Genome Assembly

Preliminary Results

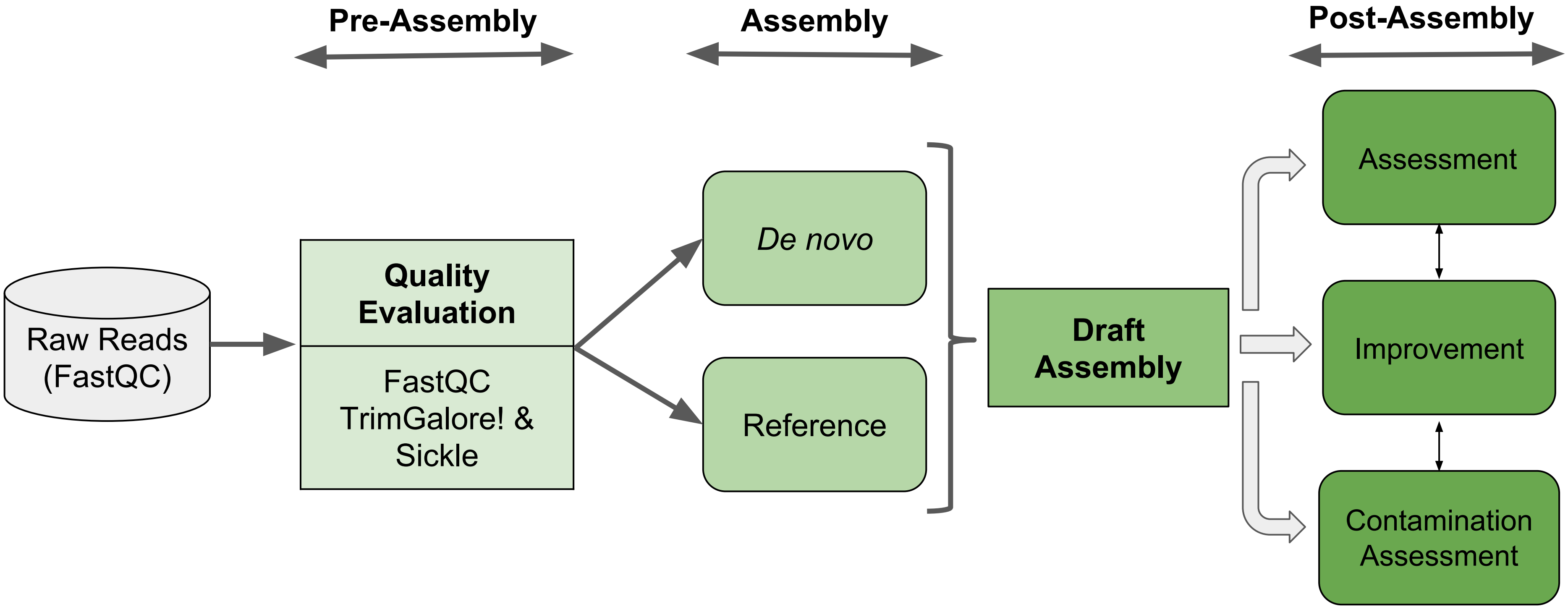
Group II

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

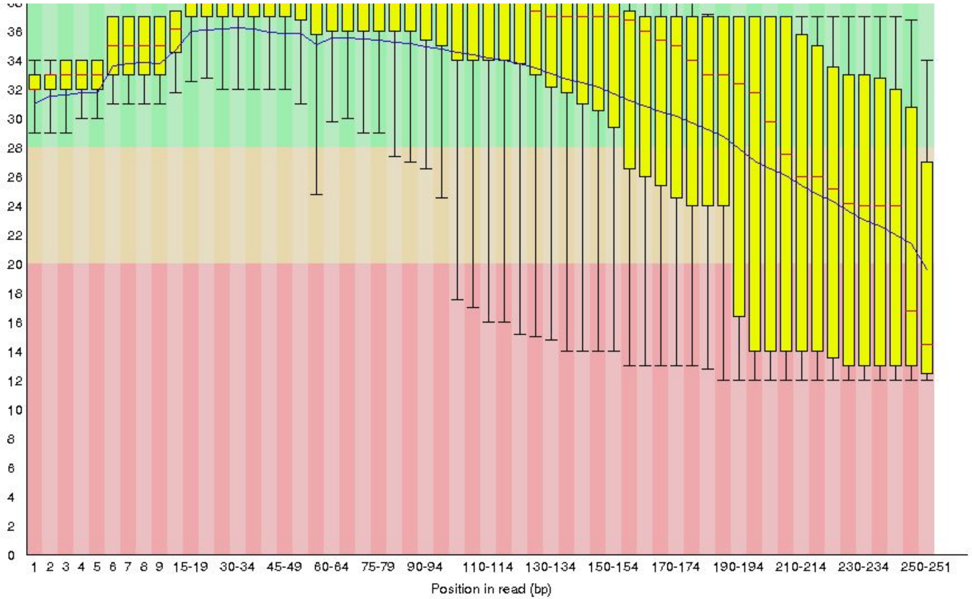
Outline

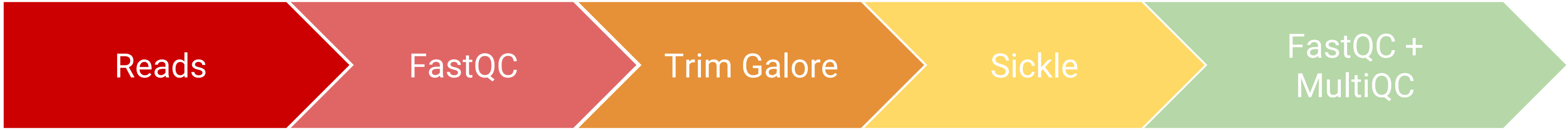
- Reminder of the task at hand
- Adapter trimming and quality control
- MultiQC
- Tool testing approach
- Reference Assembly
- De novo Assembly
