Genome statistics and annotation

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \mathbf{2} & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & \mathbf{3} & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & \mathbf{5} & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & \mathbf{7} & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & \mathbf{9} & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & \mathbf{11} & \mathbf{10} & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & \mathbf{12} \end{pmatrix}$$

By Dr. Richard Allen White III

Lecture 3 - Sep 17th, 2019

Zoom! 404-899-586

Assembly QC and annotation

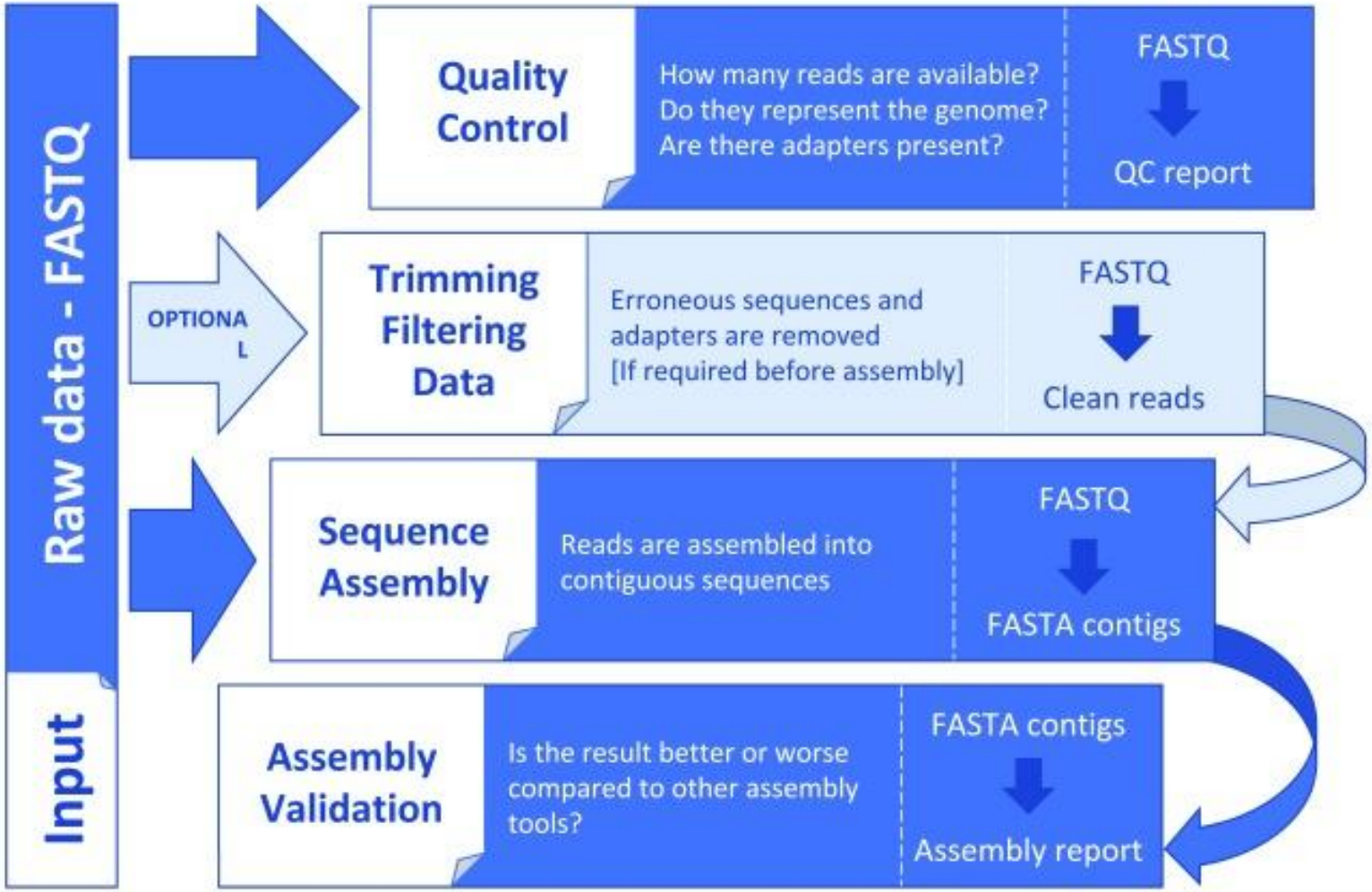
Concepts:

- Assessing genome assembly
- Genome assembly statistics
- Genome coverage
- BLAST
- Local vs. global alignment

Learning Objectives:

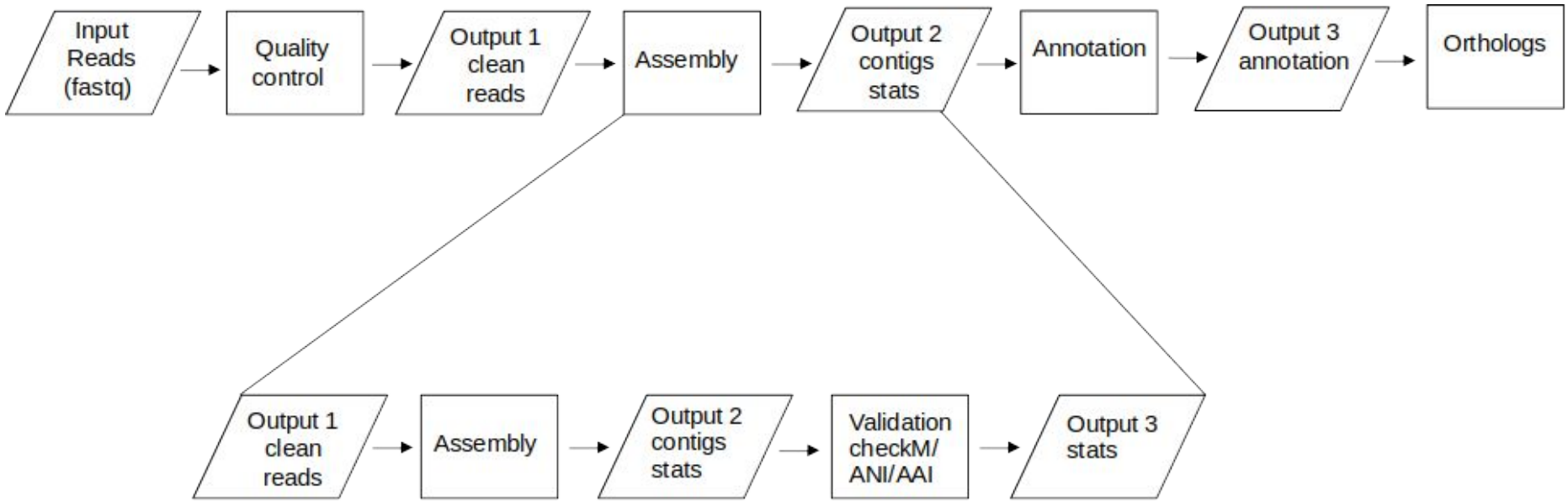
- Good vs. poor assemblies
- Check*M* for quality control of assembly and genomes
- Finding if you genome is clean or not?

