it is 50% of the known or estimated genome size that must be of the NG50 length or longer. This allows for meaningful comparisons between different assemblies.

Assembly-1 size: 50kb, size to consider for N50 is 25kb
Assembly-2 size: 100kb, size to consider for N50 is 50kb

If genome size is 100kb
Then for NG50, size to consider for NG50 is 50kb for both assembly

Advantages of N50

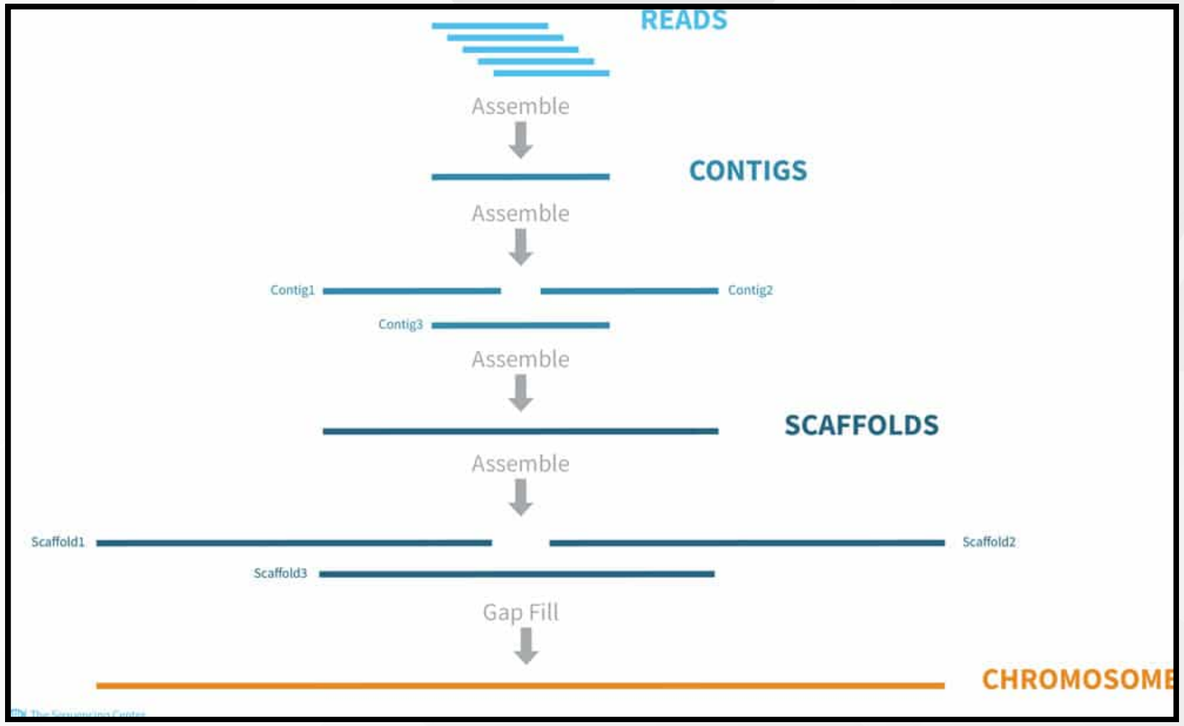
- It is still only a measure of sequence, but comparable for same genome contiguity
- There is still a limit on when it will not improve further.
- Smaller contigs can be filtered outwithout affecting the value

Disadvantages of N50

- N50 is not a measure of assembly correctness it only measures sequence contiguity.
- N50 is biased if short sequences are excluded
- N50 is not meaningful for different assembly sizes

Reads,Contigs and Scaffolds

1. **Reads : Read is a short sequence obtained after NGS & is obtained from fragmented DNA,It is a smallest and basic unit of sequencing.**
2. **Contigs: These small reads are assembled to form large contigs these are based upon overlapping region of reads.**
3. **Scaffolds: When the contigs are connected together by incorporating gaps we get scaffold**



Tools for Assembly quality assessment

- **QUAST**
- **GenomeQC tool**
- **BUSCO**
- **CEGMA**

QUAST

QUAST (QUality Assessment Tool) is another tool that computes various metrics to evaluate genome assemblies, comparing them with a reference genome if available. It provides statistics like contig numbers, largest contig length, total length, GC content, and more

Quast (QUality Assesment Tool) , evaluates genome assemblies by computing various metrics, including:

1. N50
2. L50

QUAST TOOL

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

Installation | QUAST | MetaQUAST | QUAST-LG | Icarus | Web interface | Manual | Publications

QUAST – Quality Assessment Tool for Genome Assemblies

The project aim is to create easy-to-use tools for genome assemblies evaluation and comparison.
Currently, we are working on four tools which are distributed inside one package:

