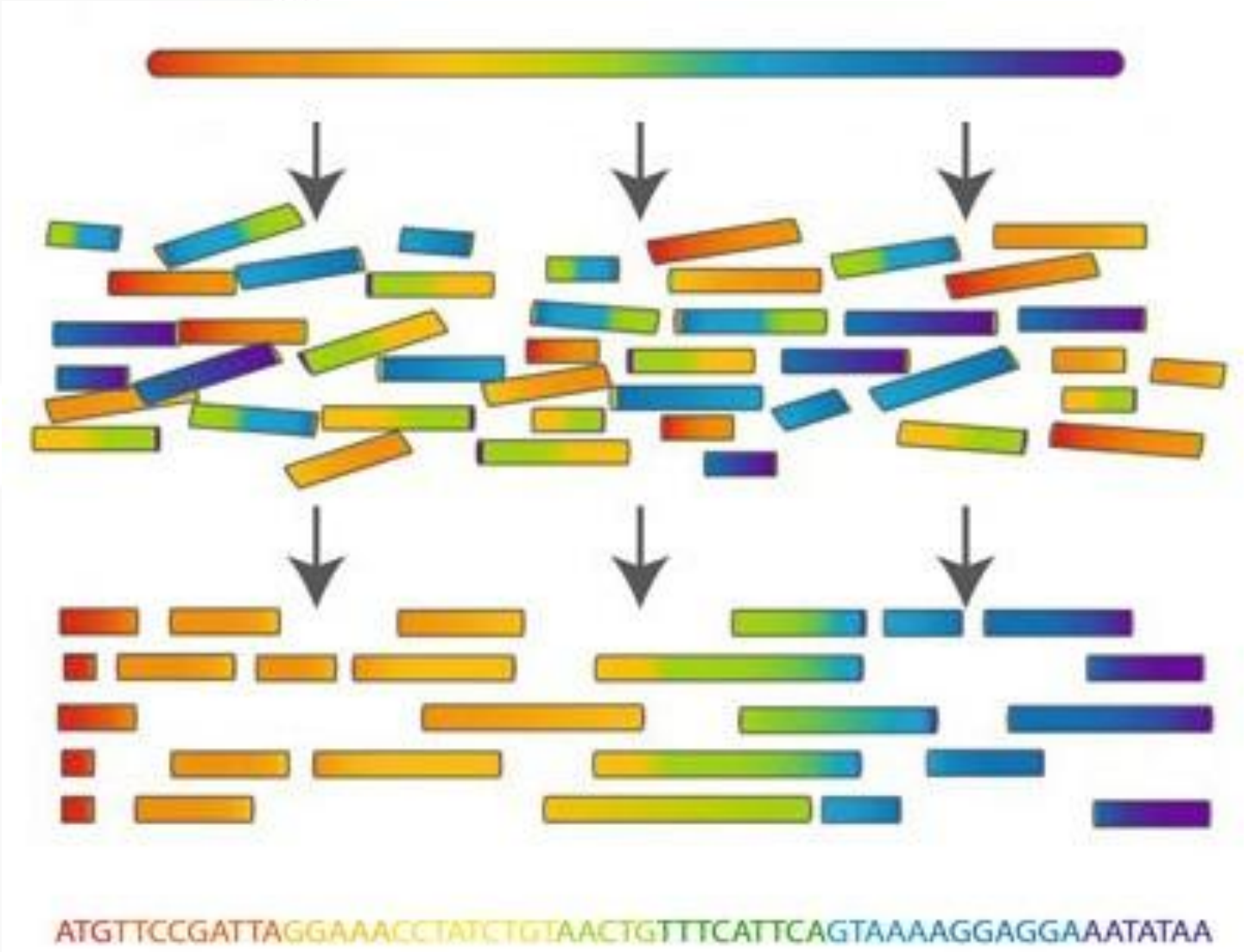

# Genome Assembly Team 3: Background and Strategy

Maddala Aparna
Yang Ruize
Kundnani Deepali
Xiao Yiqiong
Singu Swetha Gowri

January 28, 2020

# What is genome assembly?




genome sequence

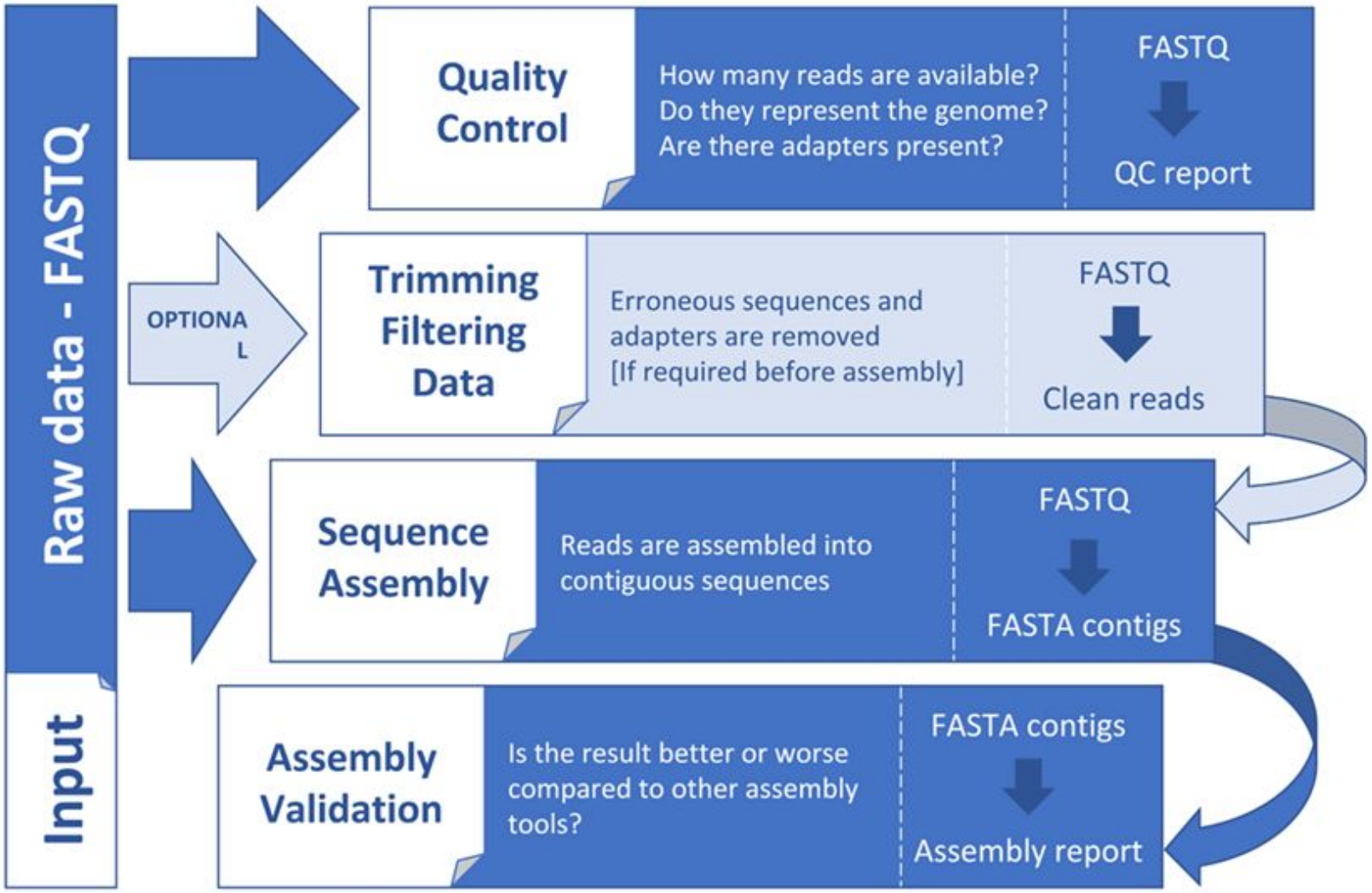
reads (output of sequencing)

alignment

contigs

Adapted from: Commins, Jennifer et al. “Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects.” Biological Procedures vol. 11 52-78. 11 Mar. 2009.