Genome sequencing - flowgraph

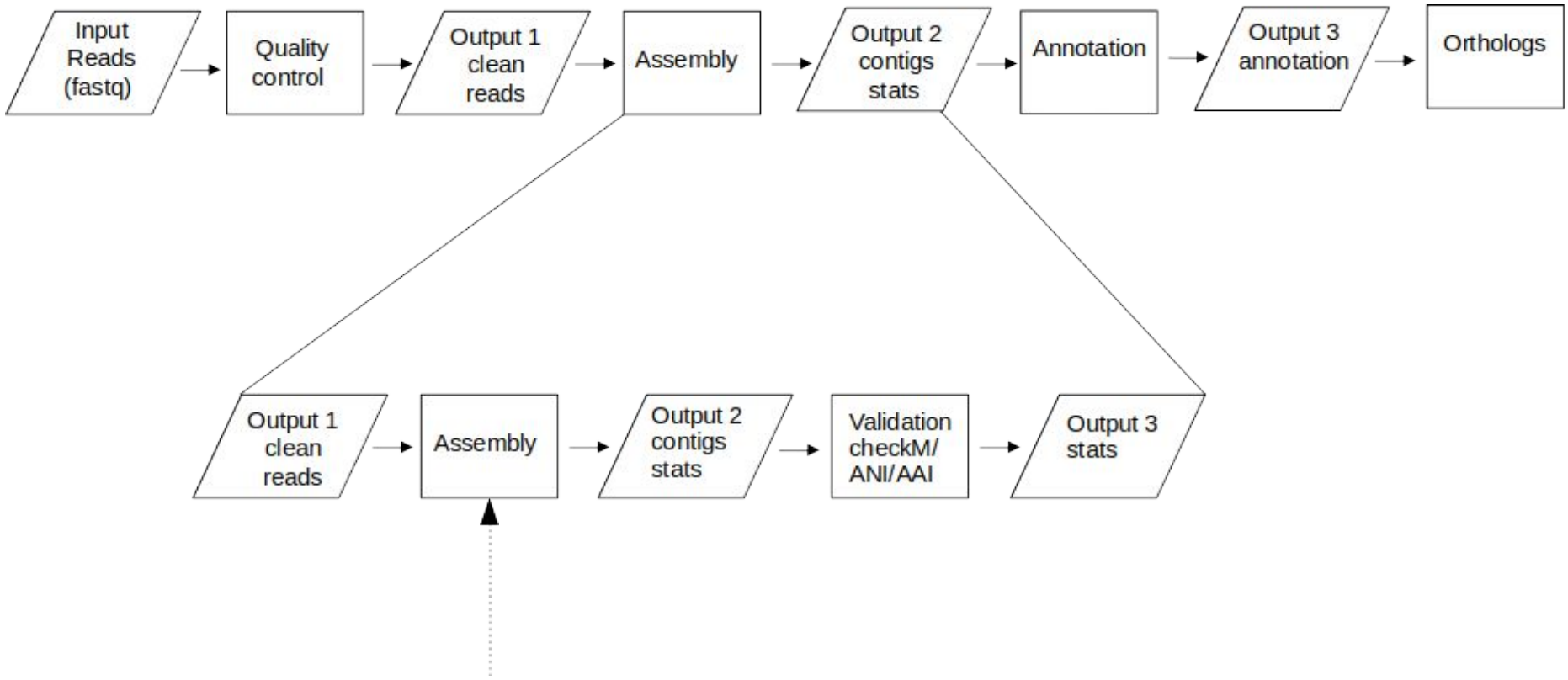


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850084/

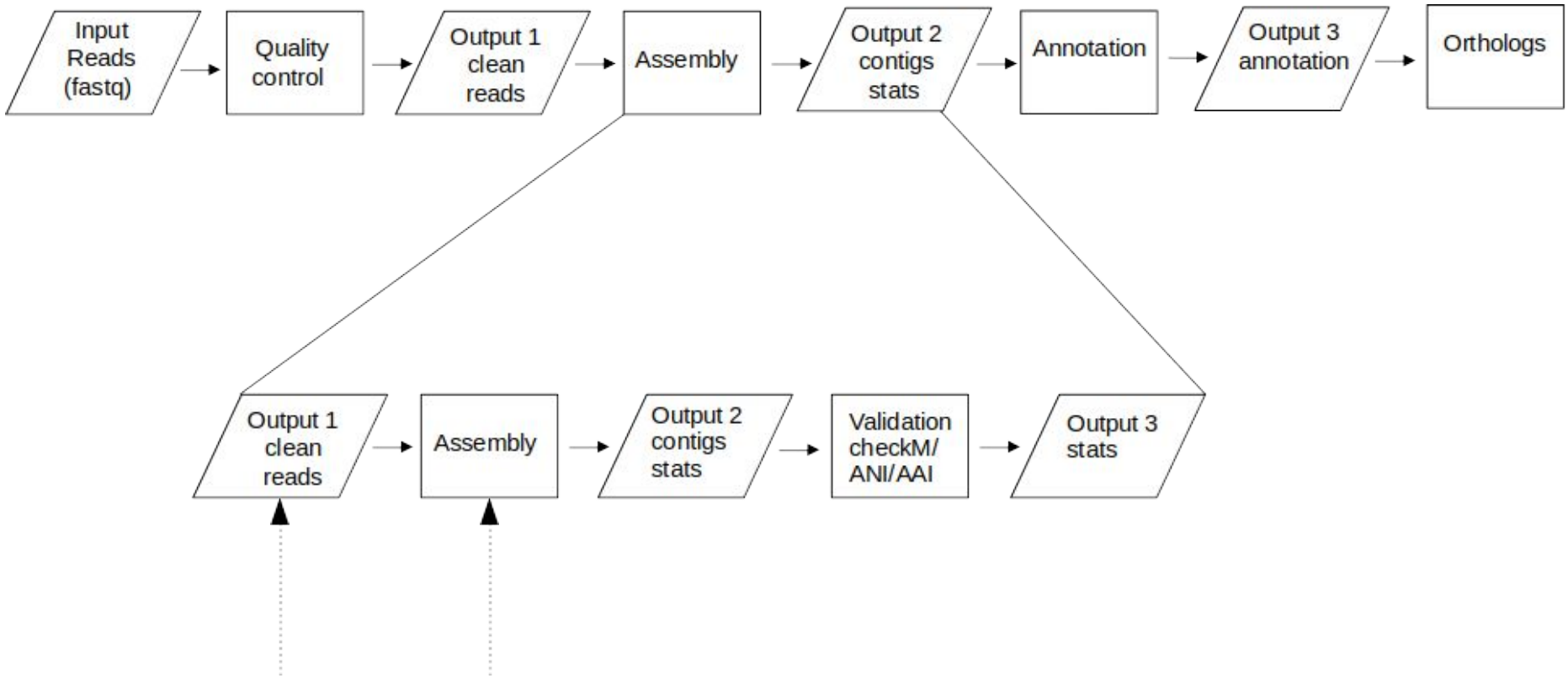
Assembly QC - flowgraph



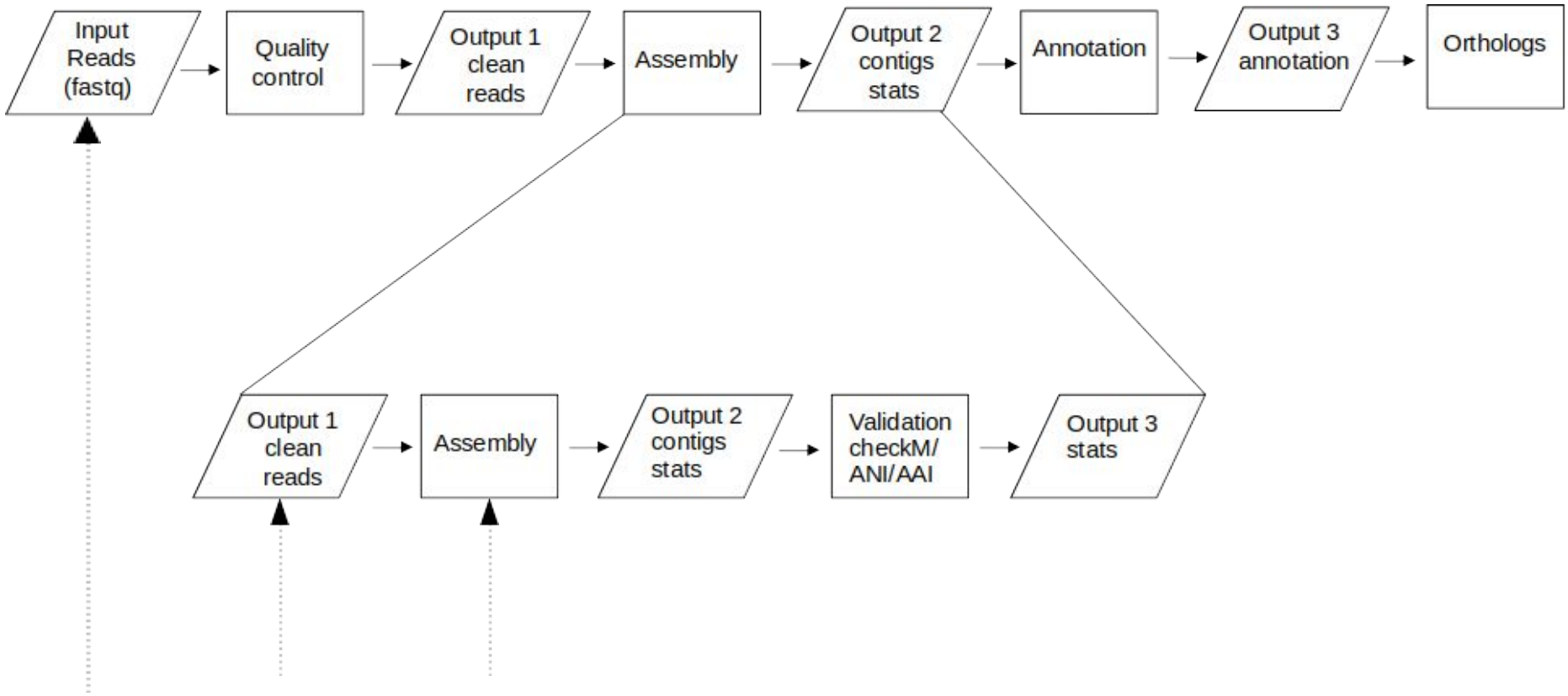
Assembly QC - flowgraph



Assembly QC - flowgraph



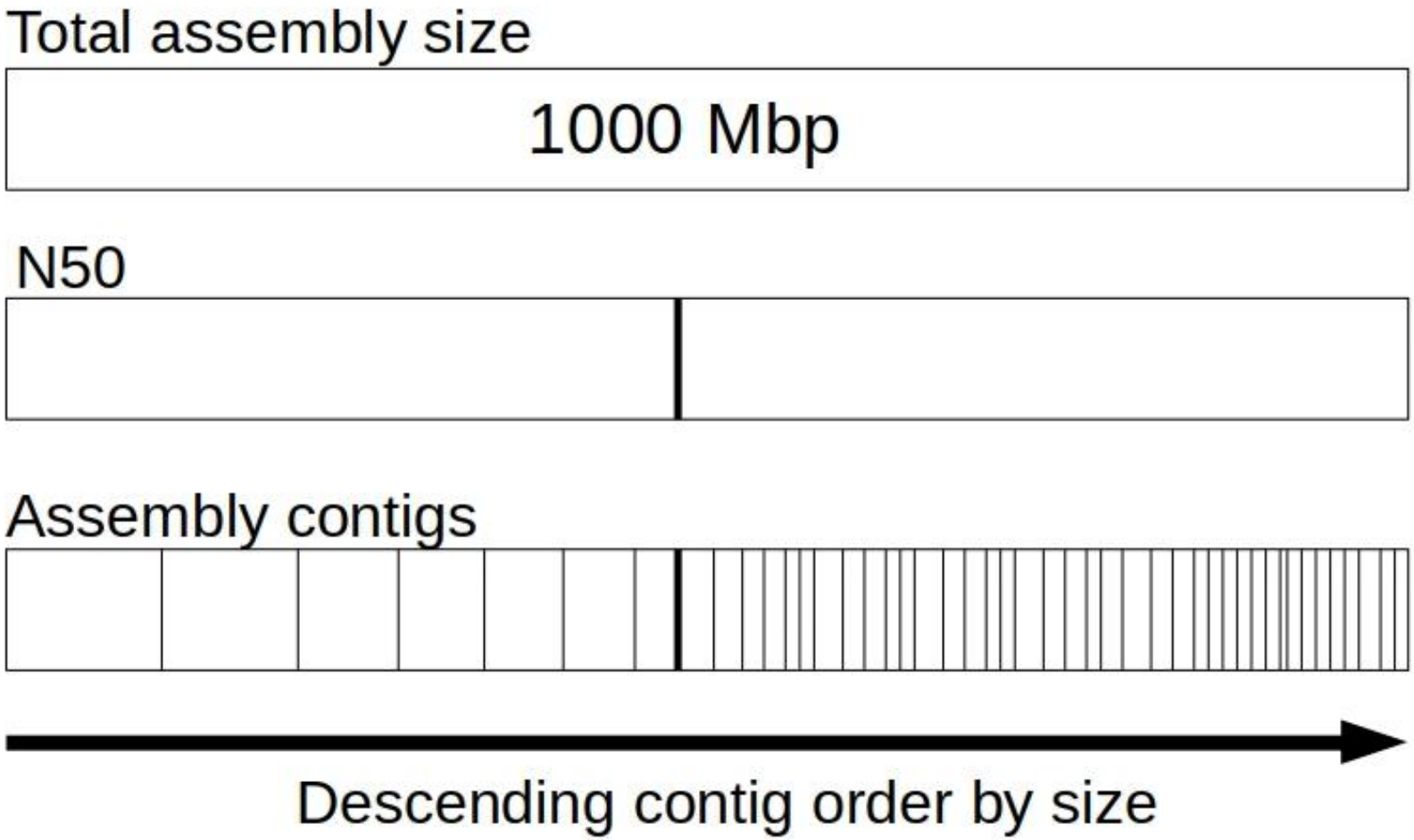
Assembly QC - flowgraph



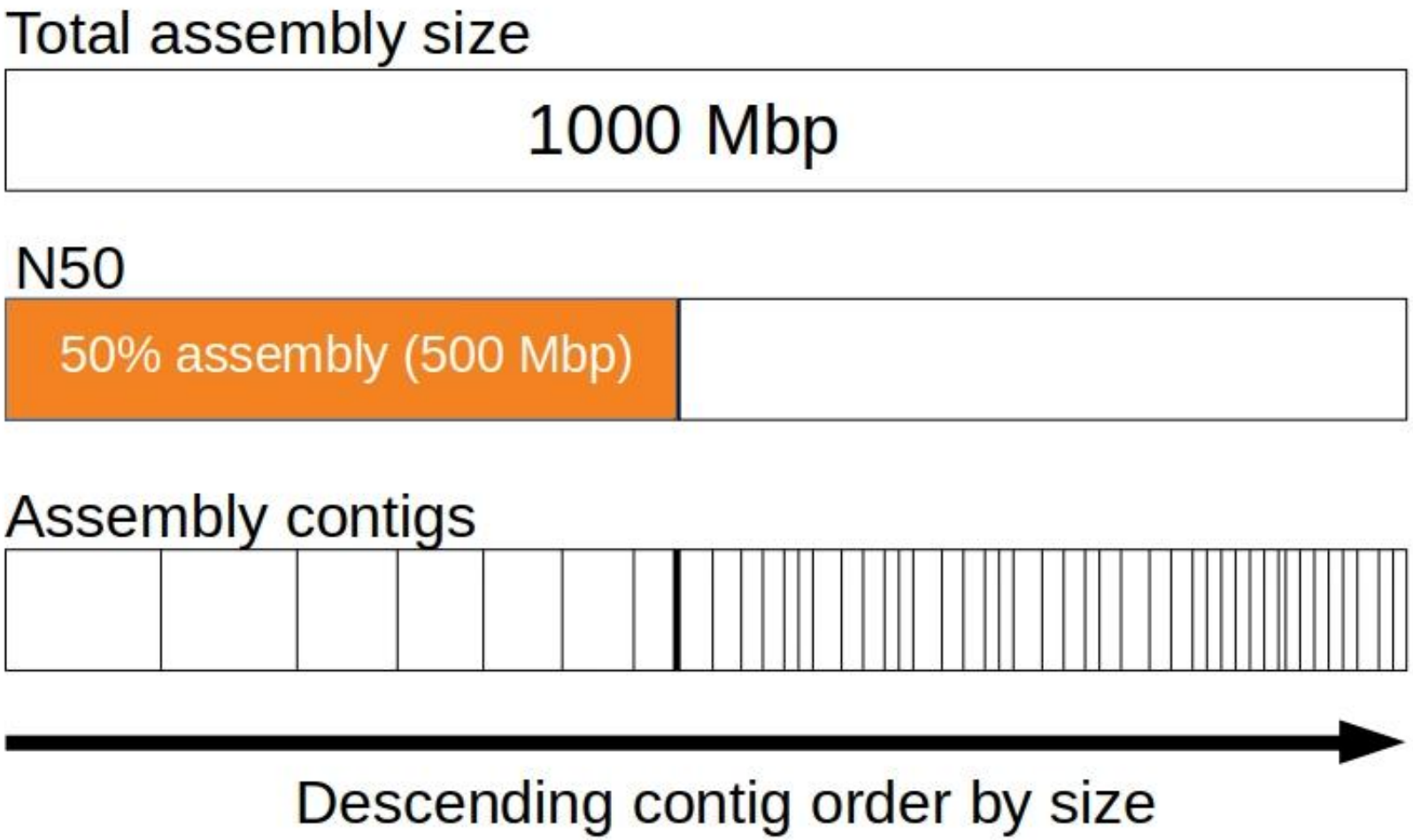
Assembly statistics

- No. of contigs and Max contig length
- N50
  - That 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value
- N90
  - Is the length for which the collection of all contigs of that length or longer contains at least 90% of the sum of the lengths of all contigs, and for which the collection of all contigs of that length or shorter contains at least 10% of the sum of the lengths of all contigs
- NG50
  - Similar to N50 but based on assembly size rather than the genome size. 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.
- L50/L90
  - Is defined as the smallest number of contigs whose length sum produces N50/N90

