Genome Assembly

Team II

Tomáš Brůna, Harshini Chandrashekar, William Harvey, Jane Lew, Yusuph Mavura, Eunbi Park, Vishnu Raghuram, Beatriz Saldana, Parisa Yousefi Zowj

Outline

- Project dataset
- Final Pipelines
  - Reference Assembly
  - De Novo
- Results and Comparison

How Antibiotic Resistance Happens

1.
Lots of germs.
A few are drug resistant.

2.
Antibiotics kill bacteria causing the illness, as well as good bacteria protecting the body from infection.

3.
The drug-resistant bacteria are now allowed to grow and take over.

4.
Some bacteria give their drug-resistance to other bacteria, causing more problems.

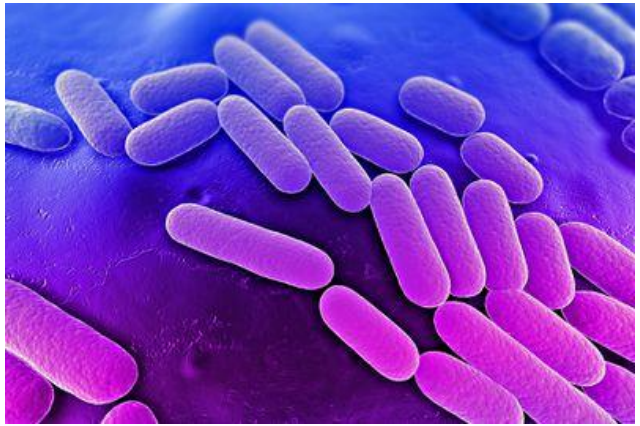
Dataset

Klebsiella:

- 5 Mb long
- 1 chromosome
- 57% GC content

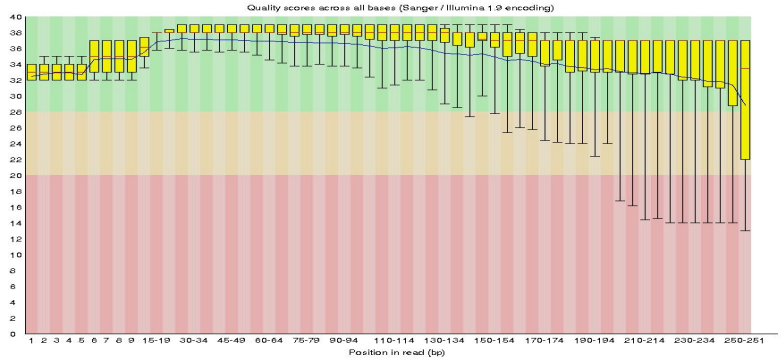
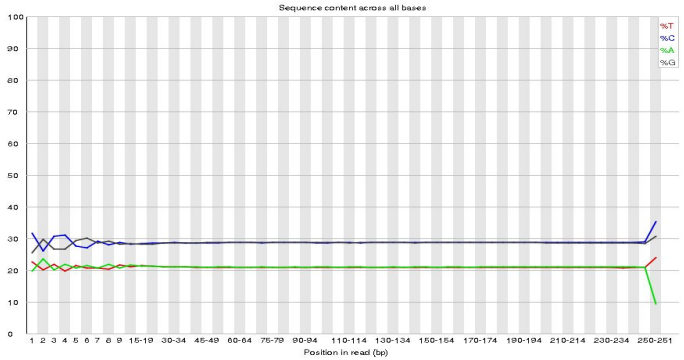
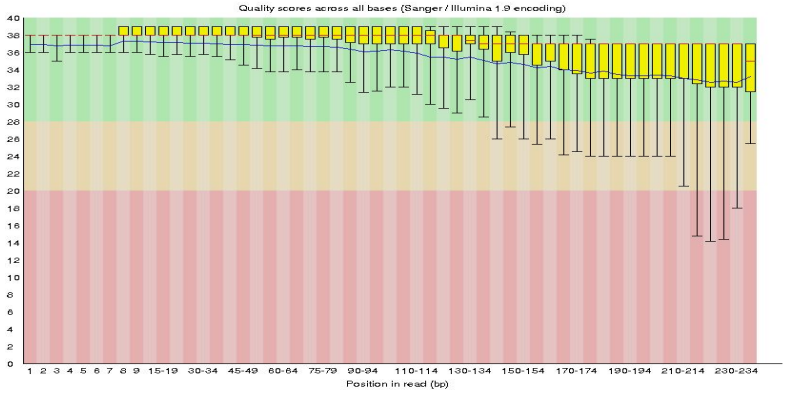
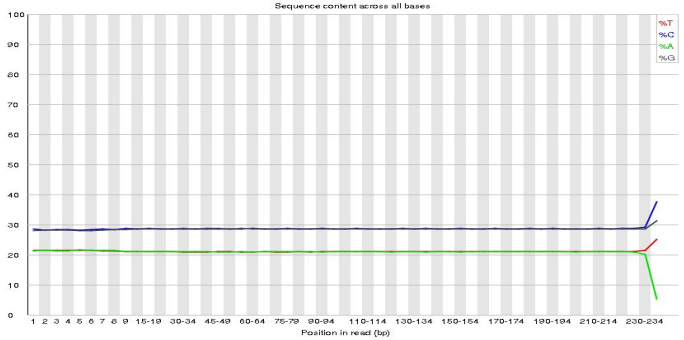
Sequencing Data:

- 262 genomes
- Paired-end, around 250 bp reads
- From Illumina MiSeq2500

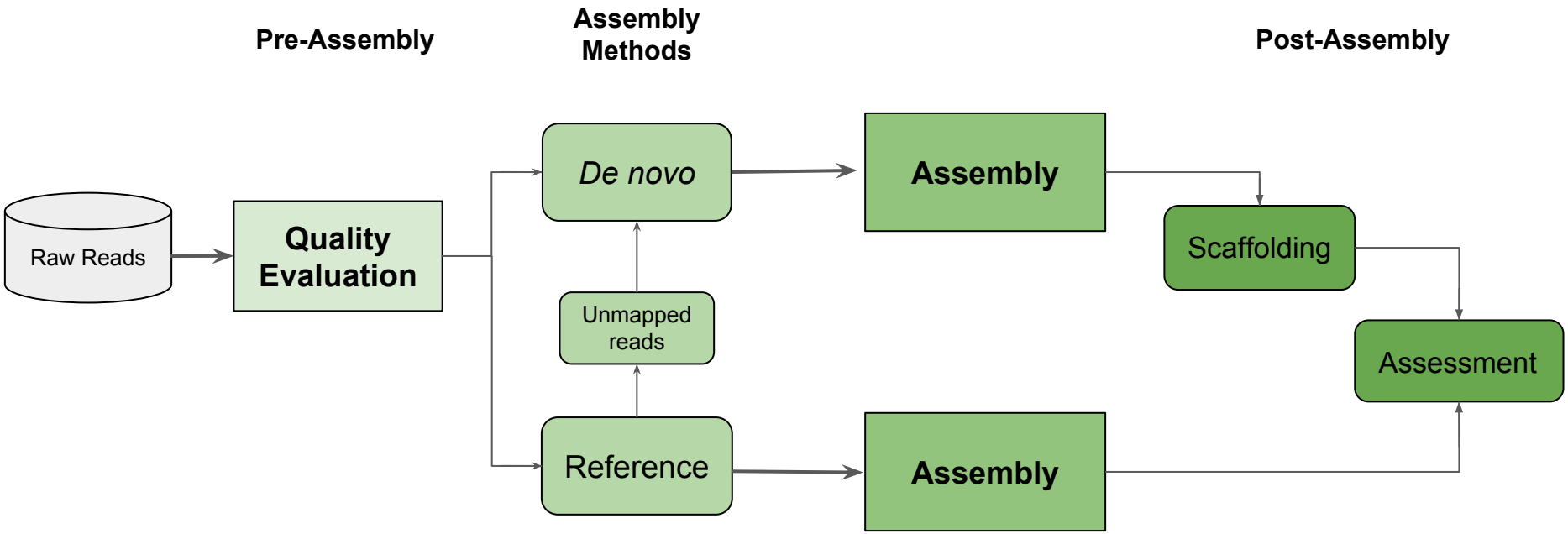


Klebsiella pneumoniae. From: Bioquell.com

Quality Evaluation

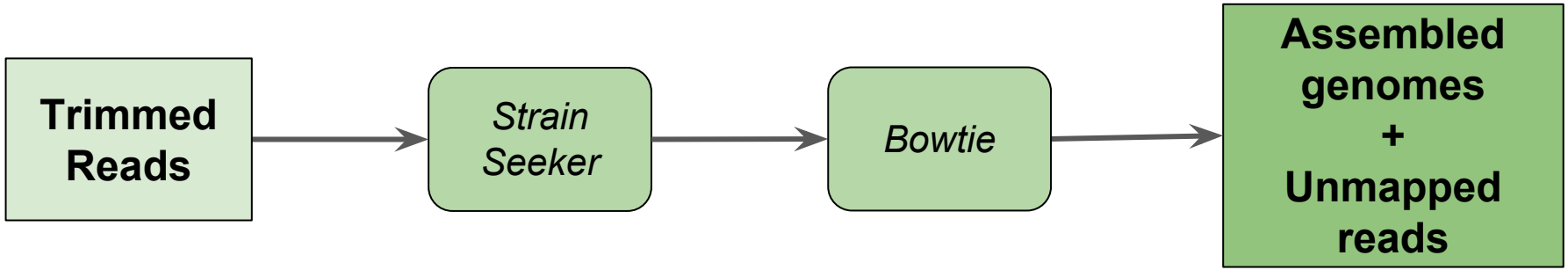
	Per base sequence quality	Base Content
Before		
After		

Pipeline

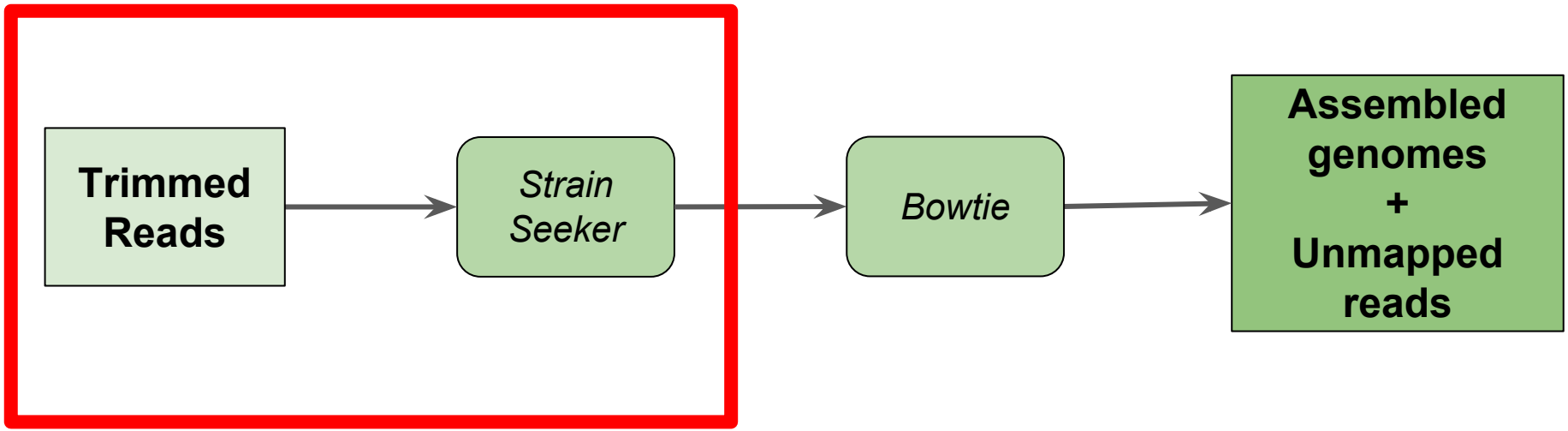

Pre-Assembly
Assembly Methods
Post-Assembly
Raw Reads
Quality Evaluation
De novo
Unmapped reads
Reference
Assembly
Assembly
Scaffolding
Assessment

Reference Assembly Pipeline

Reference Assembly Pipeline



Reference Assembly Pipeline

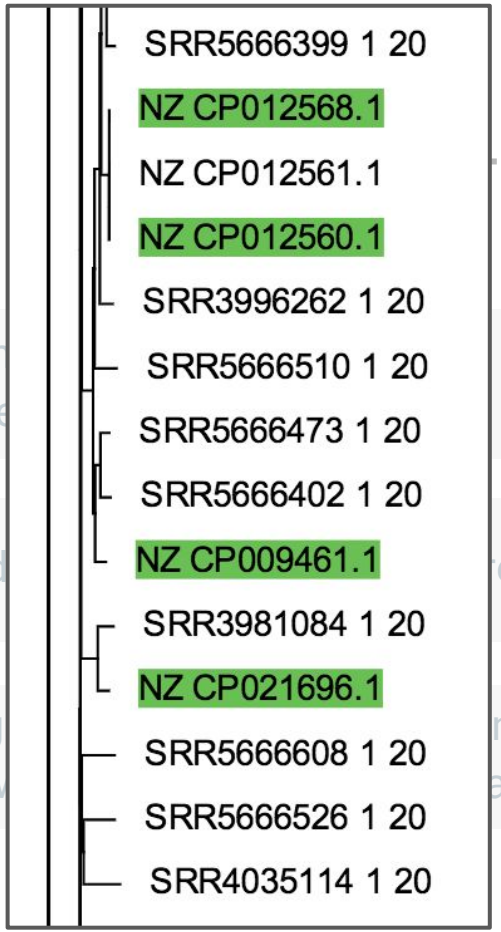


Reference Assembly Pipeline

Mashtree	Q: Which reference genomes from the 200+ *Klebsiella* genomes are closely related to our reads?
Manual Curation	Selected the most closely related references for our database
ALFRED-G	Creating a reference guide tree from the selected references (output will be the input for the strain seeker database build step)