- Scaffolding
- Next Steps

Introduction



Pre Assembly





Reads

FastQC

Trim Galore

Sickle

FastQC + MultiQC

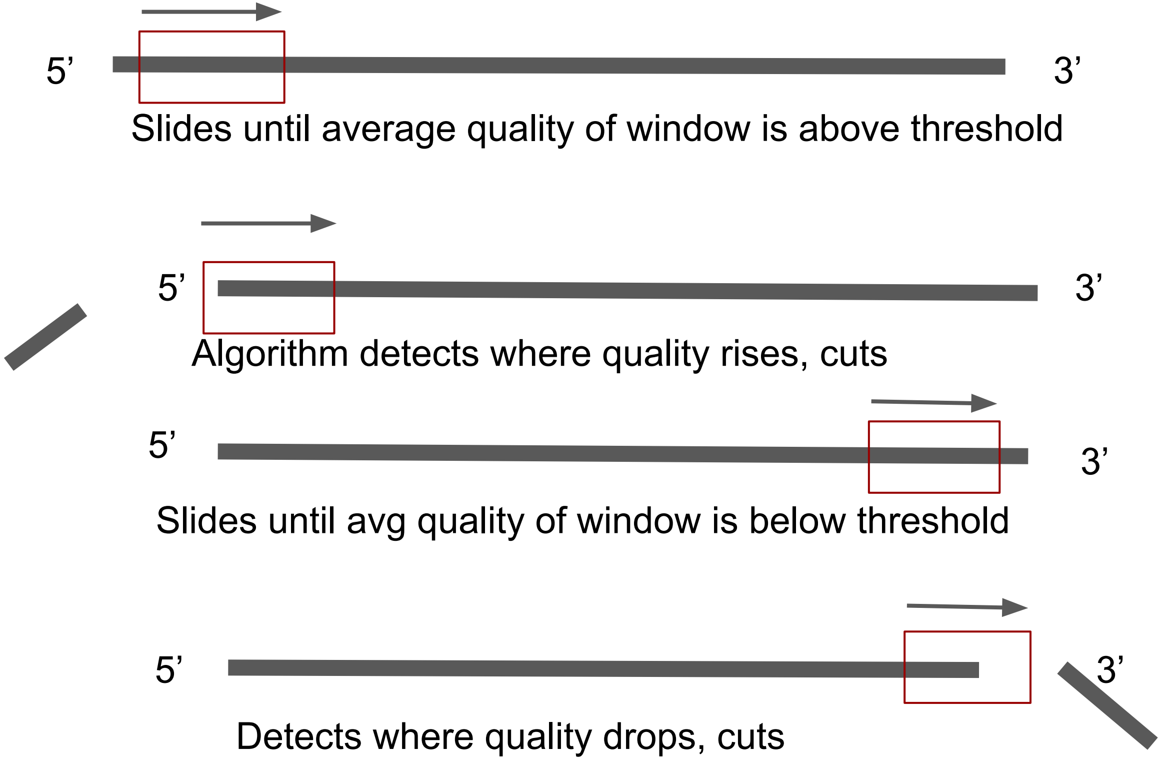
Sickle

- A windowed adaptive trimming tool for FastQ files

- Window slides along read with quality and length thresholds to determine when quality is sufficiently low to trim the 3' and 5’-end of reads