- QUAST for regular genome assemblies
- MetaQUAST for metagenome assemblies
- QUAST-LG for large genome (e.g. mammalian) assemblies
- Icarus for contig alignment visualization

downloads 89k

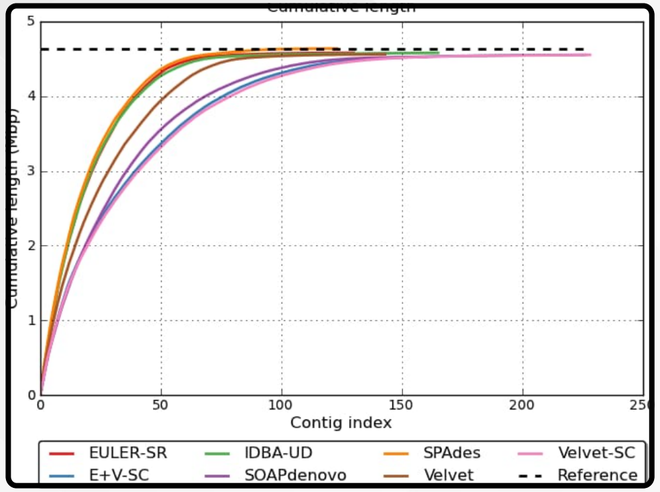
About us:

- GenomeWeb
- BioStar
- Homologus
- SEQwiki

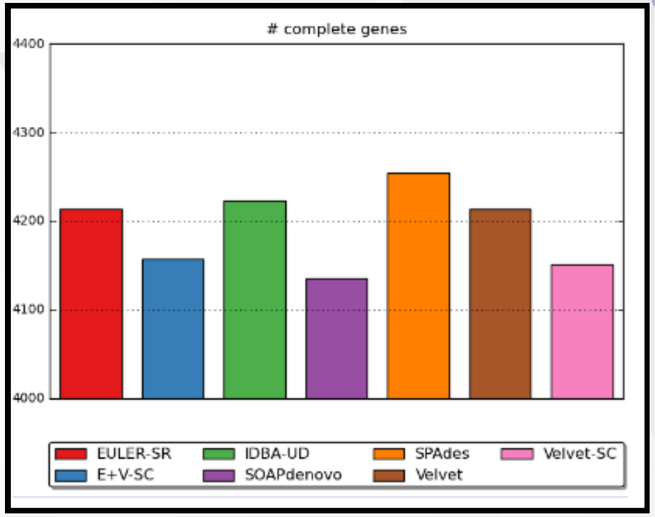
Key news:

- **June 7, 2022** — the version 5.2 is released!
- **August 31, 2021** — CZI acknowledged QUAST as an essential software tool for biomedicine and supported with a grant (jointly with SPAdes)
- April 3, 2020 — QUAST web server moved to a new location and working again!