Reference Assembly Pipeline

Trimmed Reads



Strain Seeker

Mashtree

Q: Which … the 200+ *Klebsiella* genomes are close…

Manual Curation

Selected … references for our database

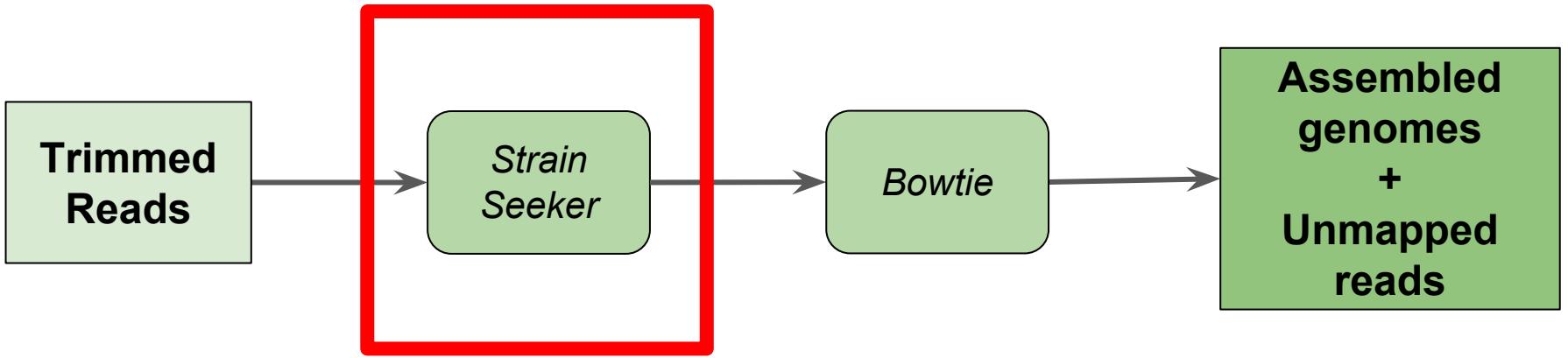
ALFRED-G

Creating … m the selected references (output … ain seeker database build step)

Reference Assembly Pipeline

Mashtree	Q: Which reference genomes from the 200+ *Klebsiella* genomes are closely related to our reads?
Manual Curation	Selected the most closely related references for our database
ALFRED-G	Creating a reference guide tree from the selected references (output will be the input for the strain seeker database build step)

Reference Assembly Pipeline



Reference Assembly Pipeline

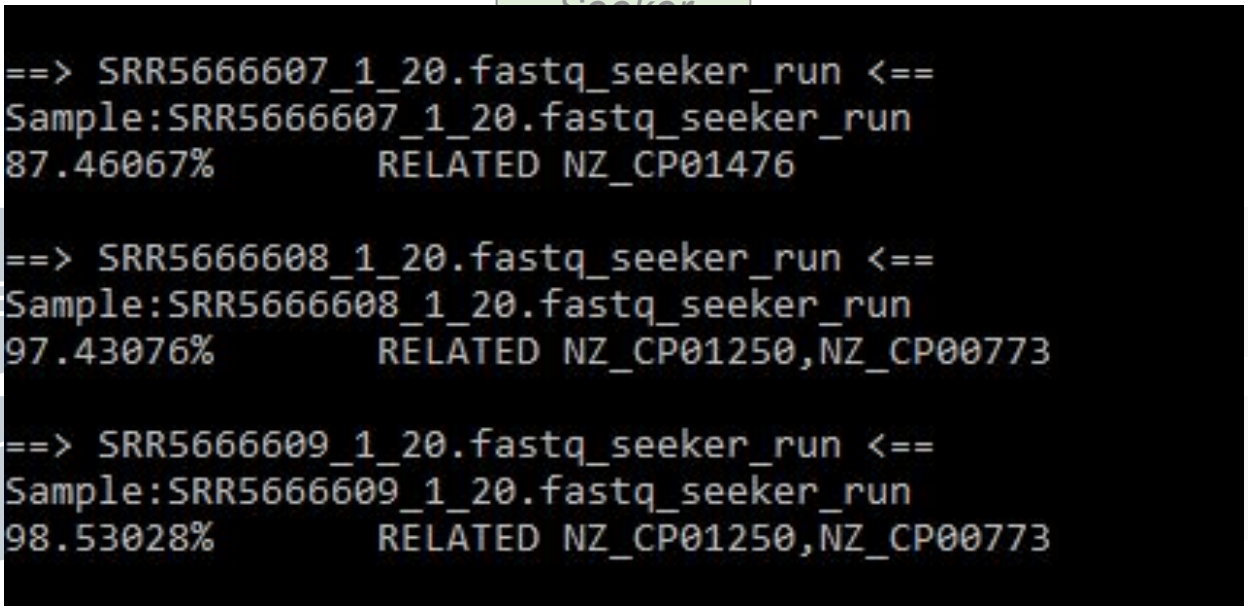
Build DB	Tree (Newick) + References (fasta)
Determine Strain	Database + Trimmed Reads (fastq)

Reference Assembly Pipeline

Strain

Build DE

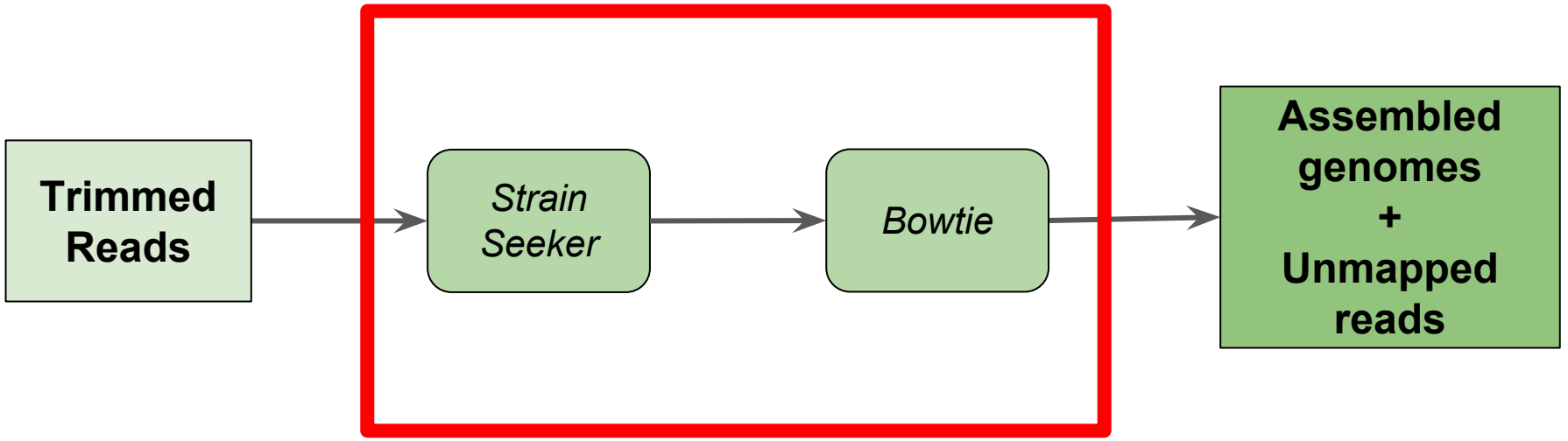
Determin
Strain



Reference Assembly Pipeline

Build DB	Tree (Newick) + References (fasta)
Determine Strain	Database + Trimmed Reads (fastq)