- Discard reads based upon the quality and length threshold

Sickle’s sliding window



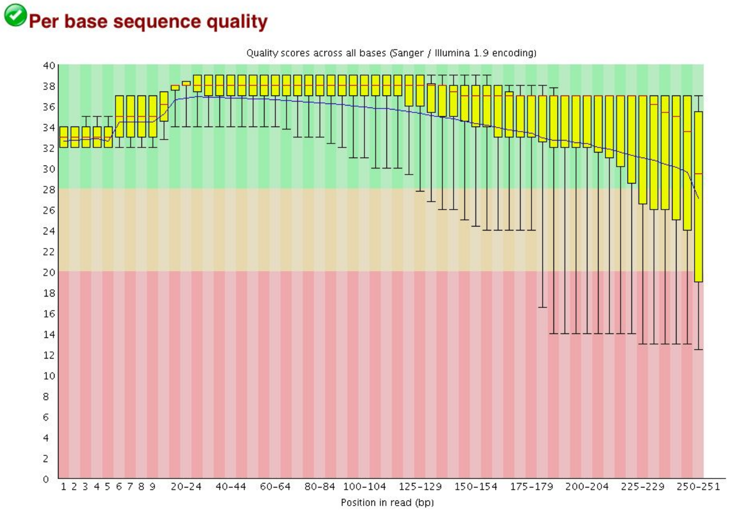
Slides until average quality of window is above threshold

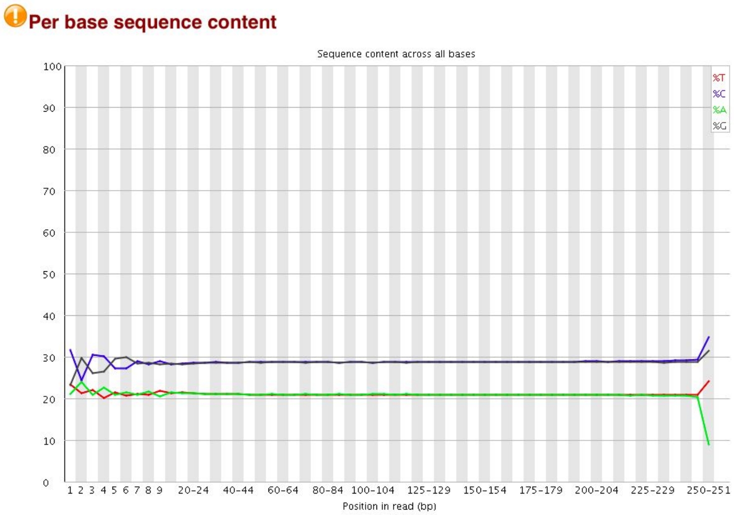
Algorithm detects where quality rises, cuts