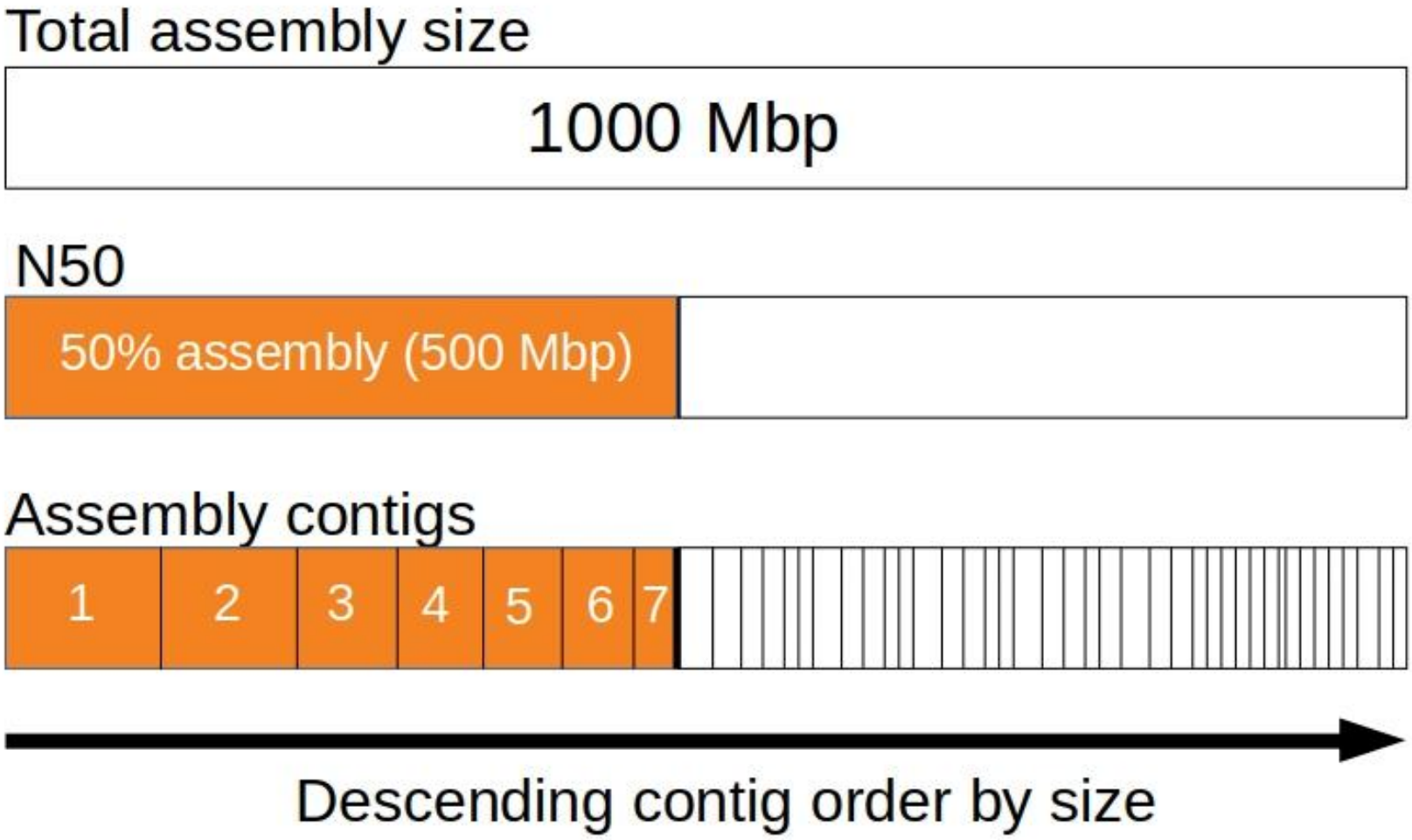
N50/L50 explained



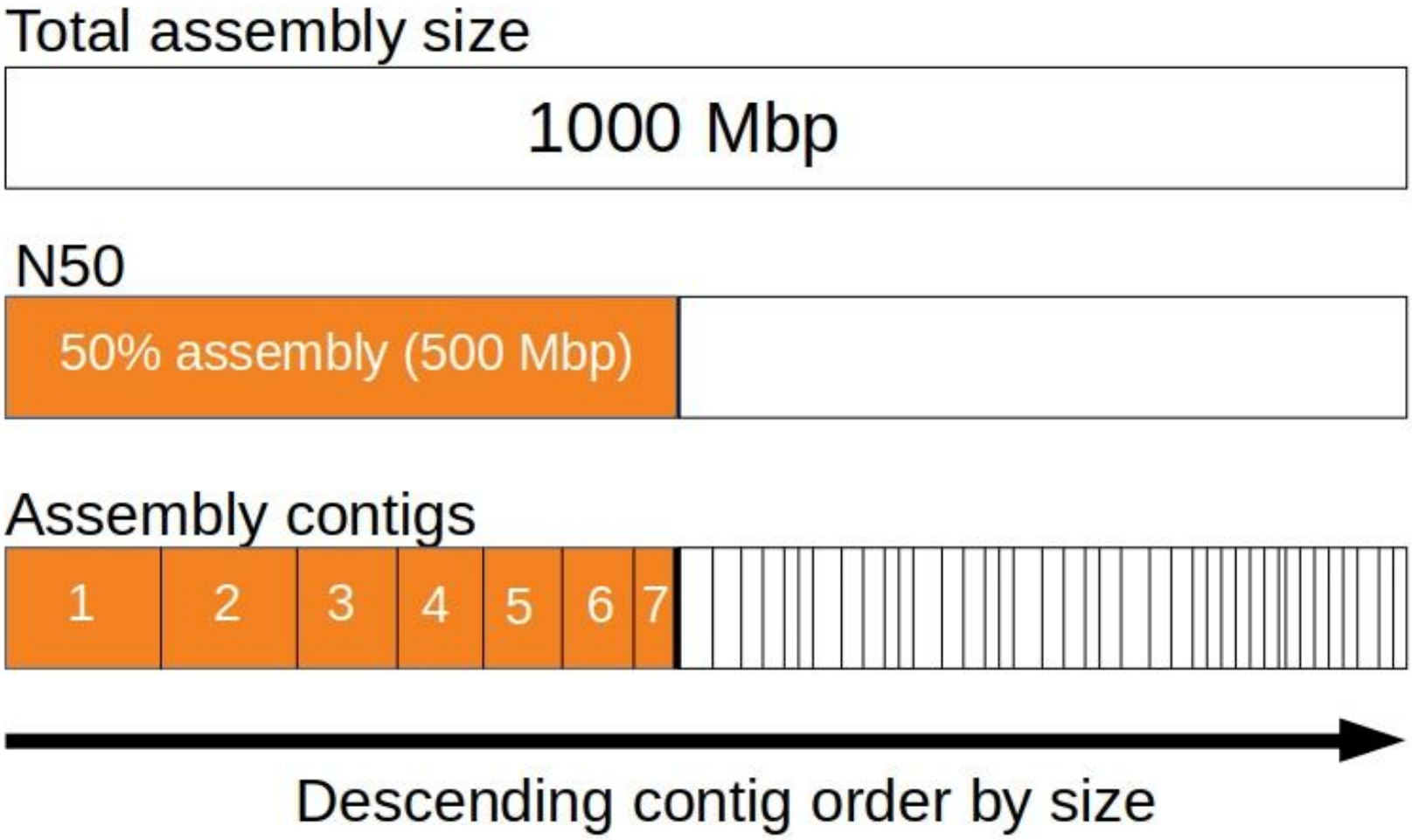
N50/L50 explained



N50/L50 explained

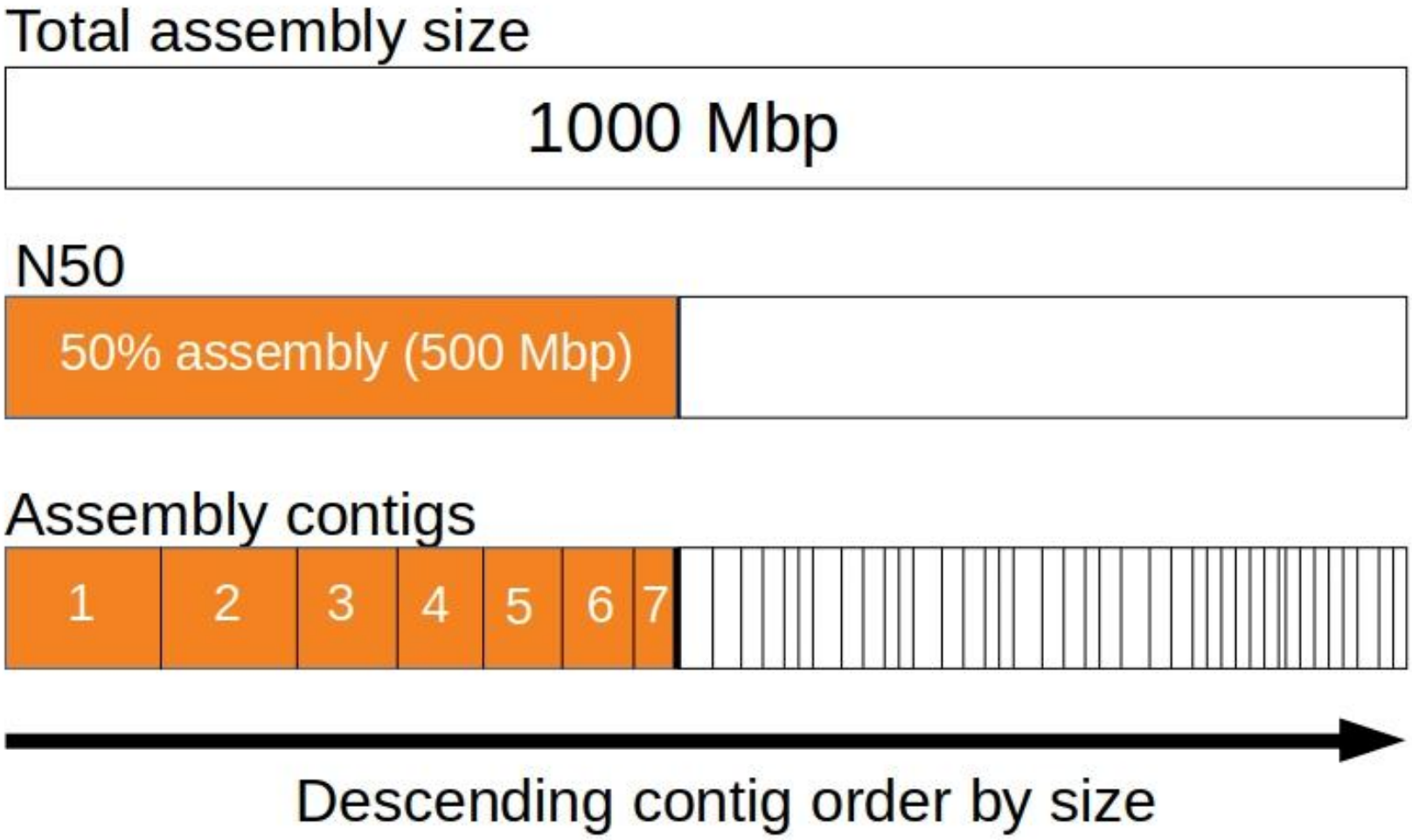


N50/L50 explained



N50 - 7 sequences, Avg contig length of those