Reference Assembly Pipeline

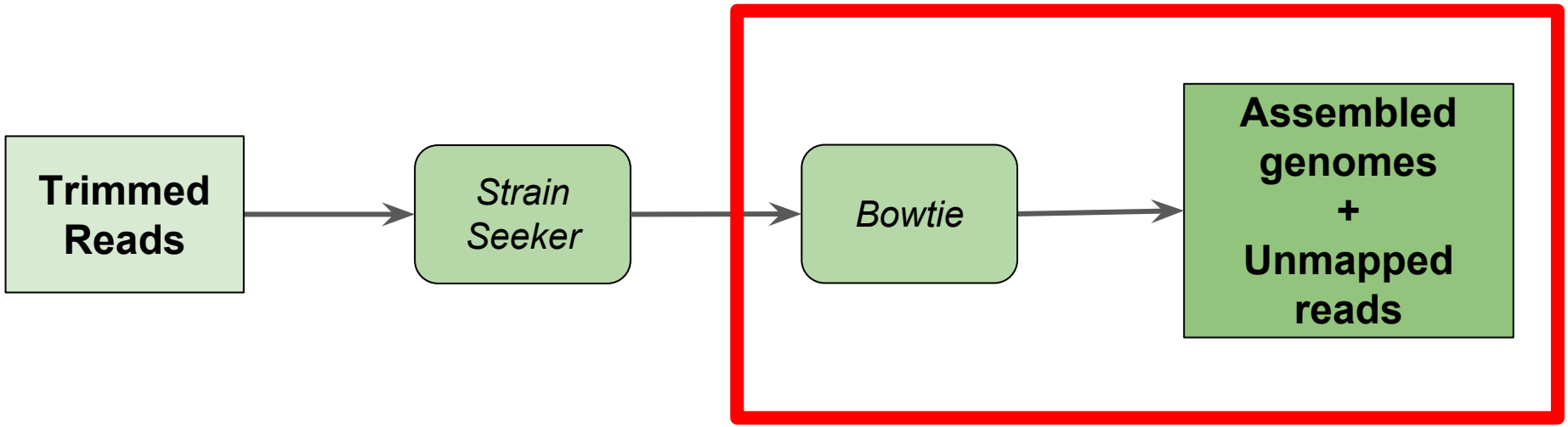


Reference Assembly Pipeline

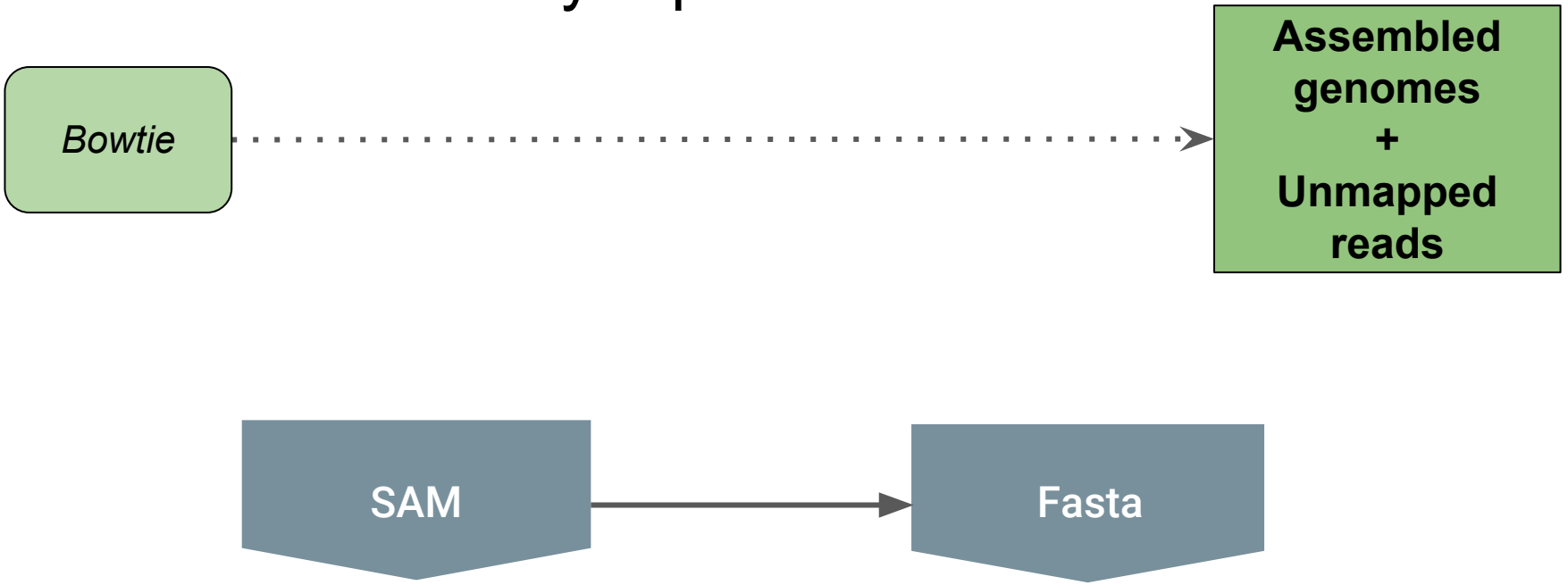
Grouping

Group results according to their most closely related reference genome

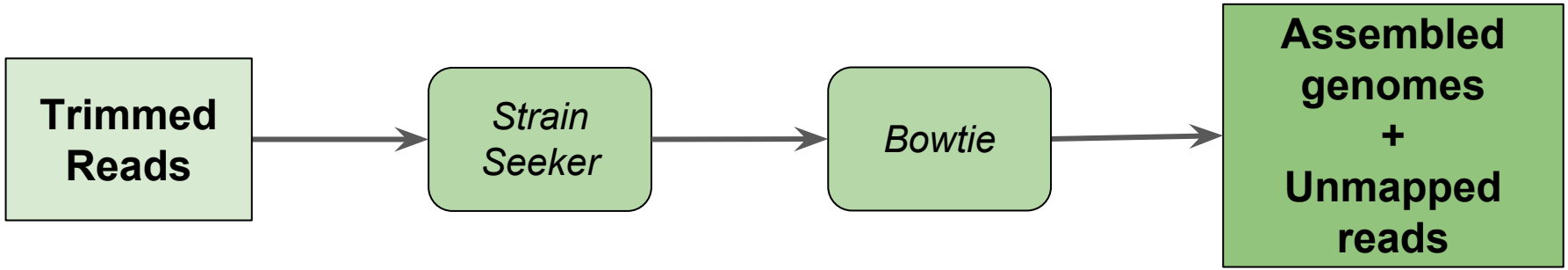
Reference Assembly Pipeline



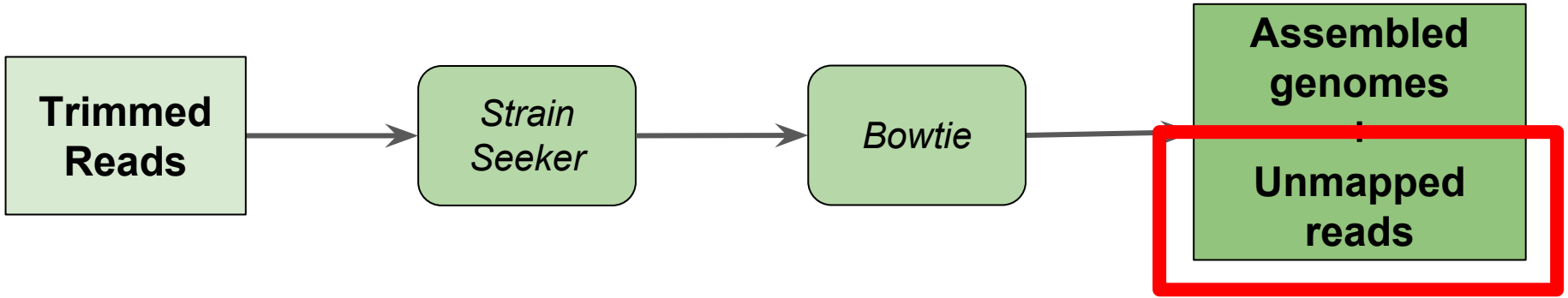
Reference Assembly Pipeline



Reference Assembly Pipeline