# Steps of Genome Assembly




Dominguez Del Angel V, Hjerde E, Sterck L et al. Ten steps to get started in Genome Assembly and Annotation [version 1]. F1000Research 2018, 7:148 (doi: 10.12688/f1000research.13598.1)

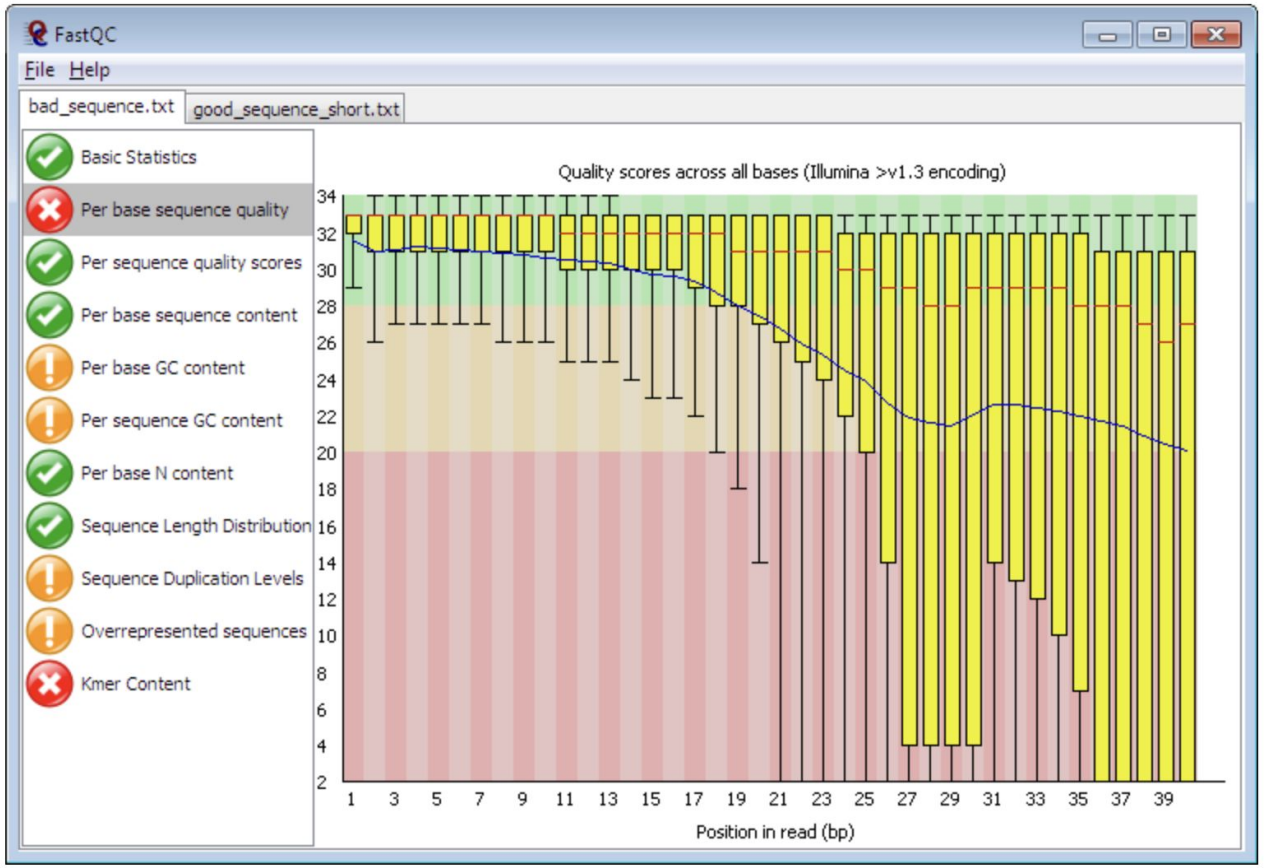
# Quality Control for Raw Reads

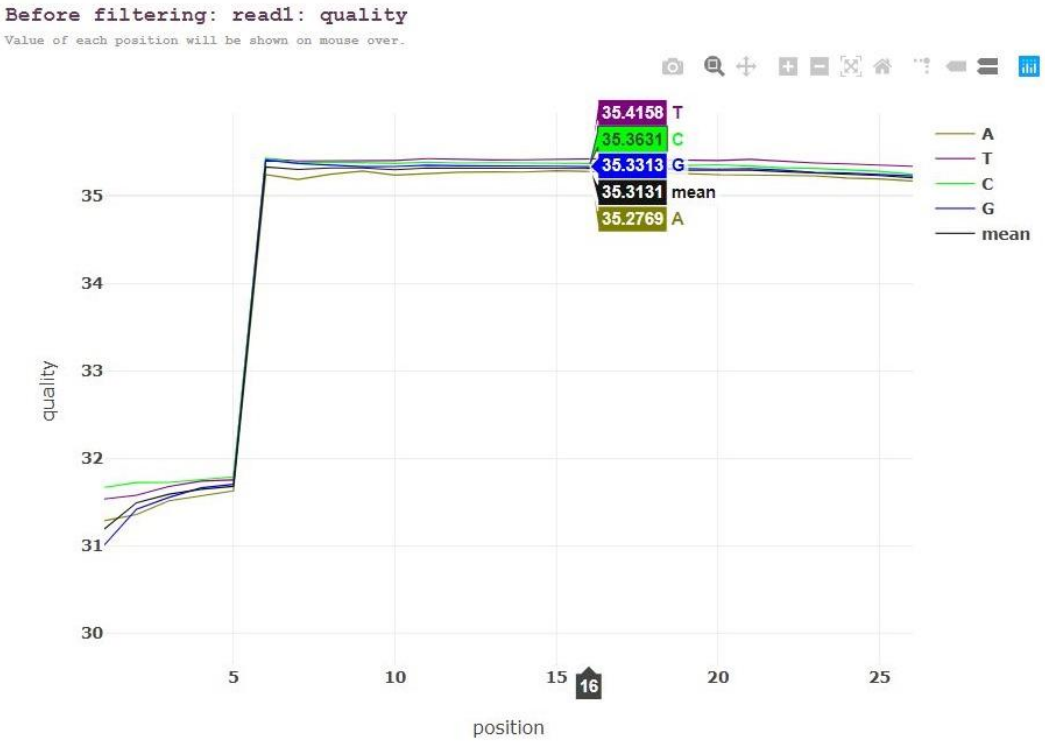
- Tools:
  - FastQC & Trimmomatic
    - FastQC - quality control on raw sequence reads
    - Trimmomatic - flexible trimming tool designed for Illumina NGS data
  - fastp
    - A tool designed to provide fast all-in-one preprocessing for FastQ files
    - integrates trimming and QC into a single tool
    - Quality control features specifically designed for paired-end data
  - MultiQC
    - consolidates multiple quality control reports generated by FastQC or fastp

# Pros & Cons

| | fastp | FastQC, Trimmomatic |
|---|---|---|
| Reliability | more recent | cited often, industry standard |
| Suitability | paired-end base correction | treats paired reads separately |
| Speed | faster than FastQC | runs fairly quickly |
| Information provided | interactive plots allow us to zoom in on regions of interest | generates per-base sequence quality scores with boxplots |