N50/L50 explained

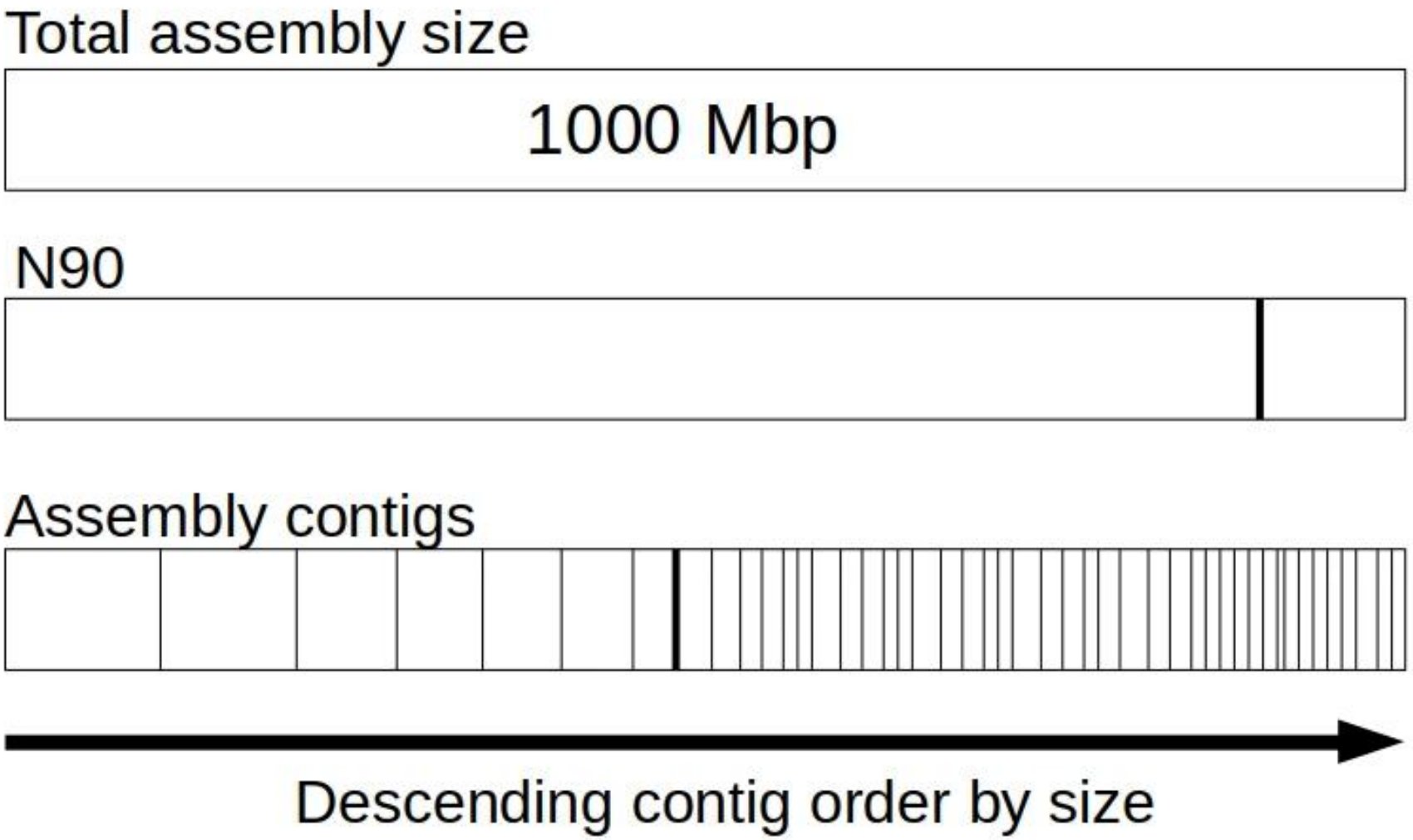


L50 - 50 Mbs

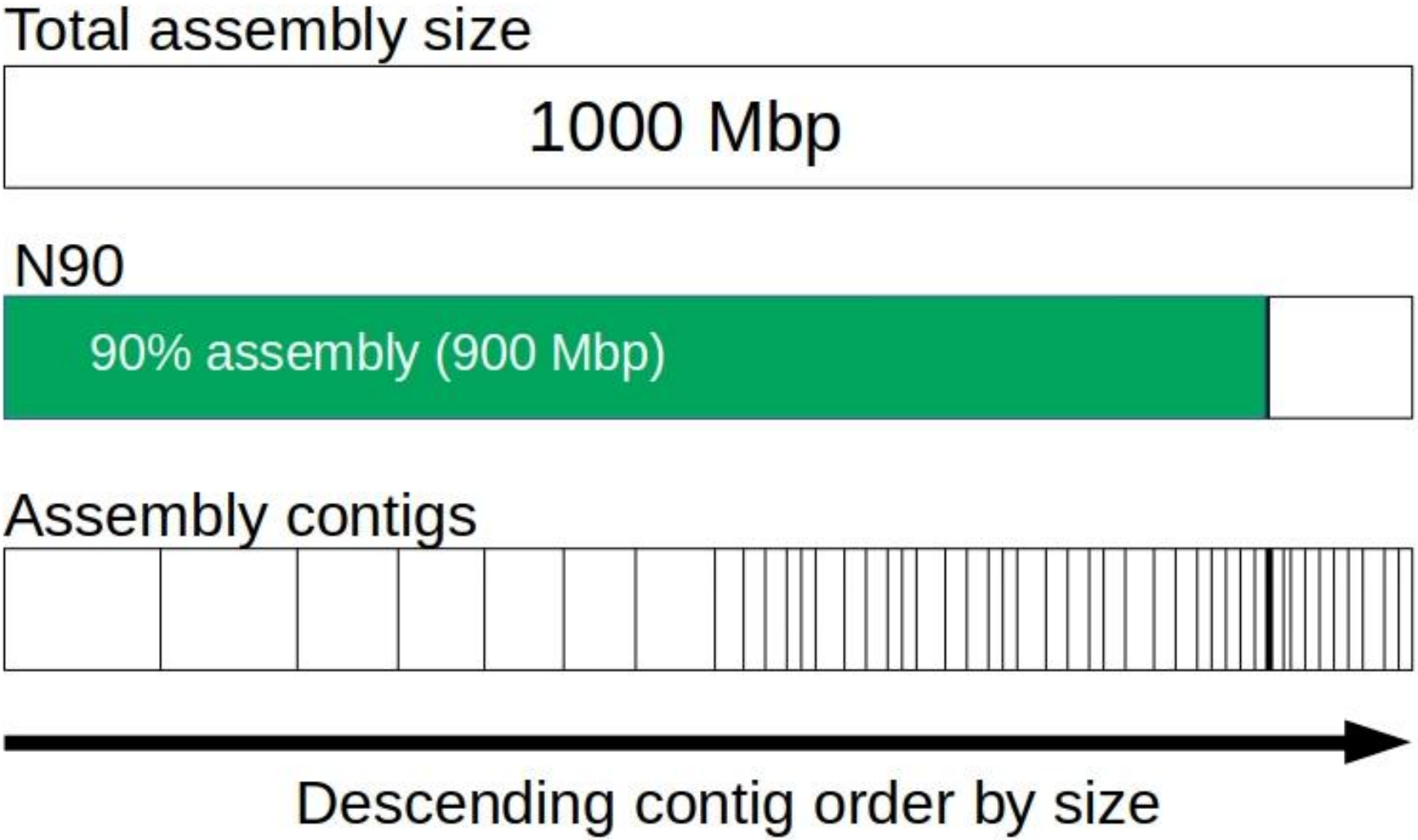
N50/L50 explained

What is the N50 for an *Rhizobium* strain with a complete genome 5 Mbp large?