Reference Assembly Pipeline

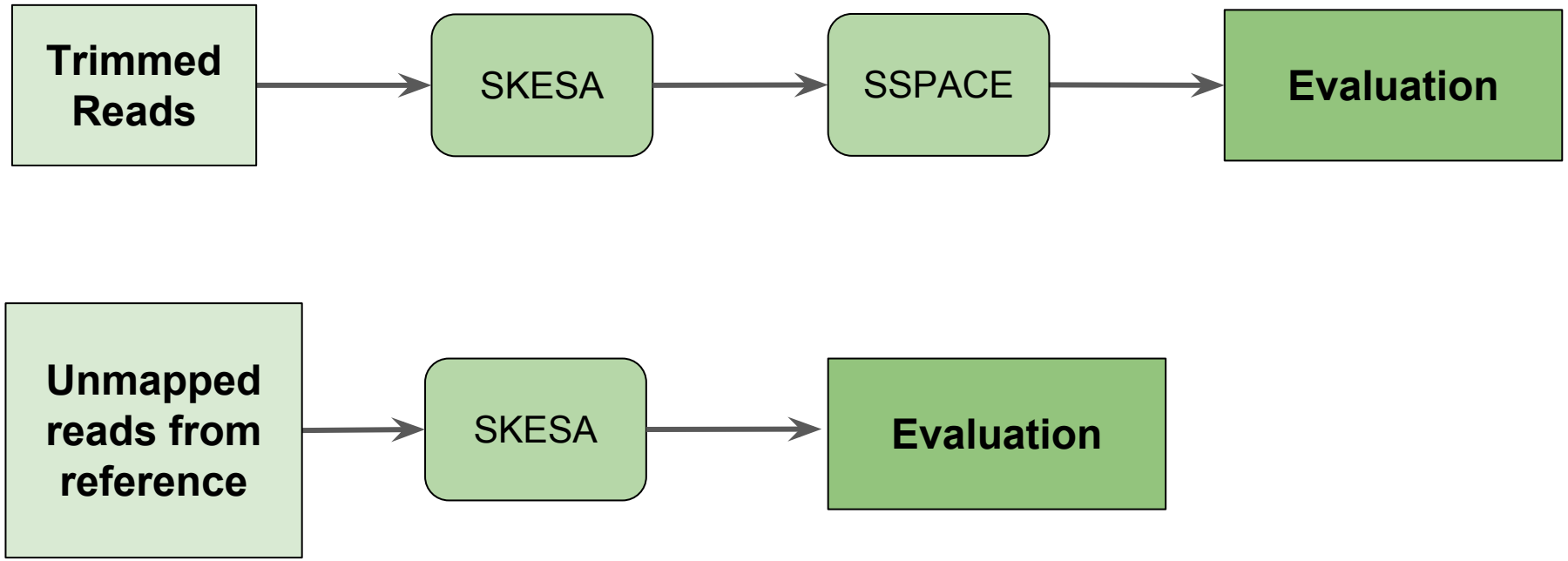


De Novo Pipeline

De Novo

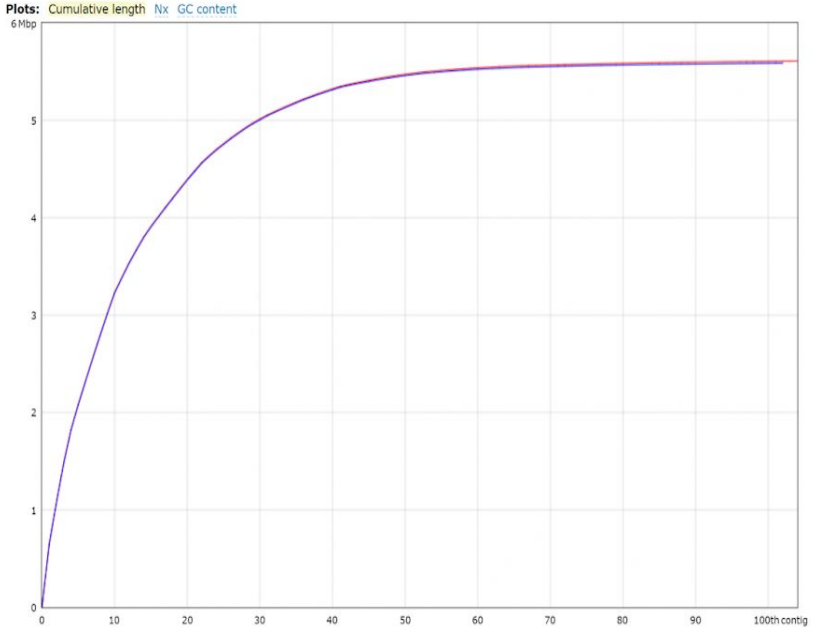
- Assembler of choice -- Skesa
- Scaffolding tool -- SSPACE

De Novo: Final Pipeline



De Novo: Results

- Scaffolding shows only marginal improvement with higher genome coverage (trimming threshold 20 compared to 27)
- We observed similar results gap filling