# FastQC vs. fastp

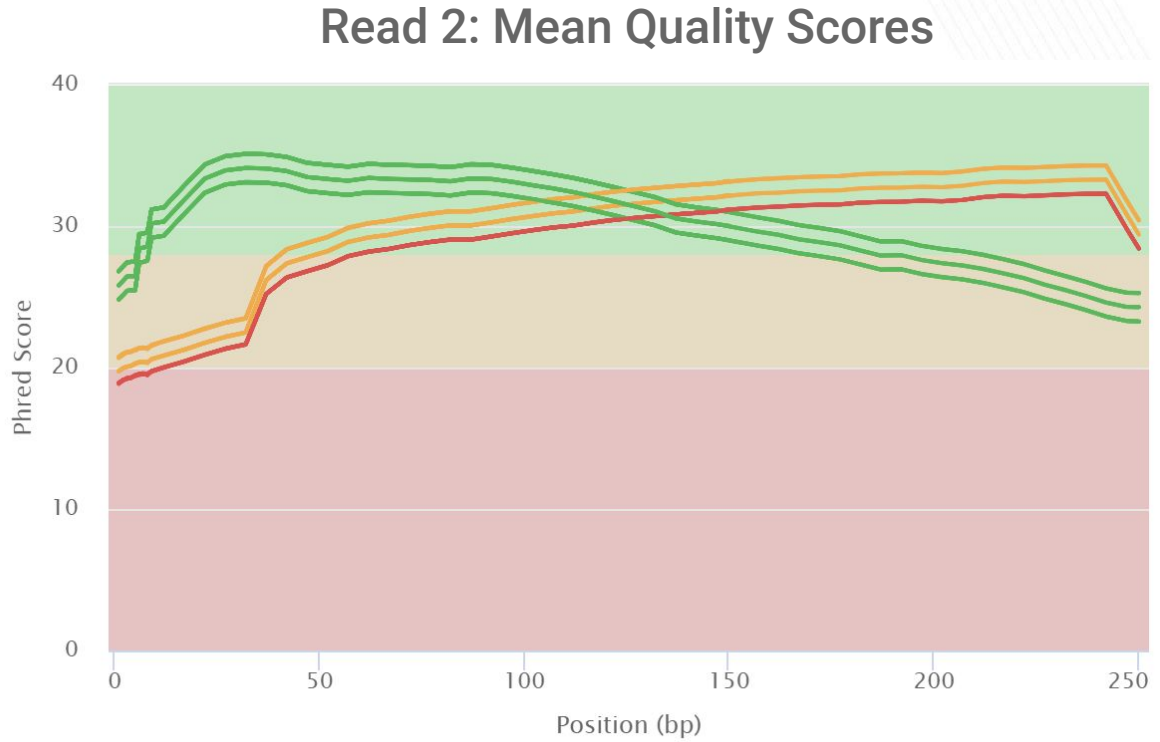





Georgia Tech
CREATING THE NEXT

# Approach to Trimming


Georgia Tech
CREATING THE NEXT

# Assembly

## Reference based assembly

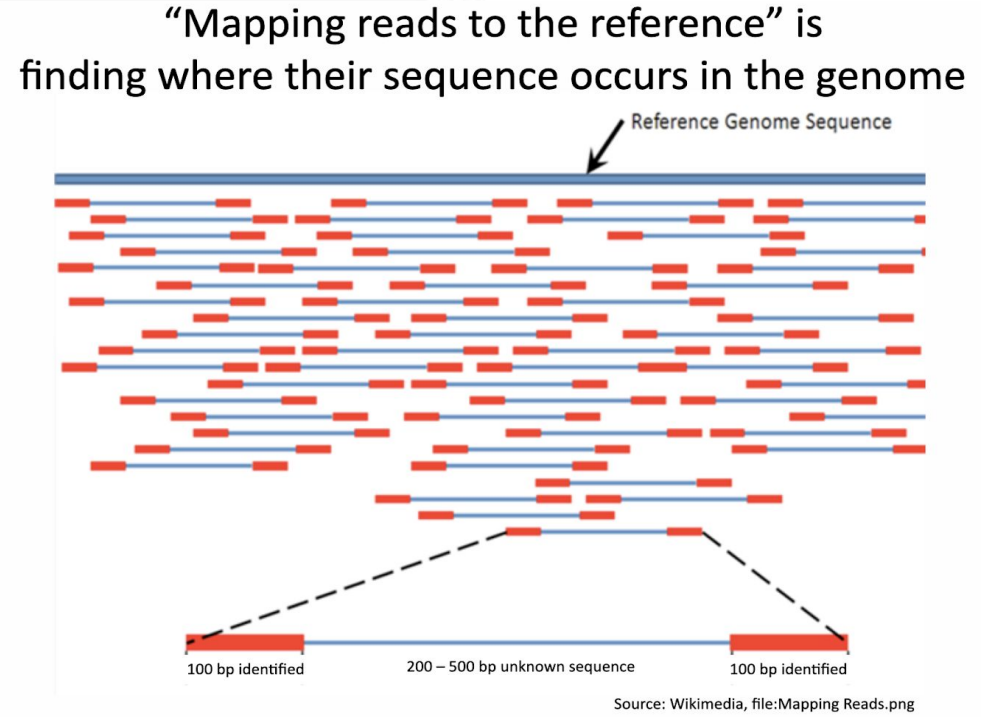



Image from :
https://www.mn.uio.no/ifi/studier/masteroppgaver/bmi/benchmarking-system.html