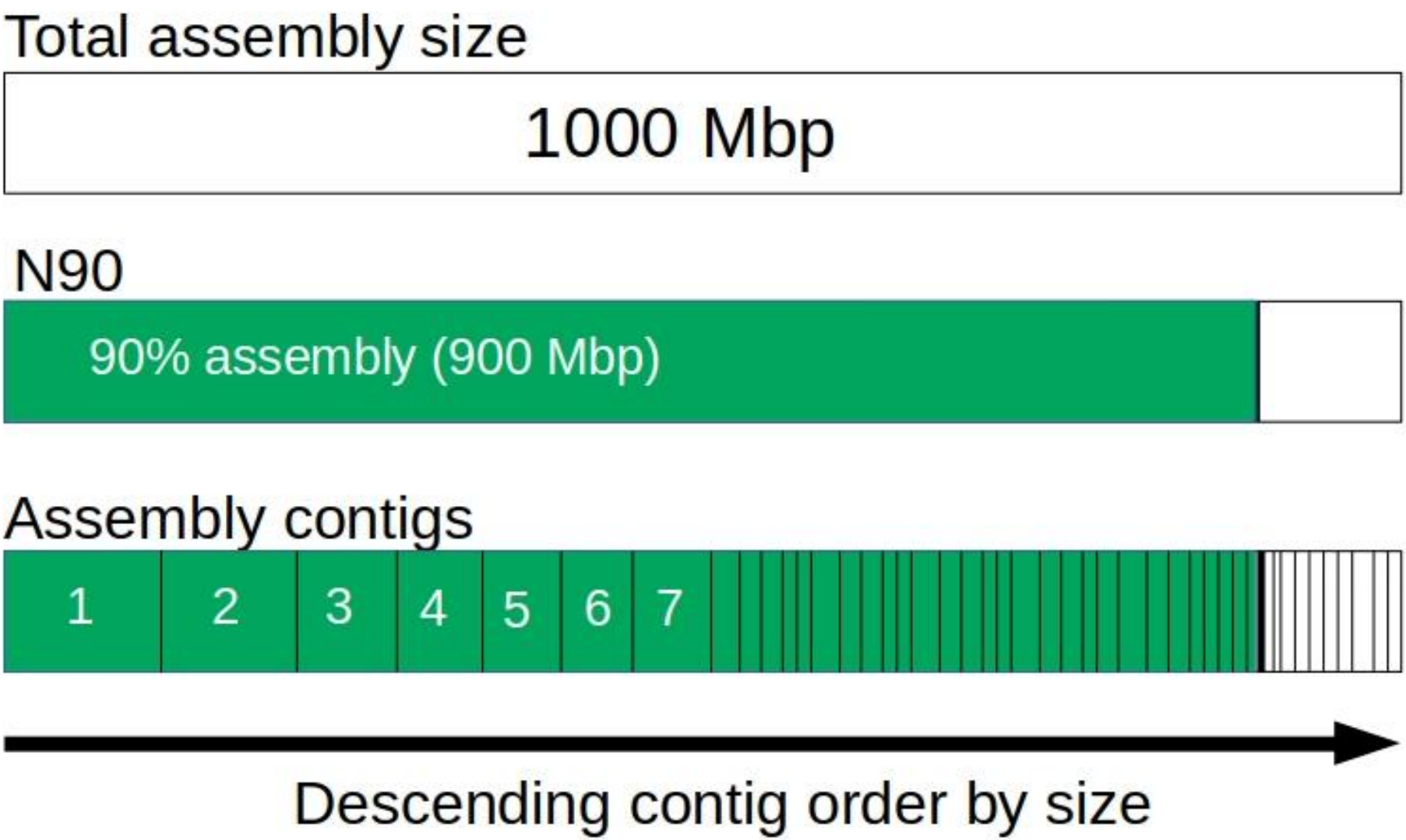
N90/L90 explained



N90/L90 explained

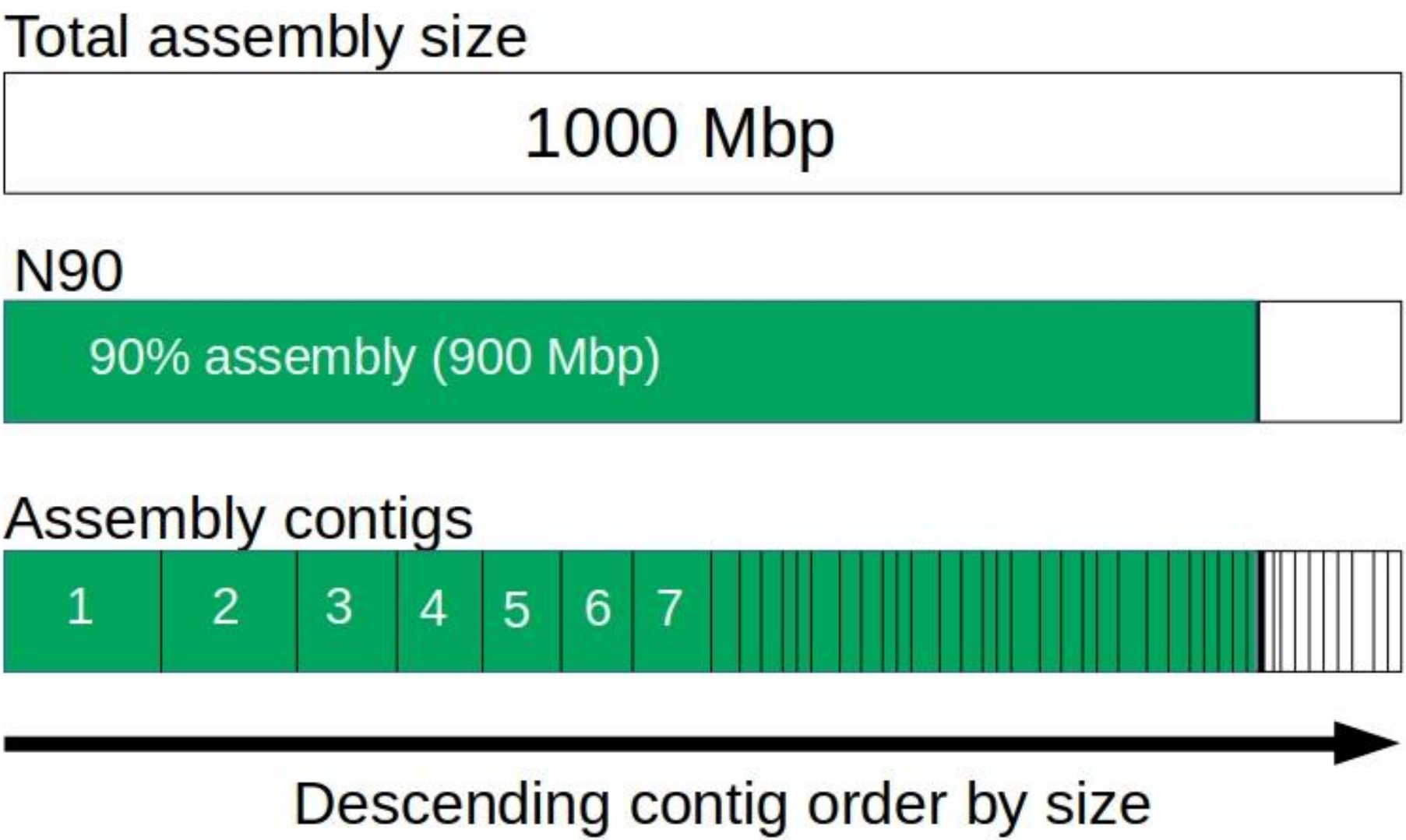


N90/L90 explained



N90 - 29 sequences, Avg contig length of those

N90/L90 explained

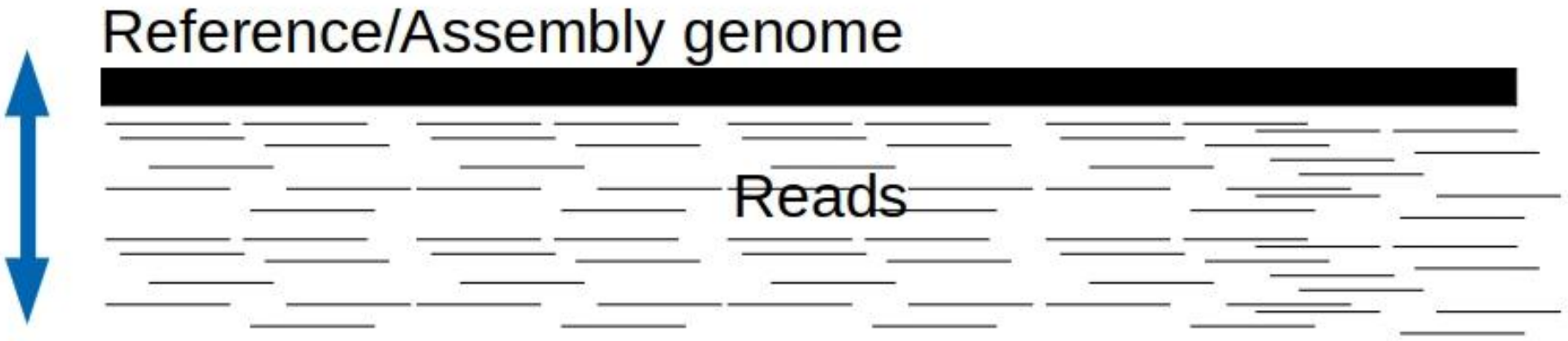


L90 - 12.5 Mbp

N90/L90 explained

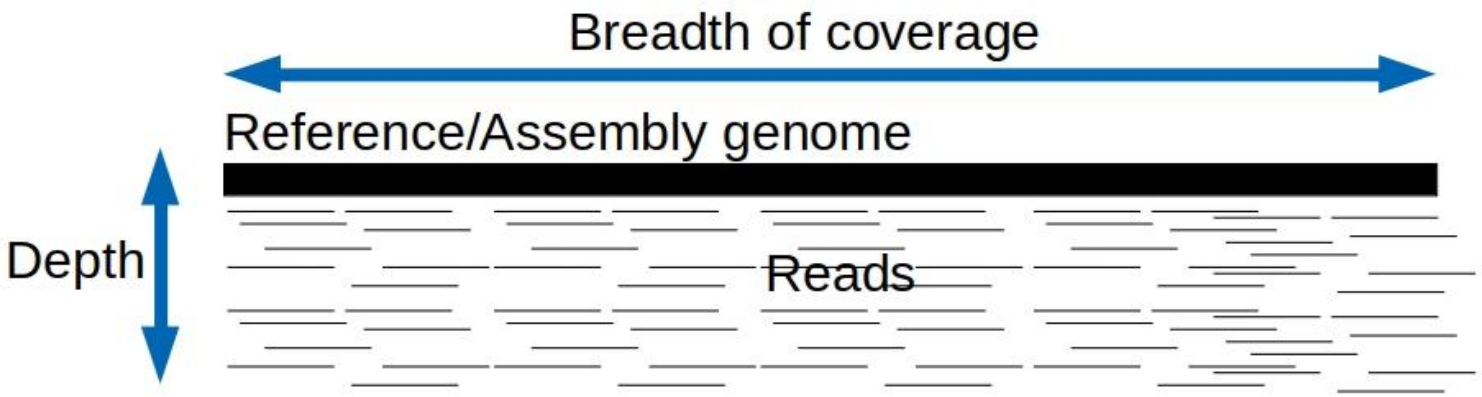
What is the N90 for an *Rhizobium* strain with a complete genome 5 Mbp large?