SRR5666600.final.scaffolds
SRR5666600.contings.skesa

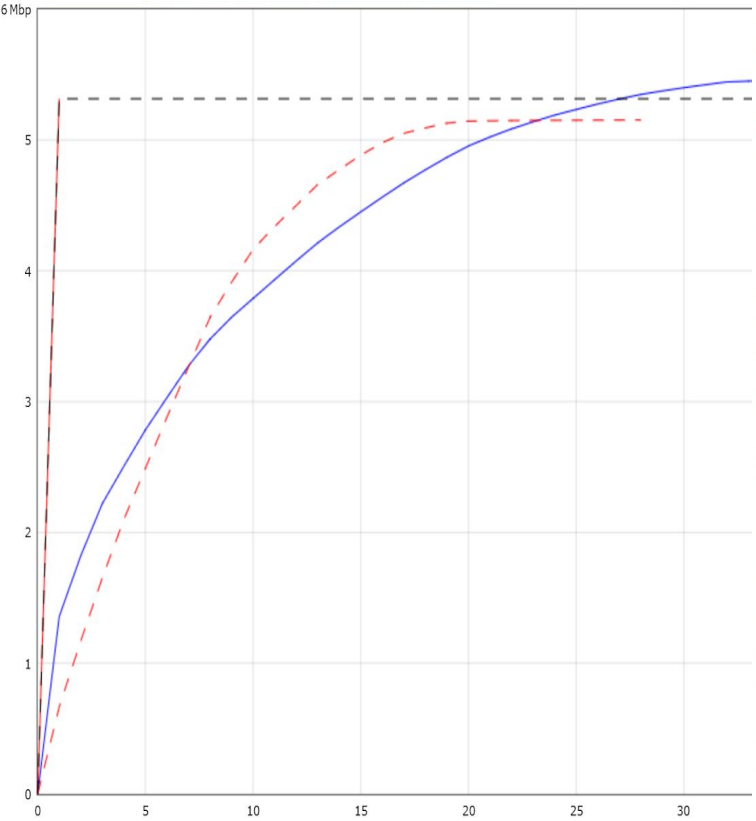
Worst Median Best

Show heatmap

Statistics without reference	SRR5666600.final.scaffolds	SRR5666600.contings.skesa
# contigs	104	102
# contigs (>= 0 bp)	111	111
# contigs (>= 1000 bp)	90	89
# contigs (>= 5000 bp)	59	57
# contigs (>= 10000 bp)	50	50
# contigs (>= 25000 bp)	39	39
# contigs (>= 50000 bp)	28	28
Largest contig	647 840	647 489
Total length	5 608 644	5 586 932
Total length (>= 0 bp)	5 611 682	5 590 775
Total length (>= 1000 bp)	5 597 892	5 577 371
Total length (>= 5000 bp)	5 534 741	5 509 675
Total length (>= 10000 bp)	5 472 195	5 458 959
Total length (>= 25000 bp)	5 299 117	5 289 710
Total length (>= 50000 bp)	4 923 330	4 915 781
N50	224 309	224 223
N75	95 595	95 571
L50	9	9
L75	19	18
GC (%)	57.21	57.22
Mismatches		
# N's	0	0
# N's per 100 kbp	0	0