## De novo assembly

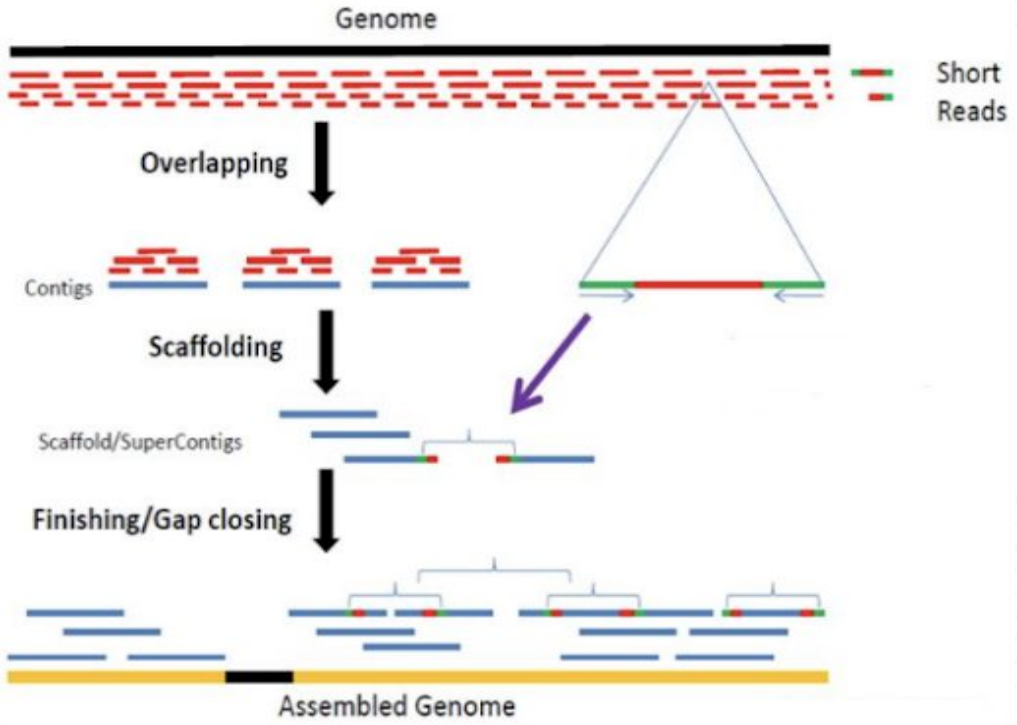



Image from: https://www.arraygen.com/De-novo-Assembly.php

# Algorithms

## OLC Graph

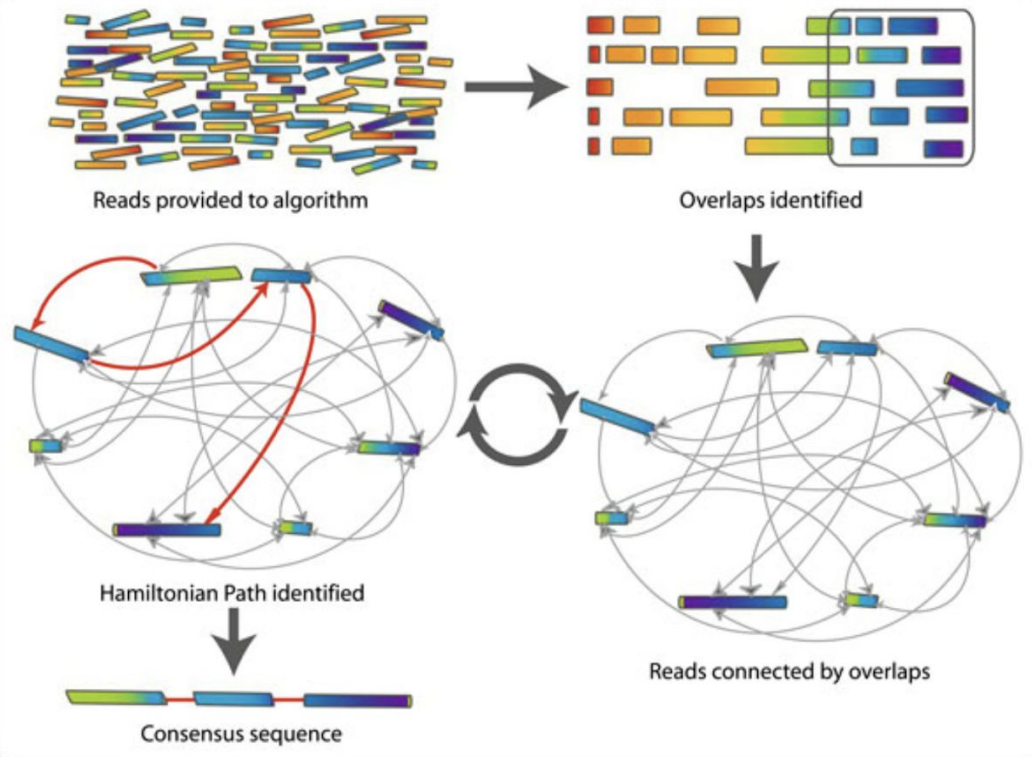



Image from: https://www.researchgate.net/figure/Overlap-layout-consensus-genome-assembly-algorithm-Reads-are-provided-to-the-algorithm_fig2_26266221

## De Bruijn Graph

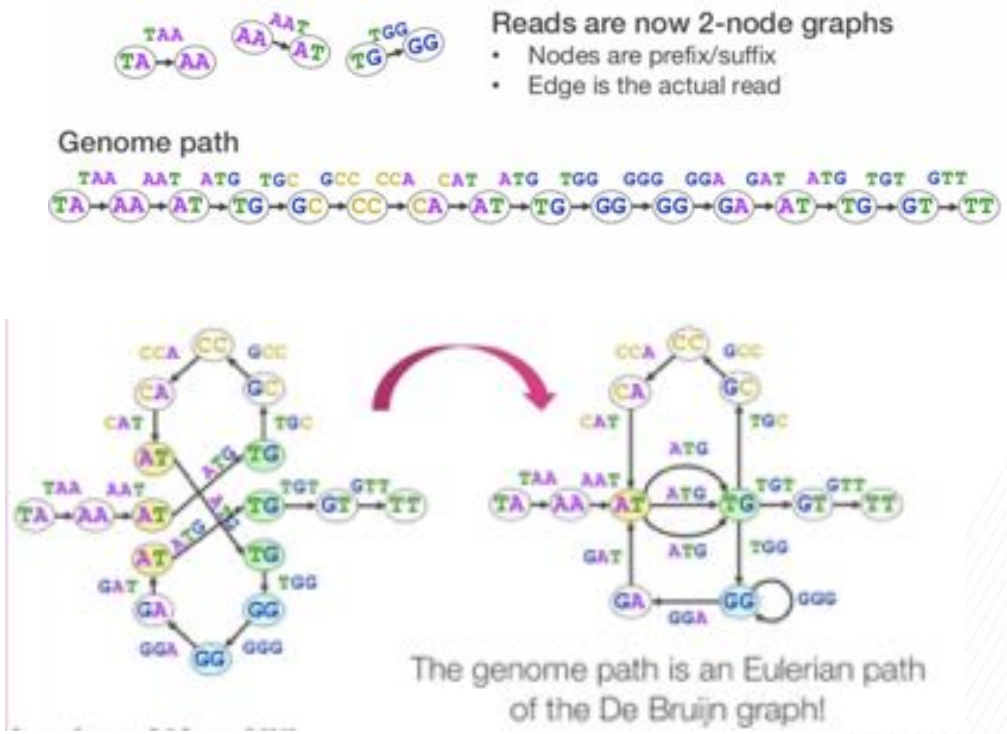



Image from : https://www.slideshare.net/JosHctorGlvez/basics-of-genome-assembly

# How to select a De novo tool?

## Why not a single tool?

The GAGE-B - Genome Assembly Gold-standard Evaluation for Bacterial study [2] shows that assembly software that performs well on one organism often performs poorly on other organisms.

Test several approaches and also with different parameter settings

# Parameters considered when selecting tool:

- Known: Genome size, Ploidy, paired end, short reads.
- Coverage
- Which algorithm to use
- Low Computational resource consumption
- Strongest performance - **common heuristics** for selecting the best assembly when the true genome sequence is unknown:
    - higher N50 contig size
    - higher sequence coverage
    - low assembly error rates

# Tool Comparison table

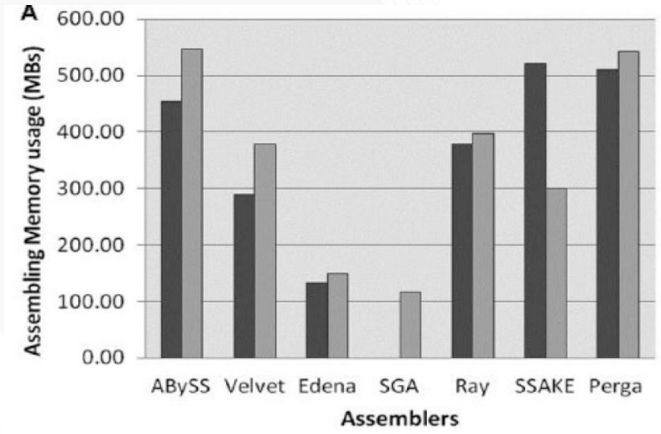
Study : A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective

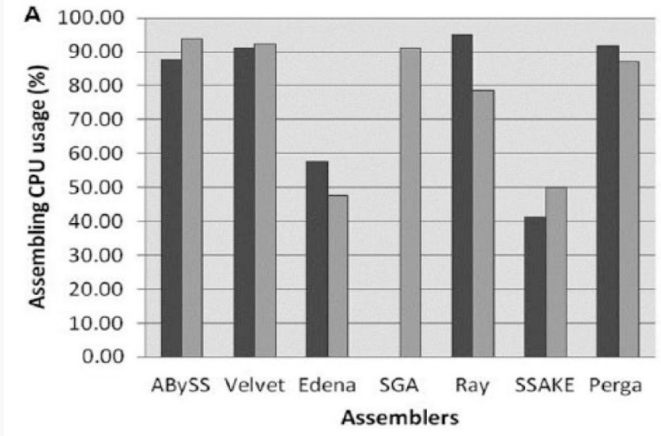
Efficiency evaluation, Accuracy evaluation and Statistical analysis is done

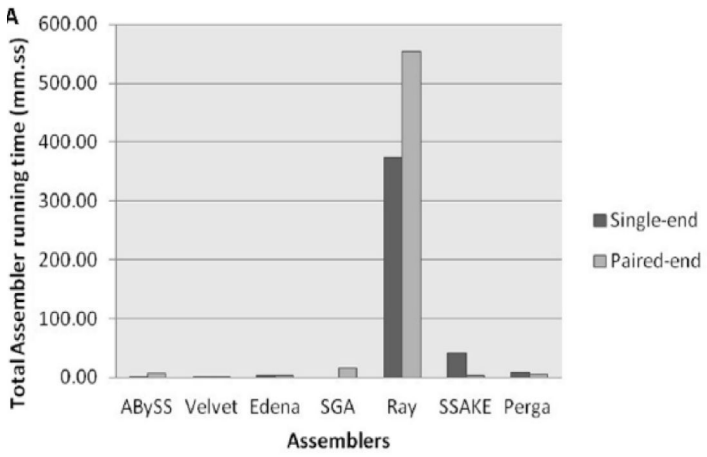
| De novo Tools | Assembler type | Good for read length | Programming language | Citations | Advantages | Disadvantages |
|---|---|---|---|---|---|---|
| **AbySS** | DBG | Single and paired end | C++ | 3078 | High N50, high genome fraction, Time efficient and High scalability | More memory and high cpu usage |
| **Velvet** | DBG | Single and paired end | C | 8667 | Time efficient, High genome fraction | High cpu usage, Low N50 |
| **Edena** | OLC | Single and paired end | C++ | 675 | Lowest Memory usage and low CPU usage | Average accuracy (mean genome fraction) |
| **SGA** | String graph | Paired end | C++ | - | Memory efficient | Only paired end, High cpu usage and Average accuracy (mean genome fraction) |
| **RAY** | Hybrid | Single and paired end | C++ | 475 | High genome fraction | Highest time [500 mins] |
| **SSAKE** | Greedy | Single and paired end | Perl | 600 | time and CPU usage efficient | Worst genome fraction |
| **Perga** | Greedy | Single and paired end | C | 11 | Good N50 contig length | High memory usage, Average accuracy (mean genome fraction) |

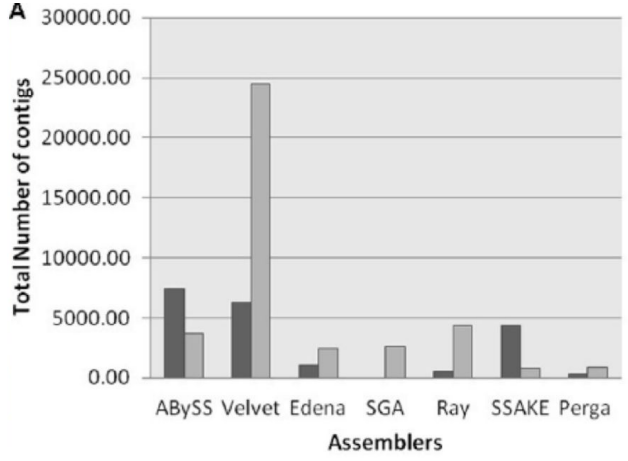
# Tool comparison plots

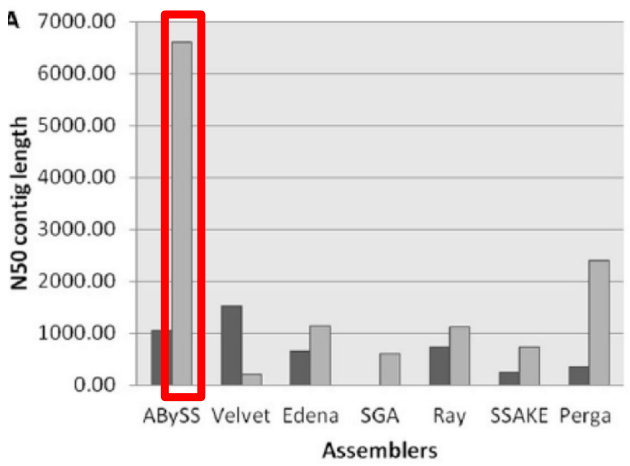
Study : A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective

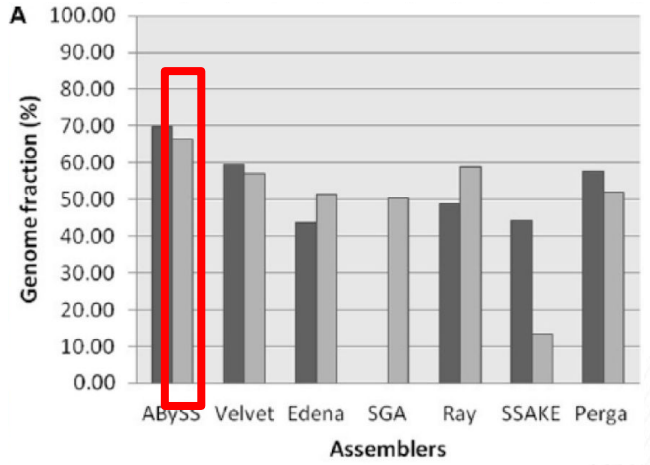



Graph plots from the study “Abdul Rafay Khan et.al [2018] - “A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective”

# Tool comparison

Study : GAGE-B: an evaluation of genome assemblers for bacterial organisms

| Assembler | Species assembled | | | | | | |
|---|---|---|---|---|---|---|---|
| | HiSeq (100 bp) reads | | | MiSeq (250 bp) reads | | | |
| | *R.sphaeroides* | *M.abscessus* | *V.cholerae* | *B.cereus* | *R.sphaeroides* | *M.abscessus* | *V.cholerae* |
| ABySS | 13.0 | 115.7 | 93.0 | 130.6 | 21.4 | 68.5 | 60.3 |
| CABOG | 11.2 | 78.2 | 48.8 | 150.5 | 30.5 | 8.3 | 32.5 |
| MIRA | 17.7 | 129.2 | 87.1 | 100.0 | 15.4 | 75.0 | 108.7 |
| MaSuRCA | **176.8** | **194.0** | **236.4** | **246.7** | **130.7** | 36.2 | 71.6 |
| SGA | 12.1 | 27.9 | 23.4 | 25.5 | 9.1 | 12.8 | 27.3 |
| SOAPdenovo | 10.5 | 147.2 | 106.5 | **246.3** | 33.5 | 113.3 | 65.5 |
| SPAdes | 83.5 | 147.9 | 77.1 | 103.7 | 118.1 | **215.4** | **246.6** |
| Velvet | 13.1 | 60.3 | 39.5 | 24.5 | 24.2 | 41.5 | 67.1 |

# Tools interested

ABySS

- Assembly by Short Sequences
- DBG, Parallel, paired-end assembler designed for short reads
- **Pros: high genome fraction, high N50, time efficient**
- Cons: more memory and cpu usage than velvet