Genome coverage



Coverage = estimate of the average number of reads covering a single base across a genome

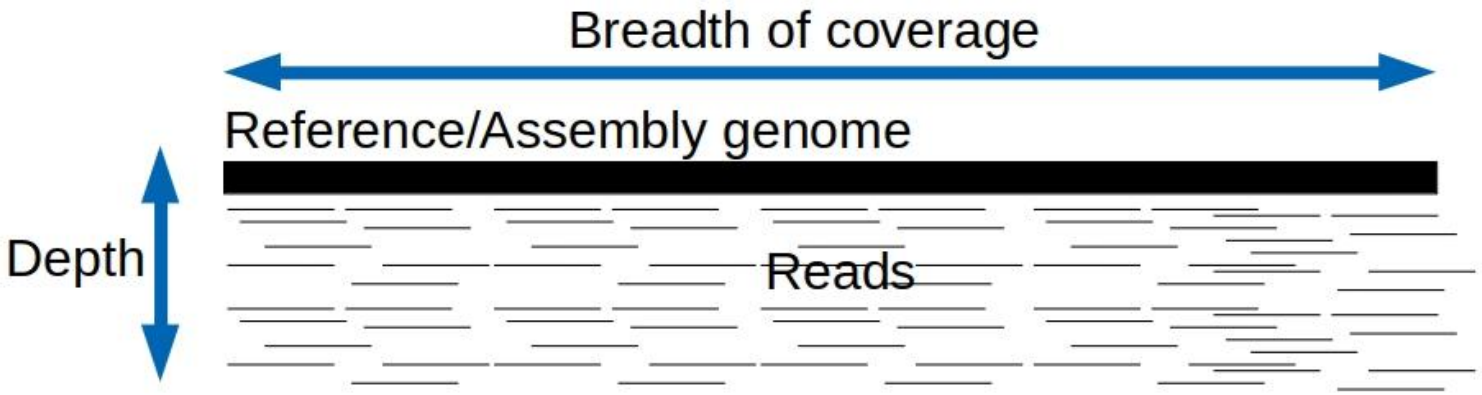
Genome coverage



$$\text{Avg coverage} = \frac{(\#\ \text{reads}) \times (\text{read length})}{\text{Size of genome}}$$

Depth of coverage = affected by the accuracy of genome alignment algorithms and by the uniqueness or the 'mappability' of sequencing reads within a target genome

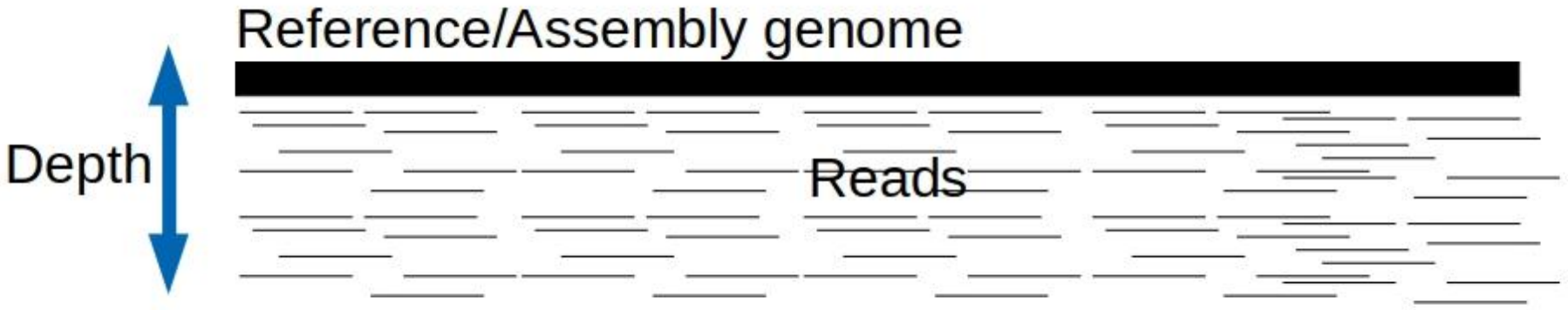
Genome coverage



$$\text{Avg coverage} = \frac{(\#\ \text{reads}) \times (\text{read length})}{\text{Size of genome}}$$

Breadth of coverage = the percentage of bases of a reference genome that are covered with a certain depth.

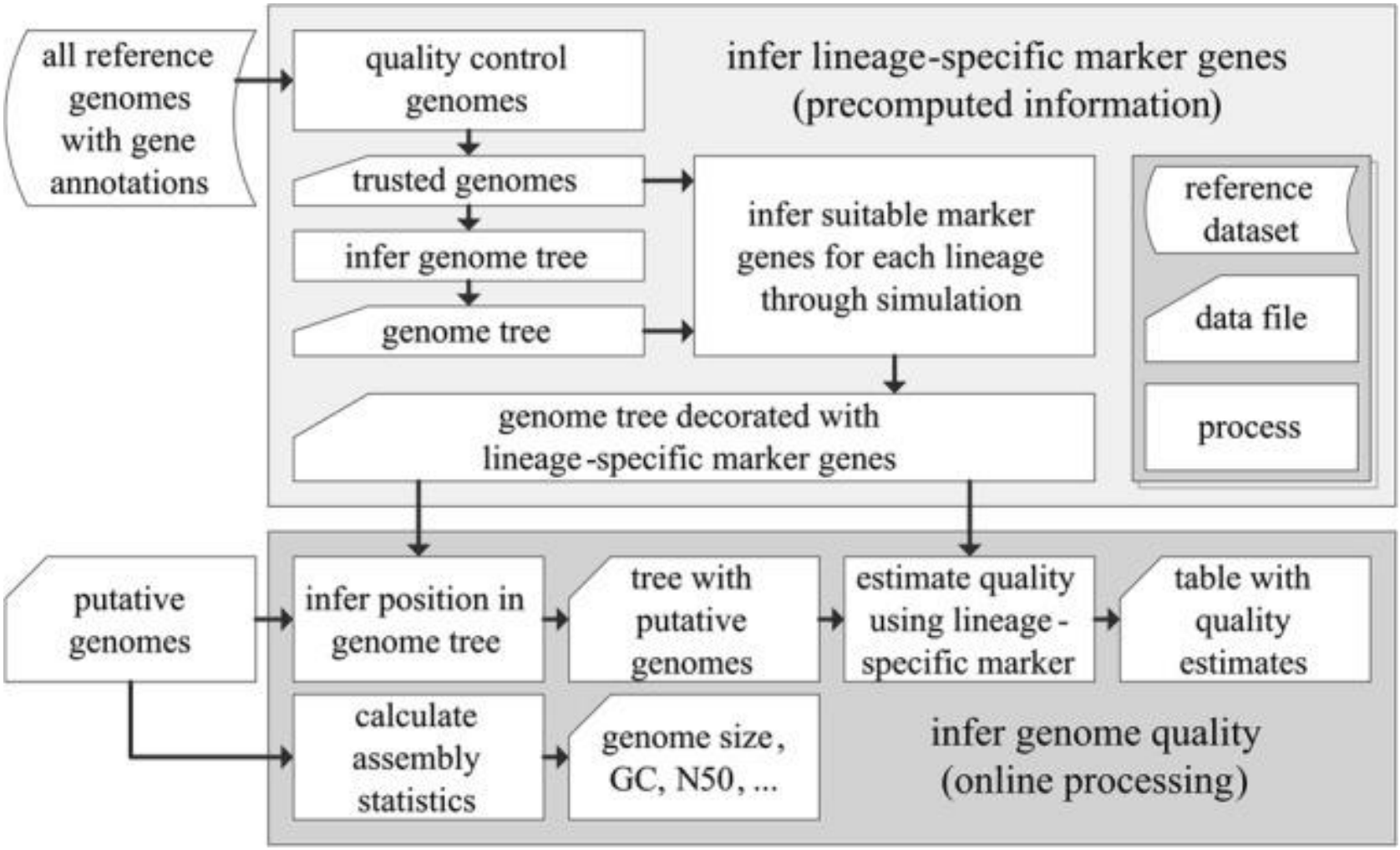
Genome coverage



$$\text{Avg coverage} = \frac{(\#\ \text{reads}) \times (\text{read length})}{\text{Size of genome}}$$

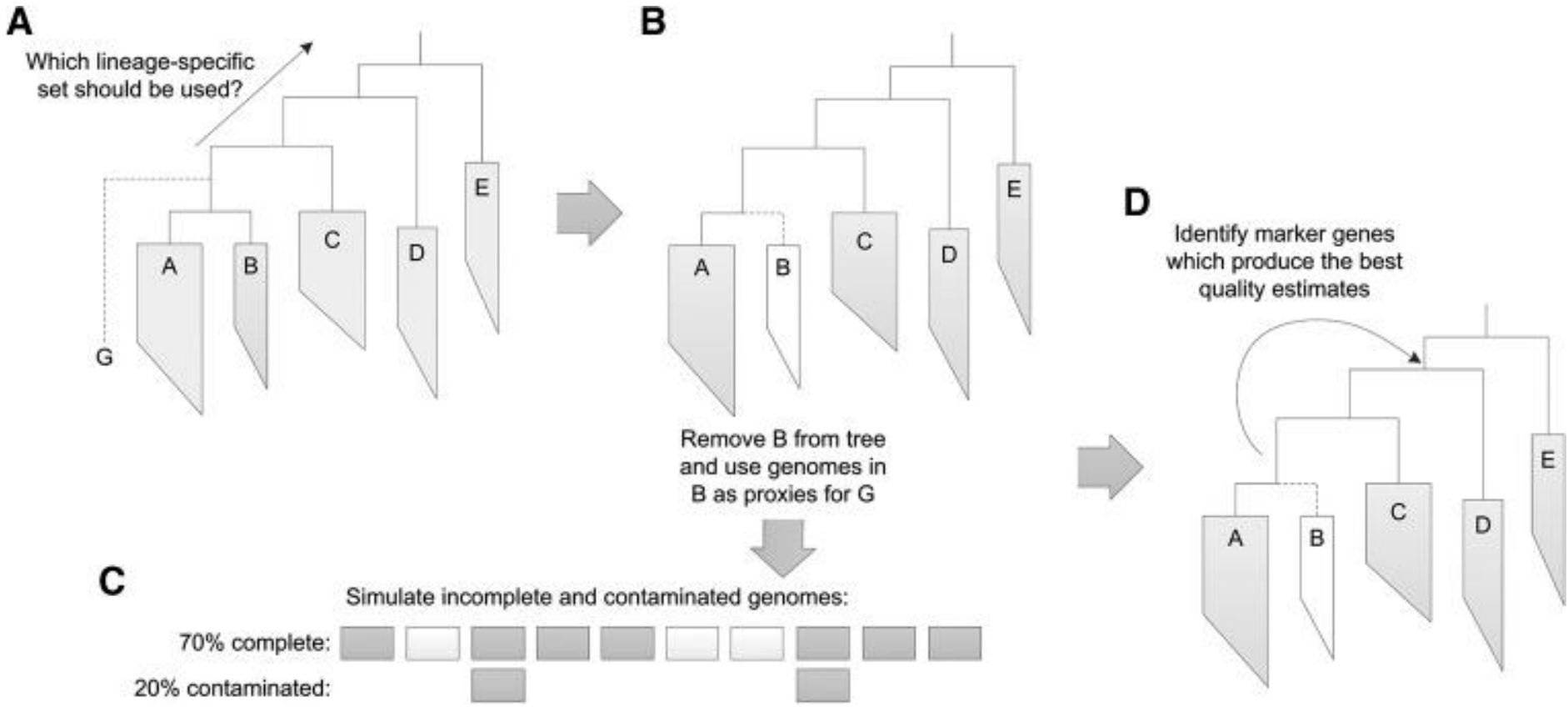
Standard human genome draft = 30x coverage
If done with Illumina for a bacterial genome read length matters thus:
150 bp paired end = >50x
250 bp paired end = >35x