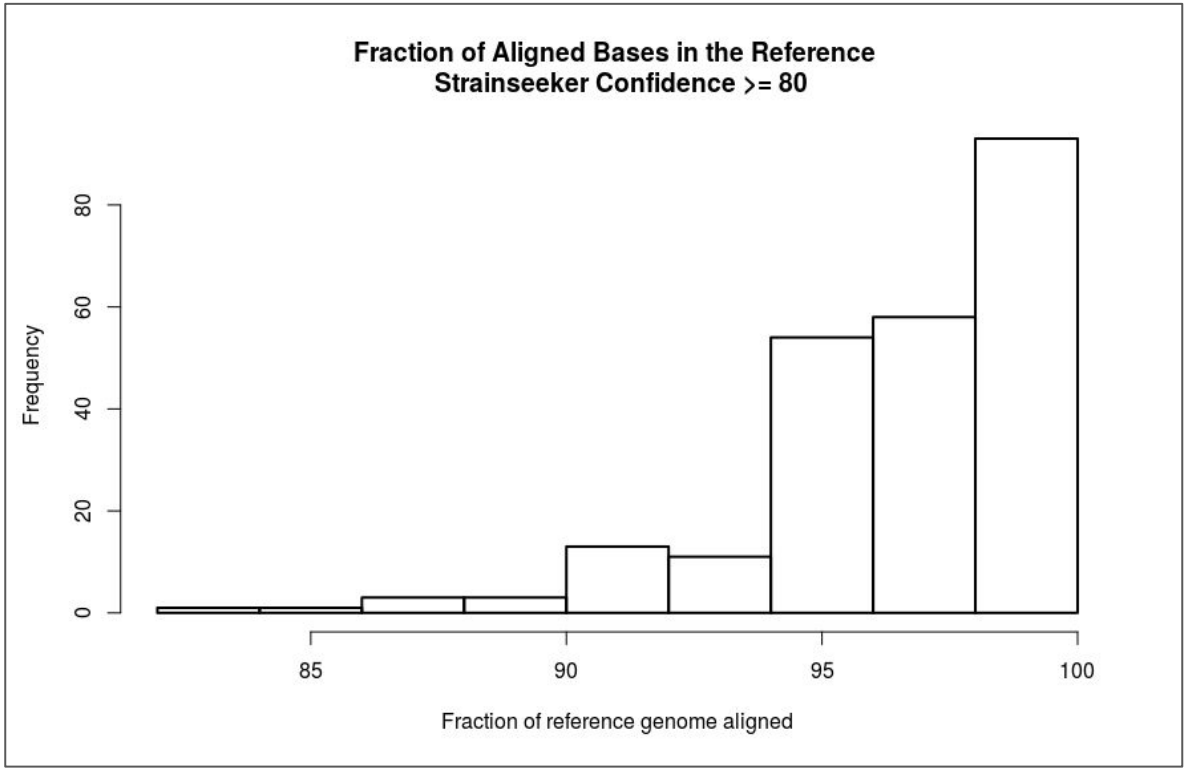
Results

Quast comparison of de novo and reference

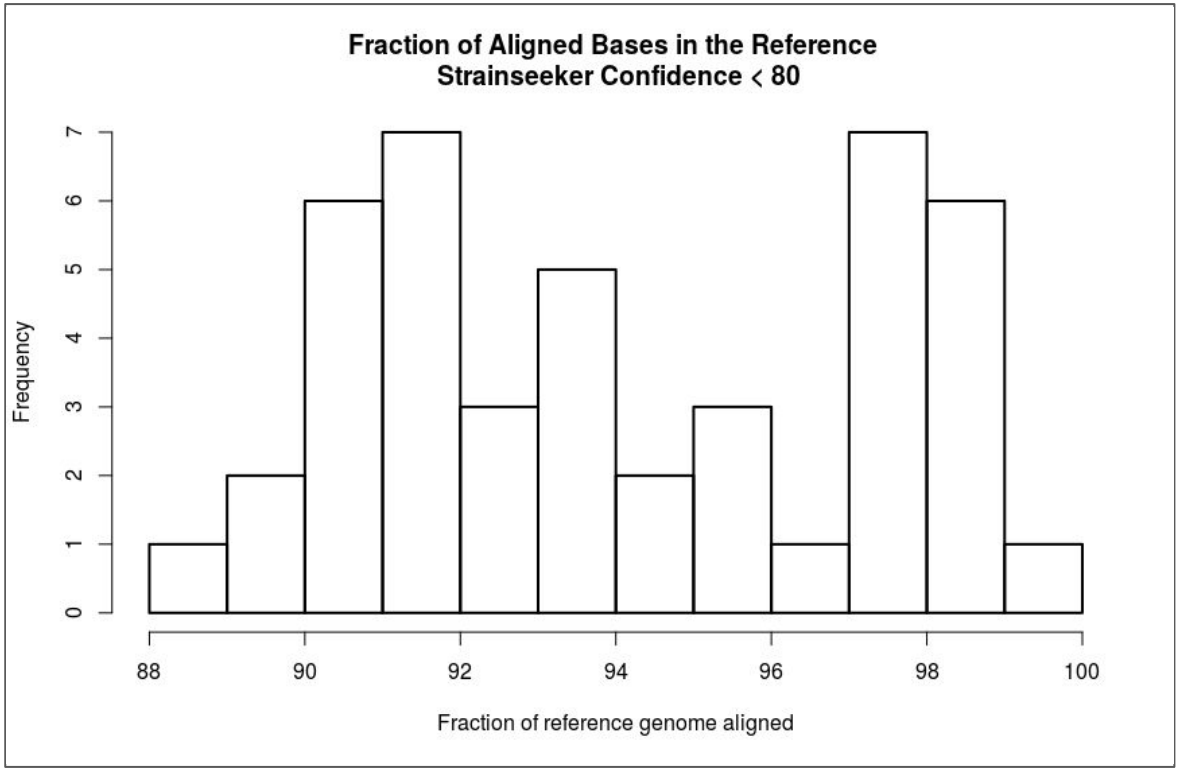


Genome statistics	SRR5666458_reference	SRR5666458_reference_broken	SRR5666458_skesa
Genome fraction (%)	96.978	96.942	95.84
Duplication ratio	1	1	1.001
Largest alignment	5 154 505	666 032	398 435
Total aligned length	5 154 505	5 152 607	5 095 972
NGA50	5 154 505	390 466	143 704
LGA50	1	6	13
Misassemblies			
# misassemblies	0	0	18
Misassembled contigs length	0	0	3 234 832
Mismatches			
# mismatches per 100 kbp	142.63	142.65	127.4
# indels per 100 kbp	0.16	0.16	3.51
# N's per 100 kbp	3028.27	0.89	0.02
Statistics without reference			
# contigs	1	28	42
Largest contig	5 315 120	666 032	1 356 900
Total length	5 315 120	5 152 607	5 466 987
Total length (>= 1000 bp)	5 315 120	5 149 589	5 464 533
Total length (>= 10000 bp)	5 315 120	5 144 940	5 444 401
Total length (>= 50000 bp)	5 315 120	5 051 115	5 190 461

Genome Fractions Aligned



Genome Fractions Aligned



Assembly of Unmapped reads

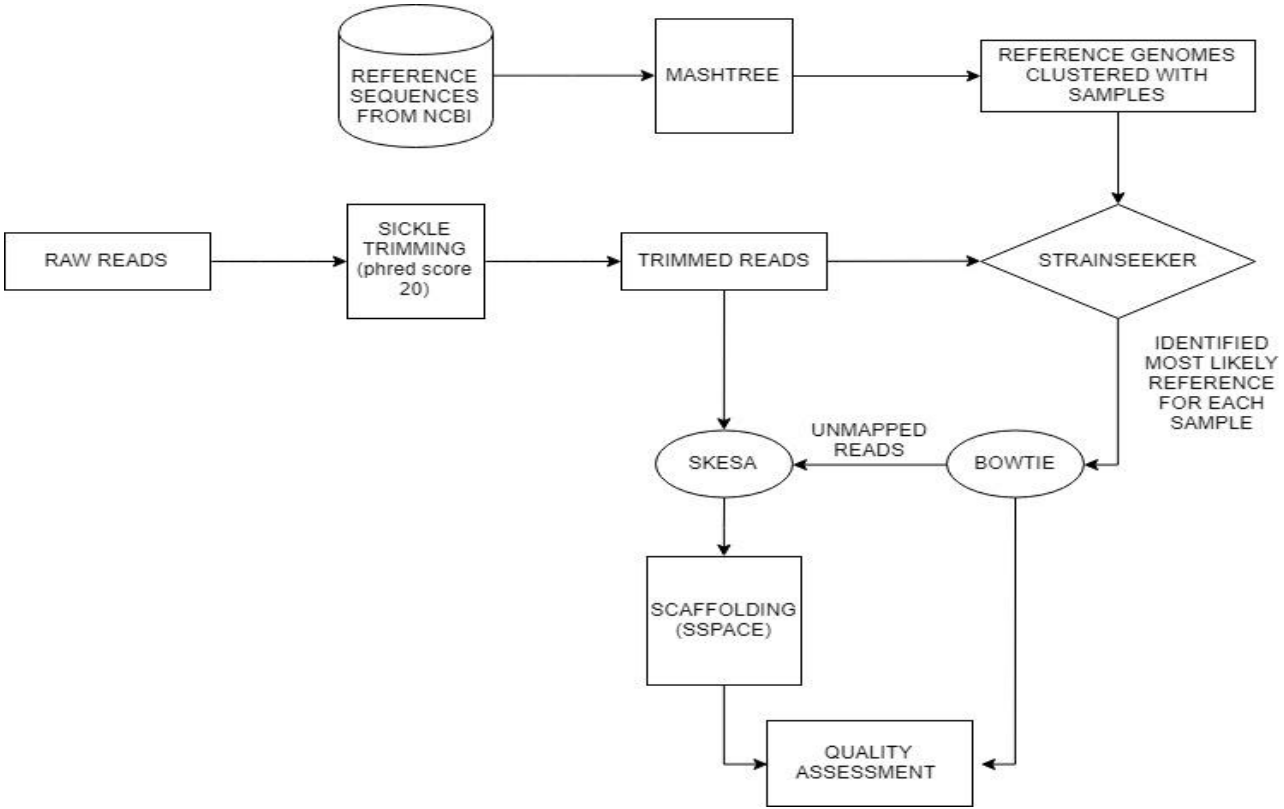
- Example: unmapped_SRR4017843.fasta
- Total number of contigs -- 38
- BLASTed the contigs after assembly
- BLAST results for one of the contigs

	Max score	Total score	Query cover	E value	Ident	Accession
Klebsiella pneumoniae subsp. pneumoniae strain ST2017:950142398 plasmid p18-43_04, complete sequence	17160	17160	99%	0.0	100%	CP023557.1
Escherichia coli strain Ecol_AZ161 plasmid pECAZ161_KPC, complete sequence	16731	17139	100%	0.0	99%	CP019010.1
Klebsiella pneumoniae strain MNCRE53 plasmid pMNCRE53_1, complete sequence	16731	17139	100%	0.0	99%	CP018433.1
Klebsiella pneumoniae strain MNCRE69 plasmid pMNCRE69_1, complete sequence	16731	17139	100%	0.0	99%	CP018425.1
Klebsiella pneumoniae 30684/NJST258_2 plasmid pNJST258C2, complete sequence	16731	17139	100%	0.0	99%	CP006919.1

Suggestions for the Next Groups

- Take reference assembled genomes for those having reference genome coverage greater than 95% and Skesa assemblies for the rest
- Quast reports for all the assemblies are on the server if a more detailed analysis is needed
- Contigs from unmapped reads will provide valuable info regarding antibiotics resistance etc. based on the number and types of plasmids/phage present