MaSuRCA

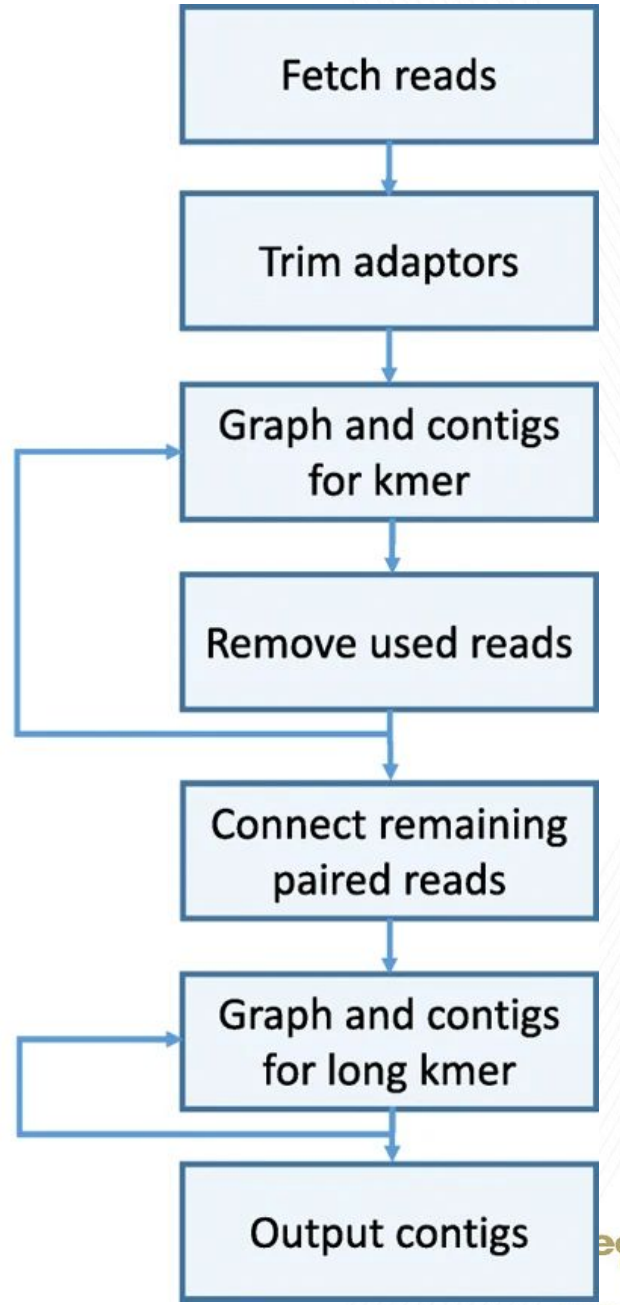
- Maryland Super-Read Celera Assembler
- Combination of OLC + De bruijn graph
- selects optimal k-mer [k-mer length comparison not necessary]
- **Pros: usually larger assembly, high N50**
- Cons: mis-assemblies

# SPAdes

- DBG assembler for small genomes
- Designed to address issues in single-cell sequencing
- **Multisized de Bruijn graph: build graph from small to large k-mer sizes**, each based on previous one. Allows SPAdes to get the advantages of both small k-mer assemblies (a more connected graph) and large k-mer assemblies (ability to resolve repeats).
- **Resolve repeats: paired-end information**. Since two reads in a pair are close to each other in the original DNA, SPAdes can use this to trace paths in the graph to form larger contigs
- **Pros:** High N50, Genome fraction, No. of complete genes.
- **Cons:** Time-consuming

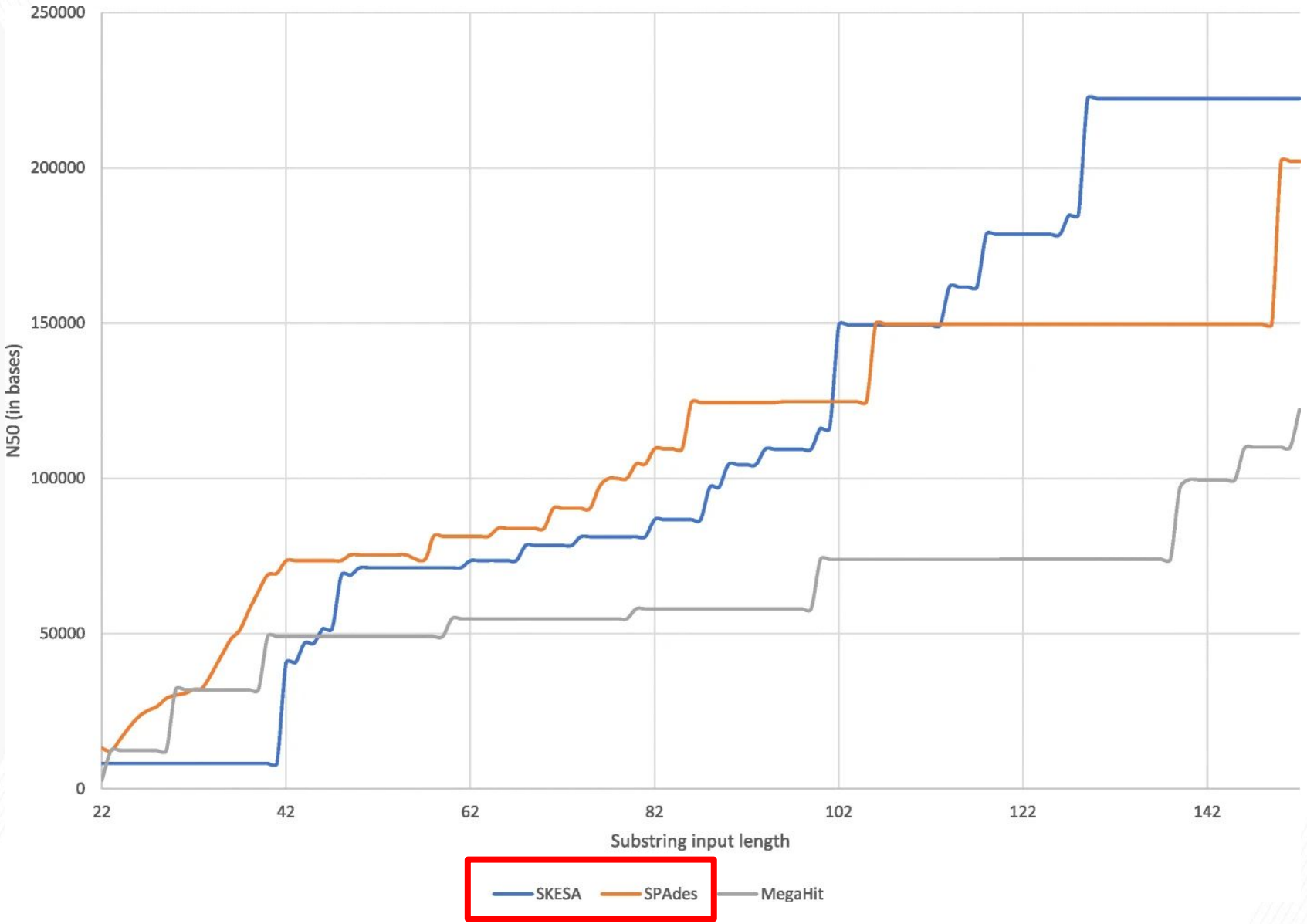
# SKESA

- DBG assembler for microbial genomes sequenced using Illumina
- **Using different k-mer sizes:**
  - From k-min (default 21) to the average read length, in a default of 11 iterations
  - Increases upto to insert size, in 3 iterations
- **Assemble repeats** shorter than insert size but longer than the mate length
- **Pros:** fast and reproducible
- **Cons:** no built-in scaffolding tool