Checking your genome (CheckM)



https://ecogenomics.github.io/CheckM/

Checking your genome (CheckM)



https://ecogenomics.github.io/CheckM/

Checking your genome (CheckM)

Table 3. Controlled vocabulary of draft genome quality based on estimated genome completeness and contamination

Completeness	Classification	Contamination	Classification
≥90%	Near	≤5%	Low*
≥70% to 90%	Substantial	5% to ≤10%	Medium
≥50% to 70%	Moderate	10% to ≤15%	High
<50%	Partial	>15%	Very high

(*) Genomes estimated to have 0% contamination can be designated as having "no detectable contamination".

https://ecogenomics.github.io/CheckM/

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.

Criterion	Description
Finished (SAG/MAG)	
Assembly quality[a]	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 or better
High-quality draft (SAG/MAG)	
Assembly quality[a]	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.
Completion[b]	>90%
Contamination[c]	<5%
Medium-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	≥50%
Contamination[c]	<10%
Low-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	<50%
Contamination[c]	<10%
This is a compressed set of genome reporting standards for SAGs and MAGs. For a complete list of mandatory and optional standards, see Supplementary Table 1.	

[a]Assembly statistics include but are not limited to: N50, L50, largest contig, number of contigs, assembly size, percentage of reads that map back to the assembly, and number of predicted genes per genome.

[b]Completion: ratio of observed single-copy marker genes to total single-copy marker genes in chosen marker gene set.

[c]Contamination: ratio of observed single-copy marker genes in ≥2 copies to total single-copy marker genes in chosen marker gene set.

https://www.nature.com/articles/nbt.3893/tables/1

Checking your genome (checkM)

Bin Id	Marker lineage	# genomes	Completeness	Contamination
049D-Bif_S2_1k	f__Paenibacillaceae (UID973)	32	92.04	0.34
113Ey_S1_1k	o__Actinomycetales (UID1593)	69	64.57	0.32
AE3022_S77_1k	g__Ensifer (UID3566)	27	98.42	0.79
AE3278P_all_1k	root (UID1)	5656	100	100

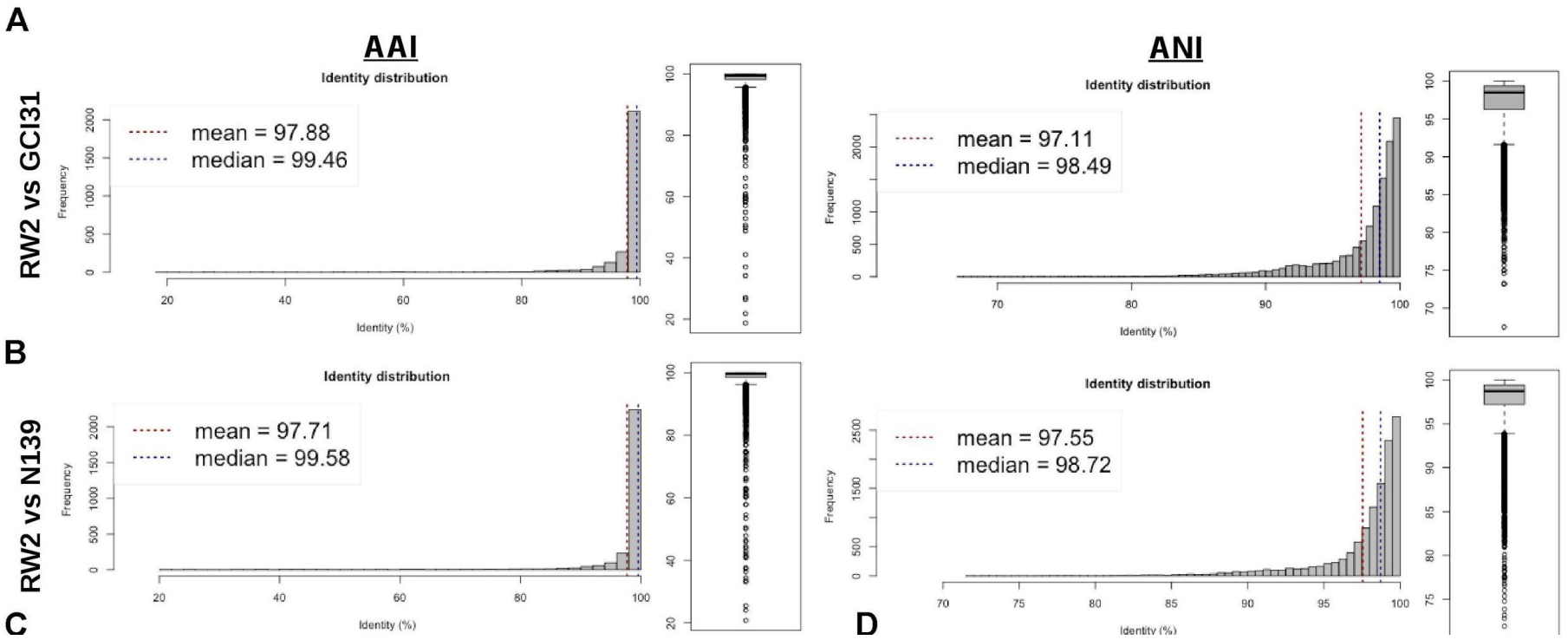
Good? Bad? Ugly?

https://ecogenomics.github.io/CheckM/

Species in bacteria ?

- Classically DNA-DNA hybridization (<70%)
  - Isolate DNA from two strains (one labeled and one unlabeled)
- Digital methods (ANI, AAI, dDDH)
  - resulting from pairwise genome comparisons and averaging the sequence identities of shared orthologous genes (amino acid or nucleotide, respectively).
- Average Nucleotide Identity (ANI)
  - >95% same species, <95% new species
  - <90% likely new genus use AAI to confirm
- Average Amino Acid Identity (AAI)
  - >95% same species, <95% new species
  - <90% new genus
- Digital DNA-DNA hybridization (dDDH)
  - >70% same species, <70% new species
  - <79% new subspecies, >79% same subspecies

Species in bacteria ?



IS RW2 the same as GCI31 (a) ?
IS RW2 the same as N139 (b)?

White 3rd et al., 2019.
https://www.frontiersin.org/articles/10.3389/fmicb.2018.03189/full

What is BLAST?

BLAST - the first annotator

J. Mol. Biol. 1990 Oct 5;**215**(3):403–10 —the primary reference for the BLAST algorithm.

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]*National Center for Biotechnology Information*
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]*Department of Computer Science*
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]*Department of Computer Science*
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straight-forward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

1. Introduction

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the DNA and amino acid sequence databases continue to grow in size they become increasingly useful in the analysis of newly sequenced genes and proteins because of the greater chance of finding such homologies. There are a number of software tools for

(Coulson *et al.*, 1987).

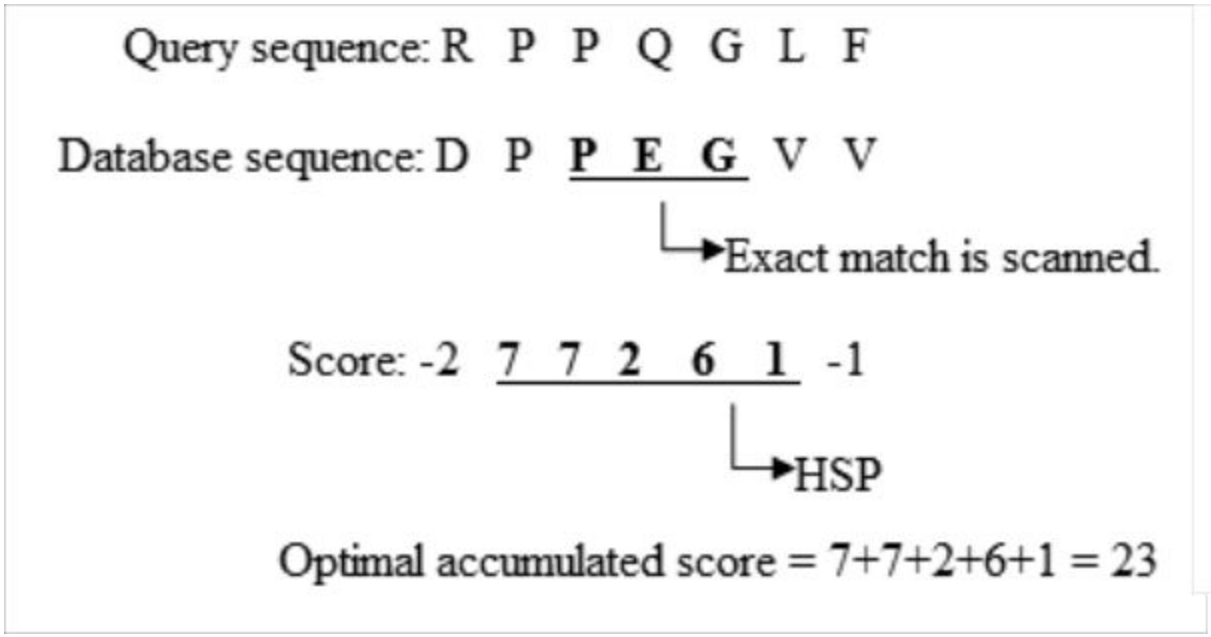
Rapid heuristic algorithms that attempt to approximate the above methods have been developed (Waterman, 1984), allowing large databases to be searched on commonly available computers. In many heuristic methods the measure of similarity is not explicitly defined as a minimal cost set of mutations, but instead is implicit in the algorithm itself. For example, the FASTP program (Lipman & Pearson, 1985; Pearson & Lipman, 1988) first finds locally similar regions between two sequences

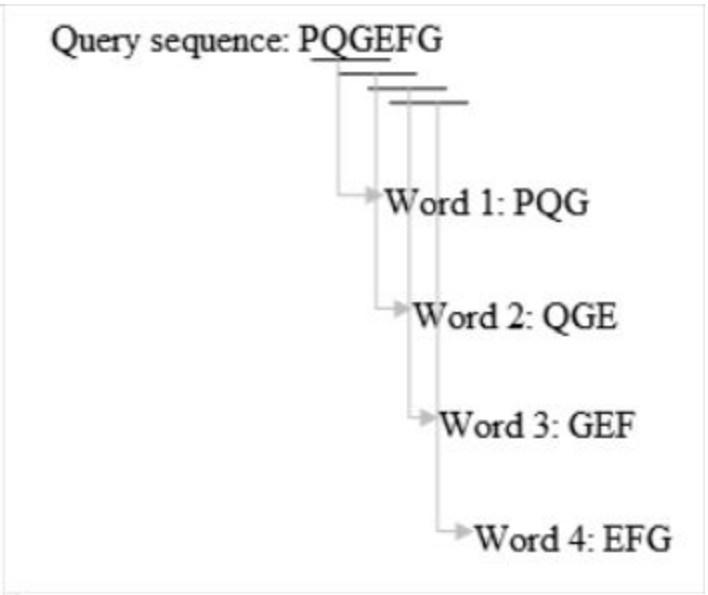
Most important and first major computational biology tool ever created…
Cited over 50,000 times...