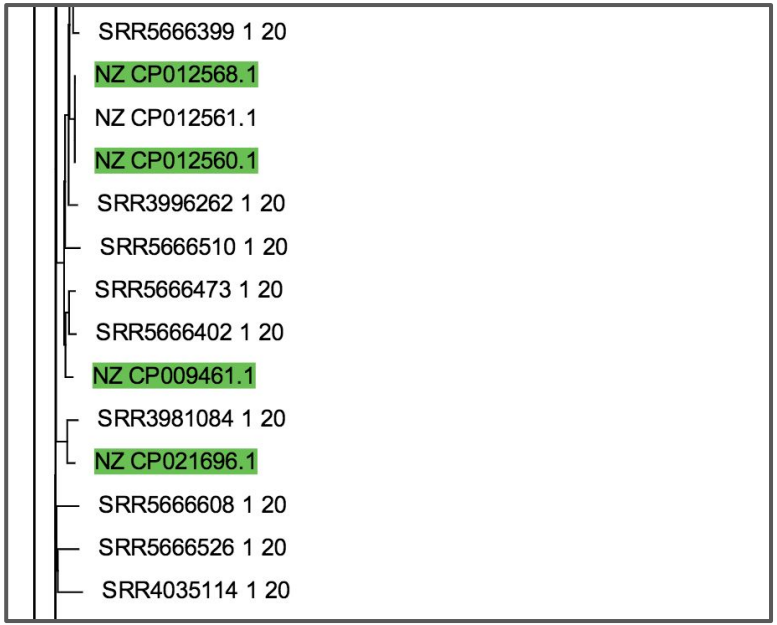
Our Genome assembly pipeline



Extra slides

Reference Assembly: Mashtree

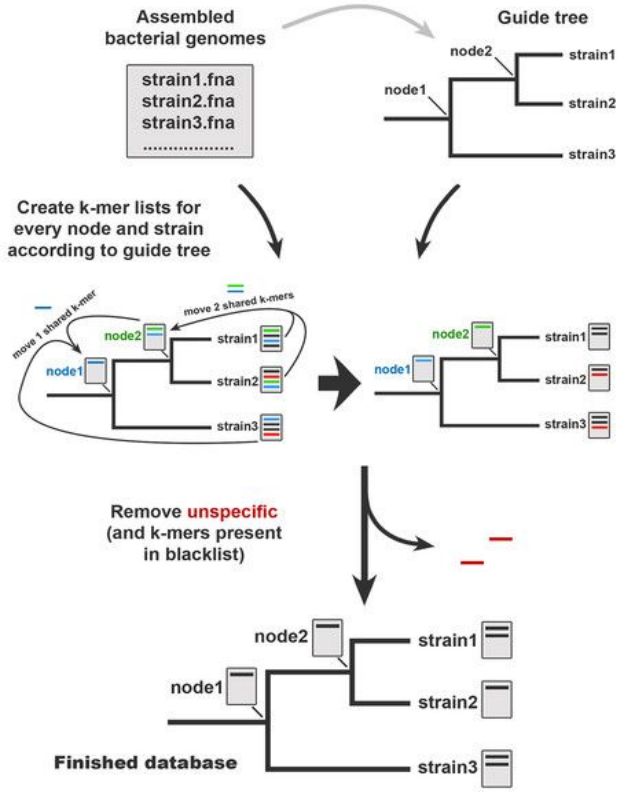
- Create tree using 250+ *Klebsiella* genomes and trimmed reads
- Identify representative references



Reference Assembly: Alfred-G

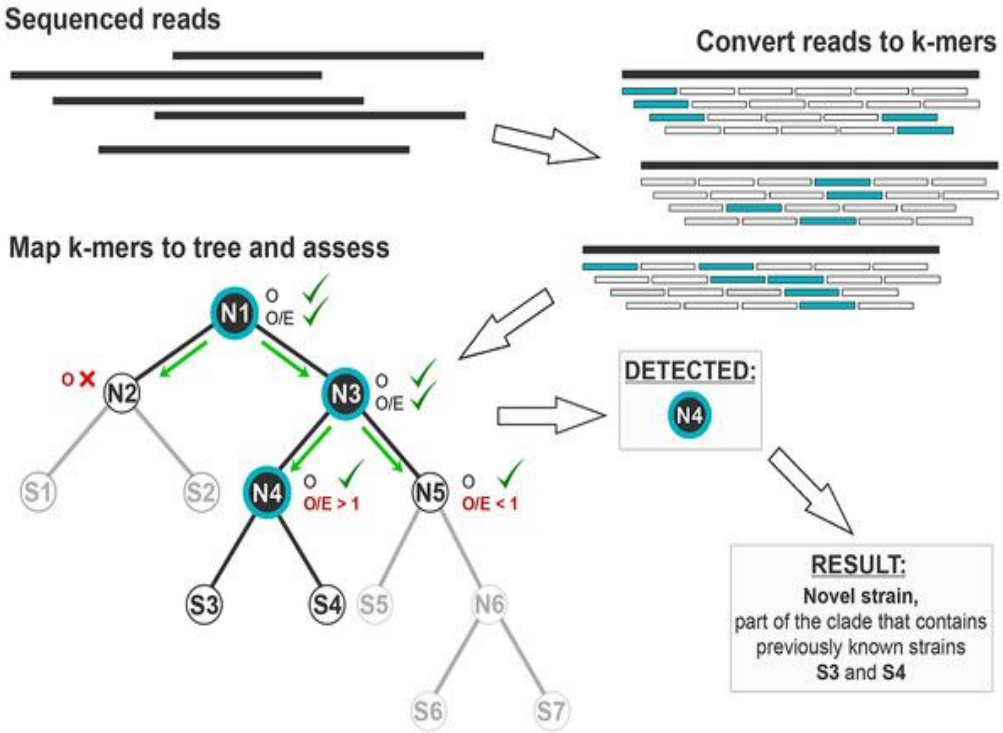
- ALignment FRee Distance (ALFRED)
- Creates Distance Matrix
- Converted to Newick Tree
    - Input for StrainSeeker

Reference Assembly: StrainSeeker (Building the Database)



Roosaare M et al. StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ. 2017;5:e3353.

Reference Assembly: StrainSeeker (Strain Identification)



Roosaare M et al. StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ. 2017;5:e3353.

Reference Assembly: StrainSeeker

- Detects bacterial strains from sequencing reads
- Establishes closest reference

```
==> SRR5666607_1_20.fastq_seeker_run <==
Sample:SRR5666607_1_20.fastq_seeker_run
87.46067%       RELATED NZ_CP01476

==> SRR5666608_1_20.fastq_seeker_run <==
Sample:SRR5666608_1_20.fastq_seeker_run
97.43076%       RELATED NZ_CP01250,NZ_CP00773

==> SRR5666609_1_20.fastq_seeker_run <==
Sample:SRR5666609_1_20.fastq_seeker_run
98.53028%       RELATED NZ_CP01250,NZ_CP00773
```

Reference Assembly: BOWTIE2

- Established reference bins
- Used BOWTIE2 to run reference assembly