# SKESA

- N50:






Georgia Tech
CREATING THE NEXT

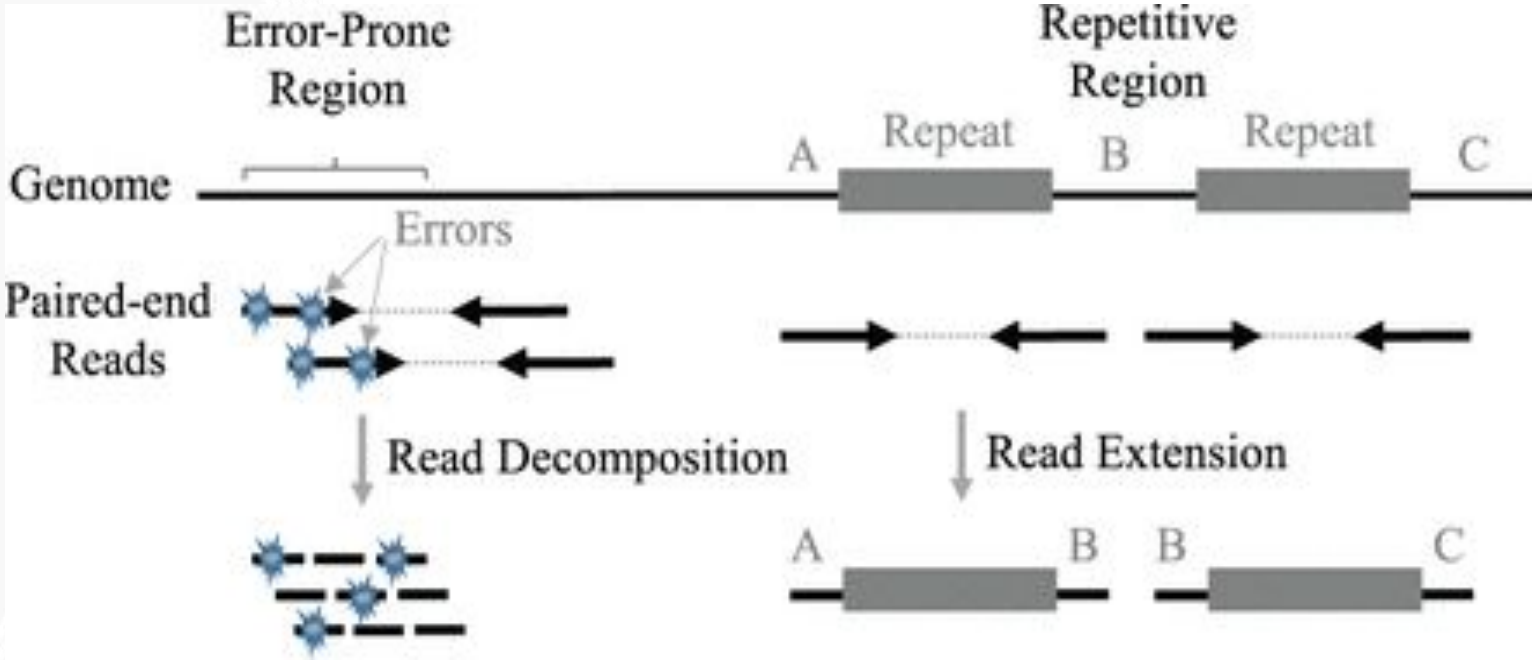
# SKESA

- Misassemblies:

| Benchmark set | | | |
|---|---|---|---|
| Measure | SKESA | SPAdes | MegaHit |
| Median | 0.08 | 2.76 | 1.89 |
| Maximum | 7.78 | 41.60 | 31.94 |
| Average | 0.40 | 3.21 | 2.79 |
| Assembly counts in benchmark set | | | |
| Mismatches range | SKESA | SPAdes | MegaHit |
| 0 | 105 | 1 | 1 |
| 0.01–1 | 247 | 40 | 80 |
| 1.01–2 | 9 | 76 | 121 |
| 2.01–3 | 9 | 89 | 58 |
| 3.01–4 | 1 | 71 | 45 |
| > 4 | 10 | 104 | 76 |

# StriDe

- Combination of string and de Bruijn graphs.
- Reads in **error-prone region** are decomposed into overlapping subreads.
- The paired-end reads are extended into long reads using an FM-index, to resolve **repeats longer than read length**.



# StriDe

- N50:

| ABySS | CABOG | MaSuRCa | SOAPdenovo | SGA | SPAdes | Velvet | StriDe |
|---|---|---|---|---|---|---|---|
| 237.7 | 278.4 | **838.5** | 243.9 | 67.1 | 237.6 | 184.4 | **827.8** |
| 41.6 | 61.1 | 75.2 | 57.9 | 20.5 | 78.6 | 38.9 | **90.6** |
| 116.3 | 94.2 | 99.7 | 116.1 | 45.0 | 127.4 | 125.2 | **151.5** |
| 128.5 | 78.2 | 147.4 | 147.2 | 28.7 | **278.4** | 60.3 | **298.0** |
| 115.8 | 11.2 | 36.4 | 10.5 | 4.8 | **173.3** | 13.1 | **175.1** |
| 99.2 | 102.8 | **228.9** | 146.3 | 39.9 | 148.1 | 122.5 | **222.2** |
| 172.6 | 48.8 | 167.9 | 106.5 | 23.8 | **344.0** | 39.5 | **356.0** |
| 74.1 | 105.8 | **115.7** | 74.2 | 48.9 | **117.2** | 83.0 | **113.0** |

- Misassemblies:

| ABySS | CABOG | MaSuRCa | SGA | SOAPdenovo | SPAdes | Velvet | StriDe |
|---|---|---|---|---|---|---|---|
| 1 (140.14) | 0 (281.24) | 3 (235.23) | 0 (97.50) | 1 (127.95) | 1 (115.88) | 0 (108.65) | 1 (115.44) |
| 3 (0.98) | 7 (5.81) | 7 (3.58) | 1 (0.43) | 9 (0.67) | 5 (0.48) | 4 (0.74) | 5 (0.43) |
| 9 (5.37) | 4 (1.55) | 5 (2.13) | 1 (0.30) | 2 (1.94) | 2 (1.19) | 2 (0.31) | 3 (0.70) |
| 3 (3.98) | 20 (6.84) | 16 (6.00) | 3 (2.60) | 21 (3.67) | 7 (3.03) | 5 (3.37) | 3 (2.85) |

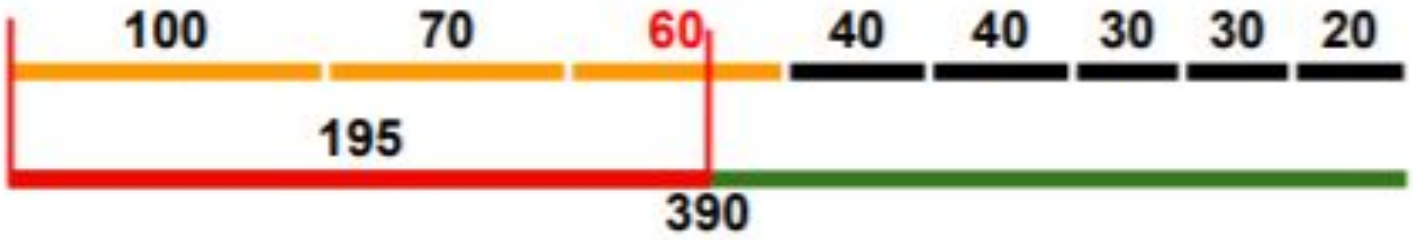
# Assembly Quality - Broad Overview

- Using Assembly information
  - **N50 family metrics** (quast)
  - remapping reads to assembled contigs (AMOSValidate, REAPR, FRCbam, Pilon, VALET)
  - computing probability of reads given the assembly (ALE, CGAL, LAP)