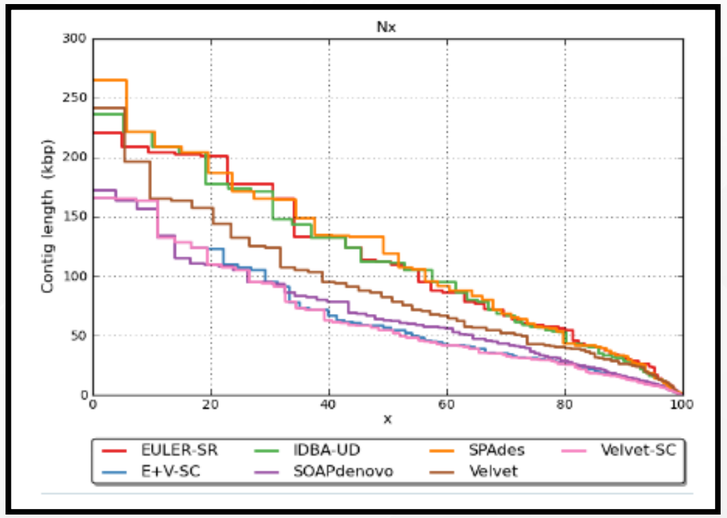
QUAST TOOL



Contig alignment plot



GC content plot



Nx-like plot

SCAFFOLD MAPPING

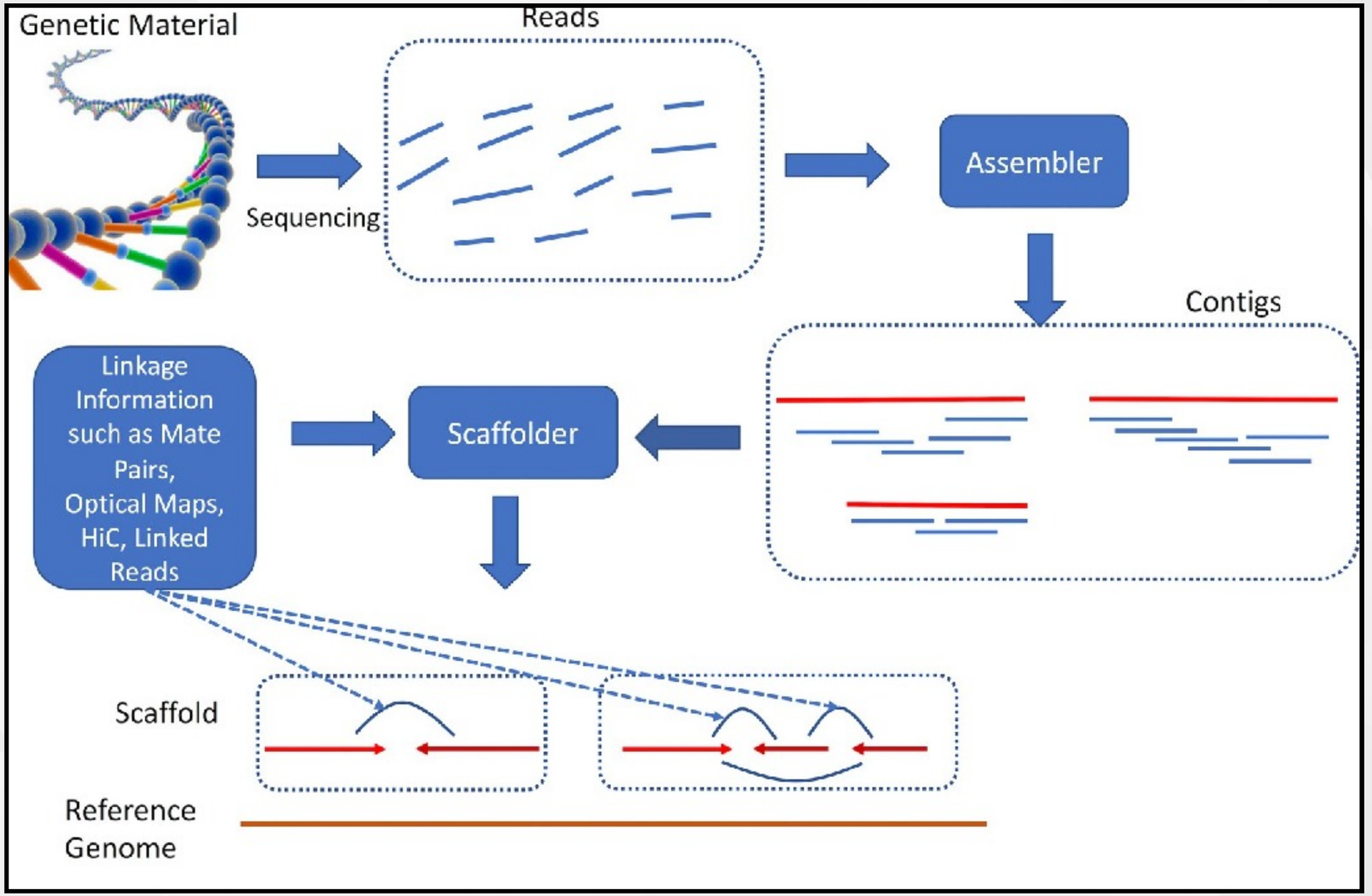
- Scaffolding is a technique in bioinformatics used to order and orient pieces of DNA sequence, providing a framework to support the construction of larger genomic sequences.

- Scaffolding is a part of bioinformatics, which integrates biology, computer science, and information technology to analyze and interpret biological data.

Key Purpose

The primary goal of scaffolding is to assemble smaller DNA sequences into a complete genome, aiding in genome assembly and the understanding of genomic structures.

STEPS OS SCAFFOLD MAPPING



Advantages of scaffold mapping

- Efficient Compound Optimization
- Accelerated Drug Discovery
- Insightful SAR Analysis

Disadvantages of scaffold mapping

- Difficulty in Mapping Flexible Molecules.
- Limited Scope of Scaffold.
- Complexity of SAR.