Basic Local Alignment Search Tool - https://blastalgorithm.com/

BLAST - the first annotator

Basic Local Alignment Search Tool



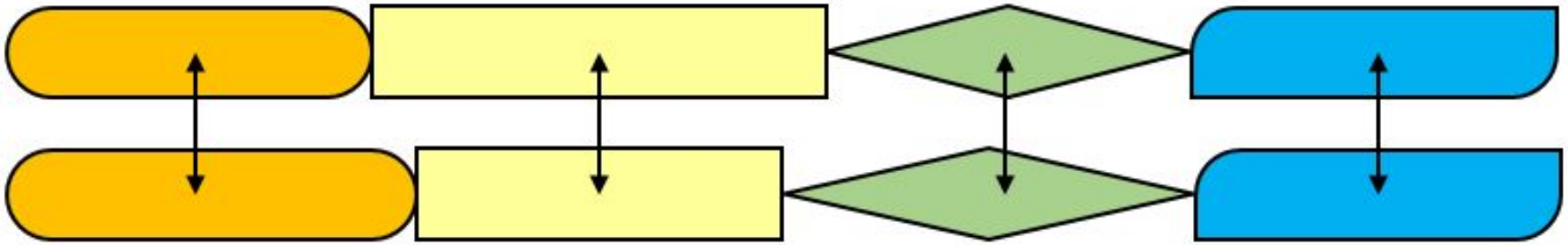


BLAST - the first annotator

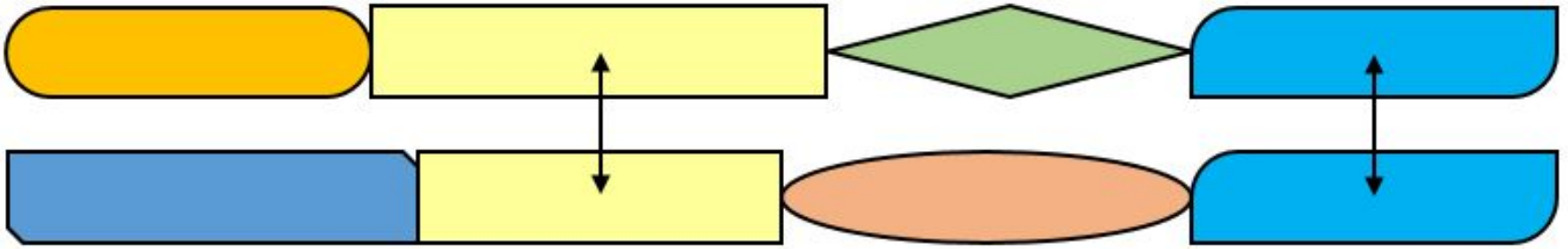
Basic Local Alignment Search Tool

- Nucleotide Blast (MegaBlast, BlastN, nuc-nuc, db-query)
- BlastX (translated nucleotide query to protein database)
- tBlastN (protein query to translated nucleotide database)
- tBlastX (translated nucleotide query and database)
- Blastp (protein query and database)

Alignment (Global vs. Local)



Global Alignment



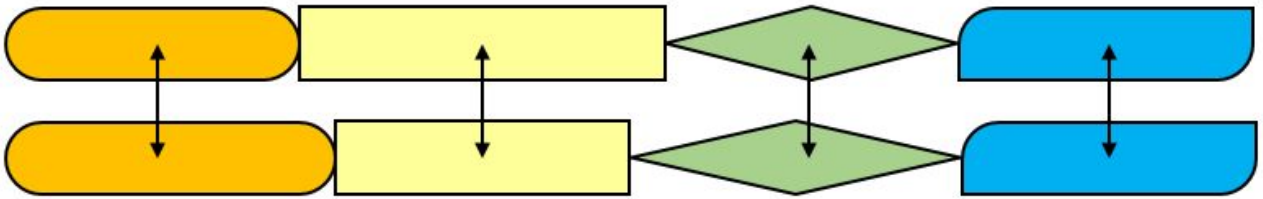
Local Alignment

Alignment (Global vs. Local)

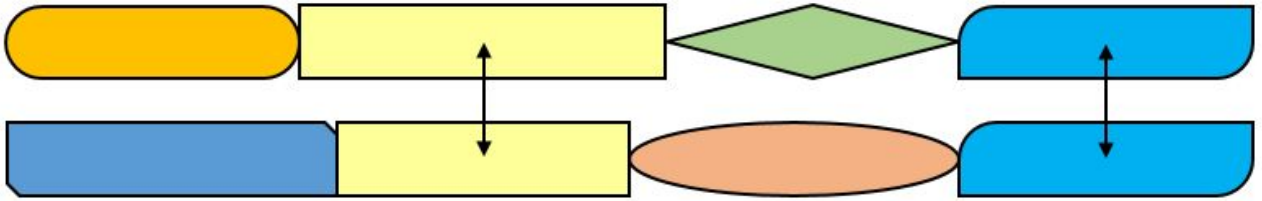
-Pre-blast-

Smith-Waterman algorithm (Local alignment)

Needleman–Wunsch algorithm (Global alignment)



Global Alignment



Local Alignment

Alignment (Global vs. Local)

-Pre-blast-

Smith-Waterman algorithm (Local alignment)

Needleman–Wunsch algorithm (Global alignment)

Basic differences between SW and NW

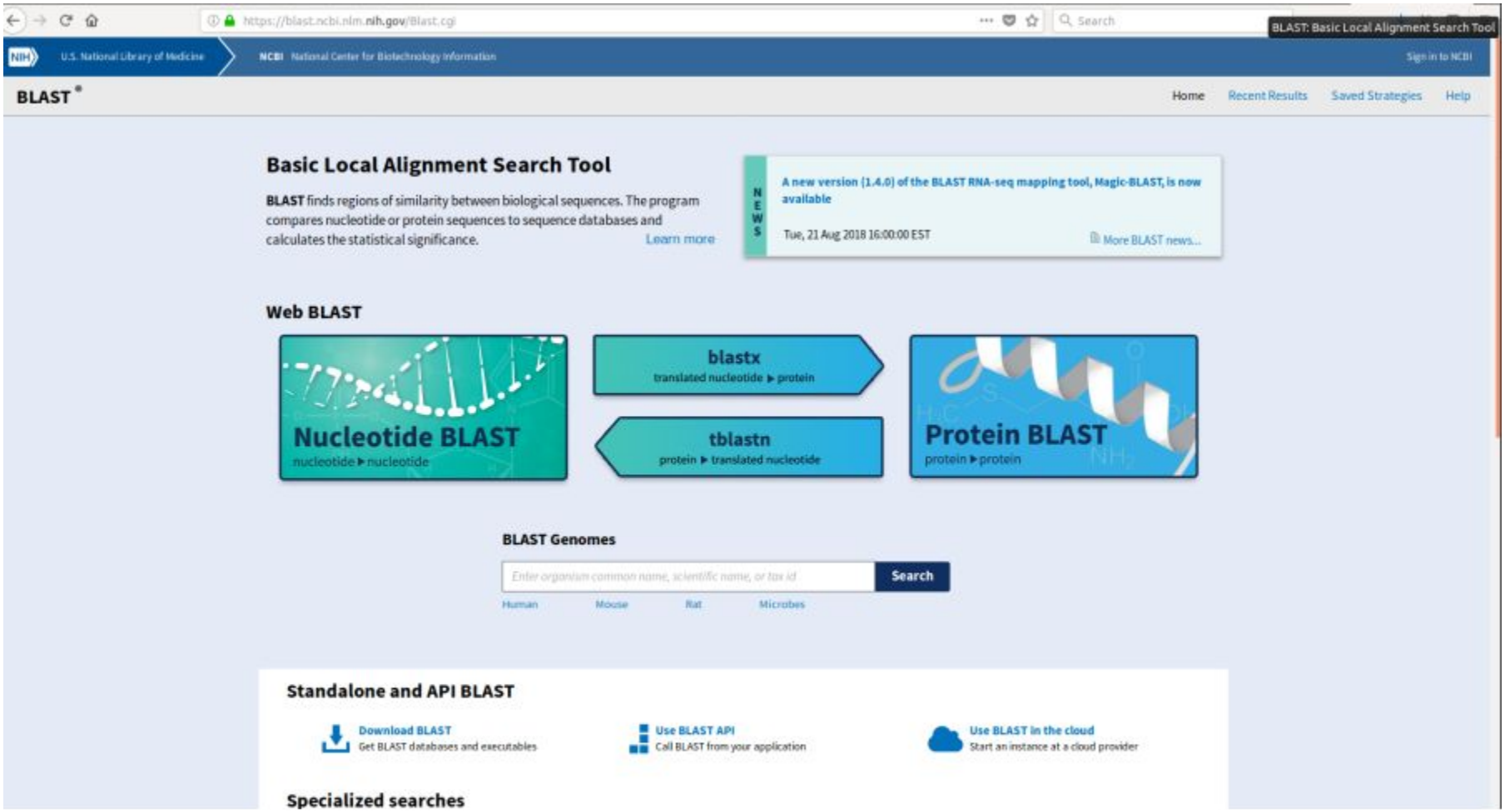
	Smith-Waterman algorithm	Needleman-Wunsch algorithm
Initialization	First row and first column are set to 0	First row and first column are subject to gap penalty
Scoring	Negative score is set to 0	Score can be negative
Traceback	Begin with the highest score, end when 0 is encountered	Begin with the cell at the lower right of the matrix, end at top left cell

Alternatives to BLAST

-Post-blast-

1. BLAT (Blast-like alignment tool)
2. PattenHunter
3. LAST (Local alignment search tool)
4. KLAST
5. Sword (awesome, both fast SW and NW)
6. USEARCH
7. MMseq2
8. DIAMOND
9. Bowtie2, BWA
10. Hmmer (based on HMM)

BLAST



Go to website - https://www.ncbi.nlm.nih.gov/BLAST/