- Using External information
  - **map to near reference genome** (quast, dnAQET or Assemblytics)
  - counting number of conserved genes in assembly (BUSCO)

# Assembly Quality - N50 family metrics

- Length of largest contig
- number of contigs
- N50 and L50
- N75 and L75



# Assembly Quality - Quast Example output

| Assembly | pit_fna | cef_fna | car_fna |
|---|---|---|---|
| # contigs (>= 0 bp) | 100 | 91 | 94 |
| # contigs (>= 1000 bp) | 62 | 58 | 61 |
| Total length (>= 0 bp) | 6480635 | 6481216 | 6480271 |
| Total length (>= 1000 bp) | 6466917 | 6468946 | 6467103 |
| # contigs | 71 | 66 | 70 |
| Largest contig | 848753 | 848766 | 662053 |
| Total length | 6473173 | 6474698 | 6473810 |
| GC (%) | 66.33 | 66.33 | 66.33 |
| N50 | 270269 | 289027 | 254671 |
| N75 | 136321 | 136321 | 146521 |
| L50 | 7 | 7 | 8 |
| L75 | 15 | 15 | 16 |
| # N's per 100 kbp | 0.00 | 0.00 | 0.00 |

# Assembly Quality with Quast

- N50 family metrics

- Conserved genes using BUSCO

**`--conserved-genes-finding`** (or `-b`)

Enables search for Universal Single-Copy Orthologs using BUSCO (only on Linux, only with Python 2.7 or Python 3). Disabled by default.

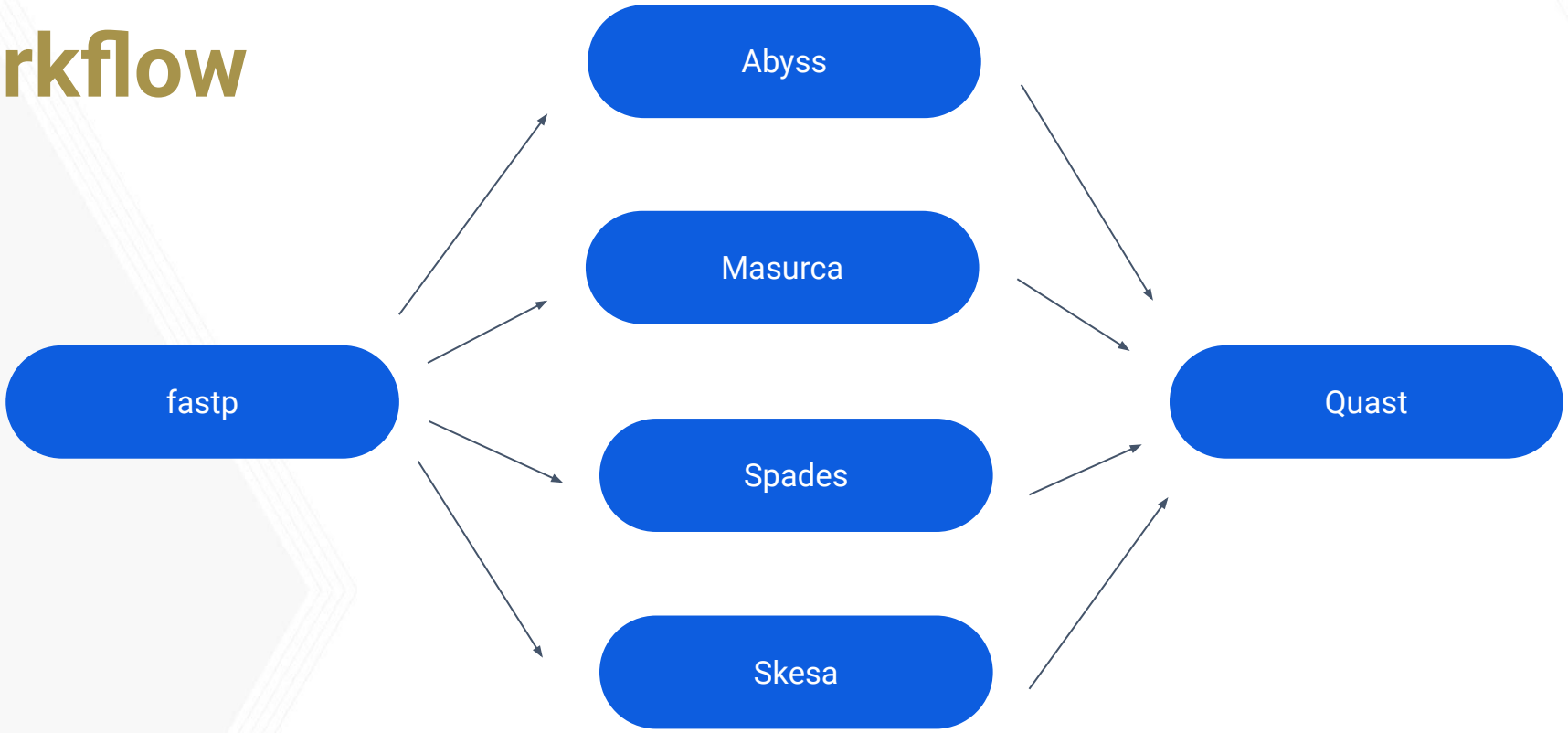
By default, we assume that the genome is prokaryotic, and BUSCO uses the bacterial database of orthologs. If the genome is eukaryotic (fungal), use `--eukaryote` (`--fungus`) option to force BUSCO to work with the eukaryotic (fungal) database.

- NG stats based on estimated reference size?

**`--est-ref-size`** `<int>`

Estimated reference genome size (in bp) for computing NGx statistics. This value will be used only if a reference genome file is not specified (see `-r` option).

# Our Workflow




**fastq**
Quality control and trimming for reads

**Feb 2nd**

**Feb 9th**

**Assembly**
Test out 4 tools and complete assembly

**Quast**
Quality control for assembled reads

**Feb 12th**

**Feb 17th**

**2nd presentation preparation**
Finish slides and update the results



Georgia Tech
CREATING THE NEXT

# References


1. Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. Biological Procedures Online, 11, 52–78. doi:10.1007/s12575-009-9004-1
2. Dominguez Del Angel V, Hjerde E, Sterck L et al. Ten steps to get started in Genome Assembly and Annotation [version 1; peer review: 2 approved]. F1000Research 2018, 7(ELIXIR):148
3. Abdul Rafay Khan et.al [2018] - “A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective” - PMID: 29511353, doi: 10.1177/1176934318758650
4. Tanja Magoc et.al [2013] - “GAGE-B: an evaluation of genome assemblers for bacterial organisms” - PMID: 23665771, doi: 10.1093/bioinformatics/btt273
5. Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, Alexey Gurevich, Versatile genome assembly evaluation with QUAST-LG, Bioinformatics (2018) 34 (13): i142-i150. doi: 10.1093/bioinformatics/bty266
6. Bankevich, A.; Nurk, S. et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 19(5), 455–477. doi: 10.1089/cmb.2012.0021
7. Huang, Y.-T., & Liao, C.-F. (2016). Integration of string and de Bruijn graphs for genome assembly. Bioinformatics, 32(9), 1301–1307. doi: 10.1093/bioinformatics/btw011
8. Souvorov, A., Agarwala, R., & Lipman, D. J. (2018). SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biology, 19(1). doi: 10.1186/s13059-018-1540-z