Applications of Scaffold Mapping

- Scaffold mapping can also be used in the context of tissue engineering and biosensing, where scaffold-based materials are used to create three-dimensional structures that can be used to study cell behavior and to develop analytical methods for biomarker detection

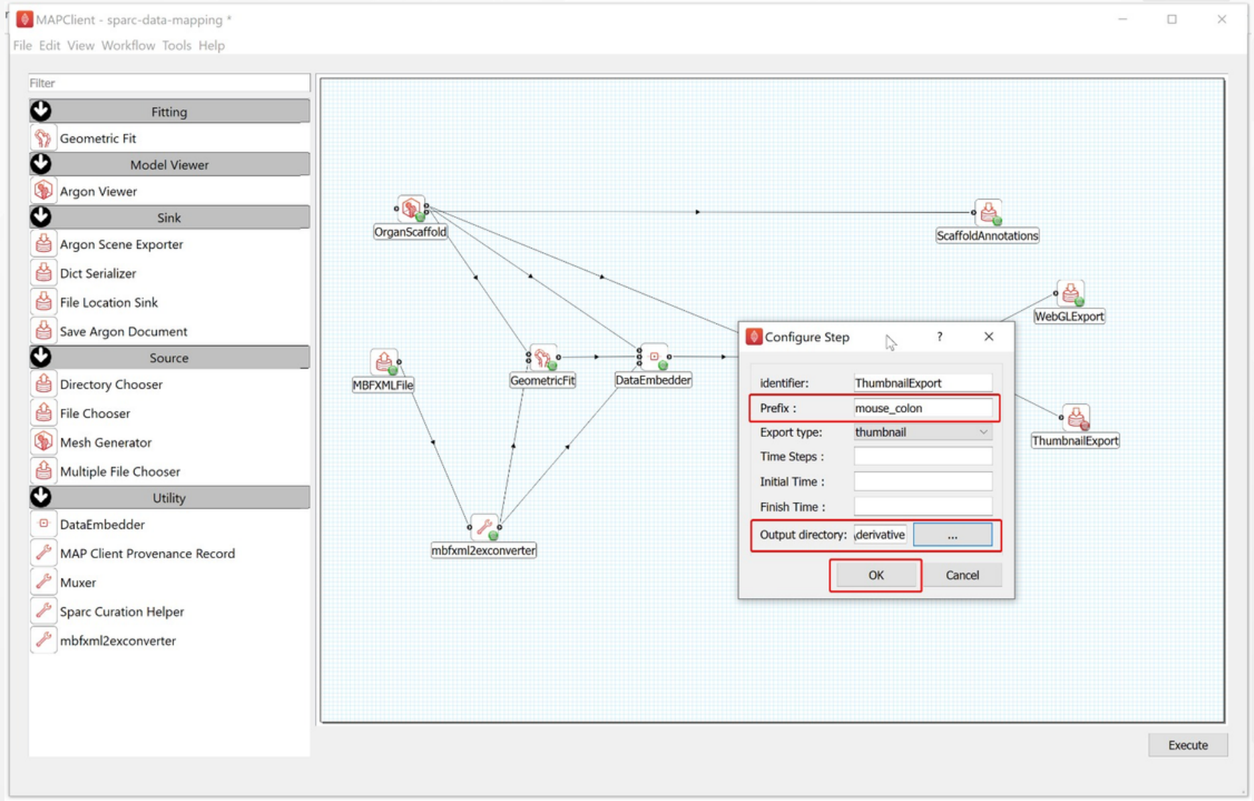
- his process can help to identify and correct mis assemblies in draft genomes, and it can also be used to identify and characterize structural variants, such as inversions and translocations, between different genomes

- scaffold mapping can be used to study the evolutionary history of genomes, by comparing the order and orientation of scaffolds between different species

Applications of Scaffold Mapping

Scaffold Mapping Tool:

The SPARC (Stimulating Peripheral Activity to Relieve Conditions) project provides a collection of tools for mapping SPARC data to organ scaffolds through a specialized release of the MAP-Client application called the MAP-Core scaffold mapping tool



REFERENCES

- Narwade, N., Patel, S., Alam, A., Chattopadhyay, S., Mittal, S., & Kulkarni, A. (2019). Mapping of scaffold/matrix attachment regions in human genome: a data mining exercise. Nucleic acids research, 47(14), 7247–7261. https://doi.org/10.1093/nar/gkz562
- Lander et al. (2001) Initial sequencing and analysis of the human genome. Nature 409.6822: 860-921.
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics (Oxford, England), 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086
- Duitama J, et al. Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. PLoS One. 2015;10(4):e0124617
- Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. 2013;2:10.
- Tang, H., Zhang, X., Miao, C. et al. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol 16, 3 (2015). https://doi.org/10.1186/s13059-014-0573-1
- Andrew J. Page, Ben Taylor, Aidan J. Delaney, Jorge Soares, Torsten Seemann, Jacqueline A. Keane and Simon R. Harris https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000123?crawler=true
- O. (2023, February 15). The Easiest way to Calculate N50 for Genome Assembly. One Stop Data Analysis. https://onestopdataanalysis.com/n50-genome/
- B. (2021, June 1). Different Assembly statistics (N50, L50, NG50, LG50, NA50, NGA50 and Misassemblies). YouTube. https://www.youtube.com/watch?v=ViXzKrQo25k