Slides until avg quality of window is below threshold

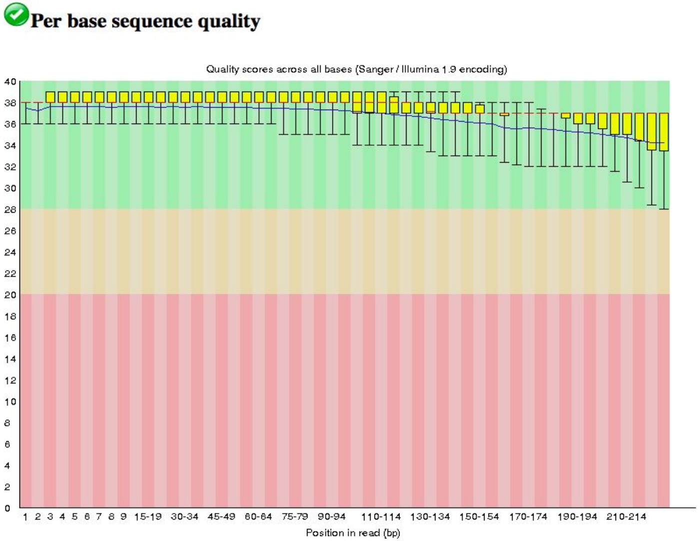
Detects where quality drops, cuts

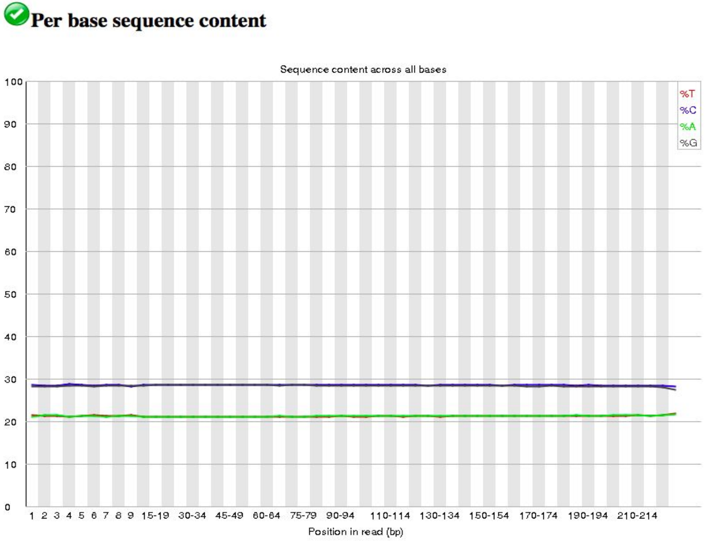
Before


Per base sequence quality


Per base sequence content

After


Per base sequence quality

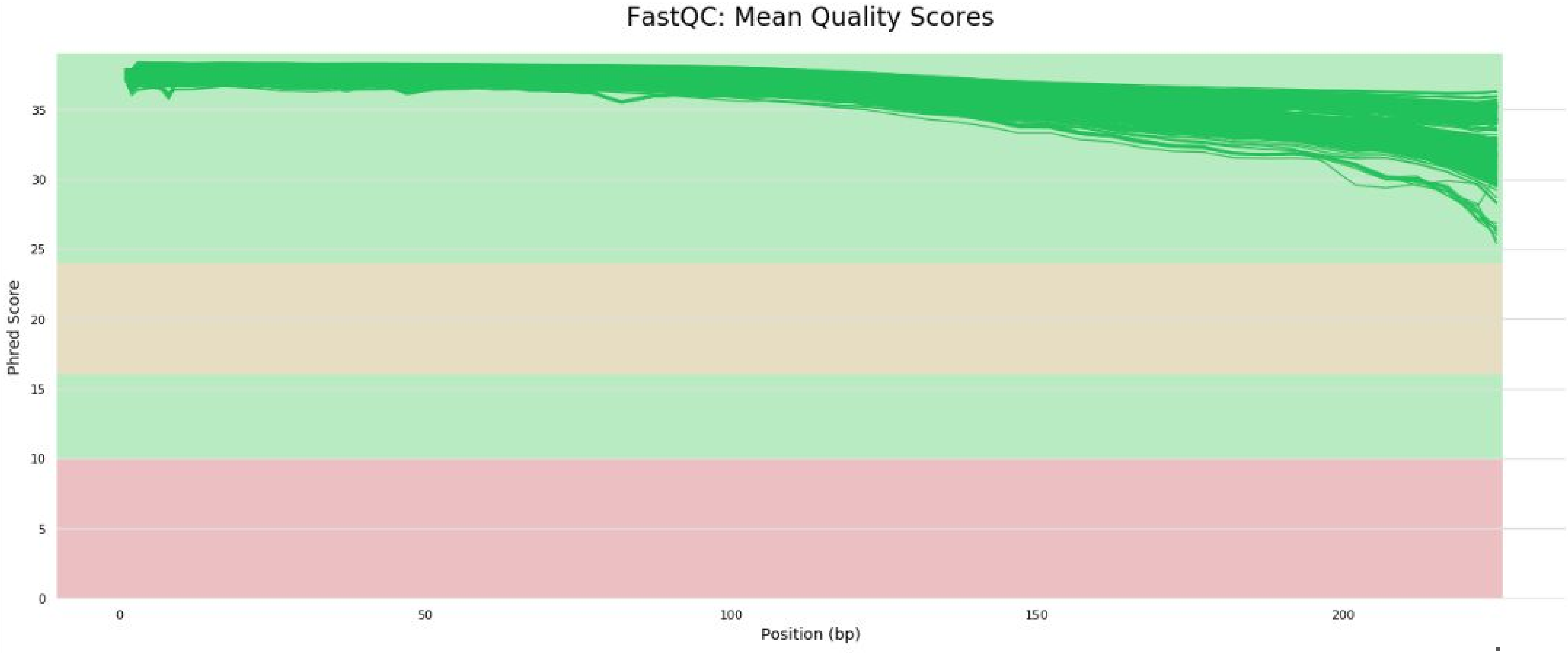

Per base sequence content

MultiQC

- MultiQC is a tool to create a single report with interactive plots for multiple bioinformatics analyses across many samples.

Aggregate results from bioinformatics analyses across many samples into a single report

MultiQC: Results



Reference Assembly

Reference Assembly

Problem:

- No information about species (cannot simply choose a *Klebsiella* species to use as reference genome)

Strategies:

- Test conservation between species (ANI + Mauve)
- Attempt to determine the species of each sample (StrainSeeker)
- Assemble samples using test reference genomes (Bowtie + SMALT)

Reference Assembly: Testing Conservation

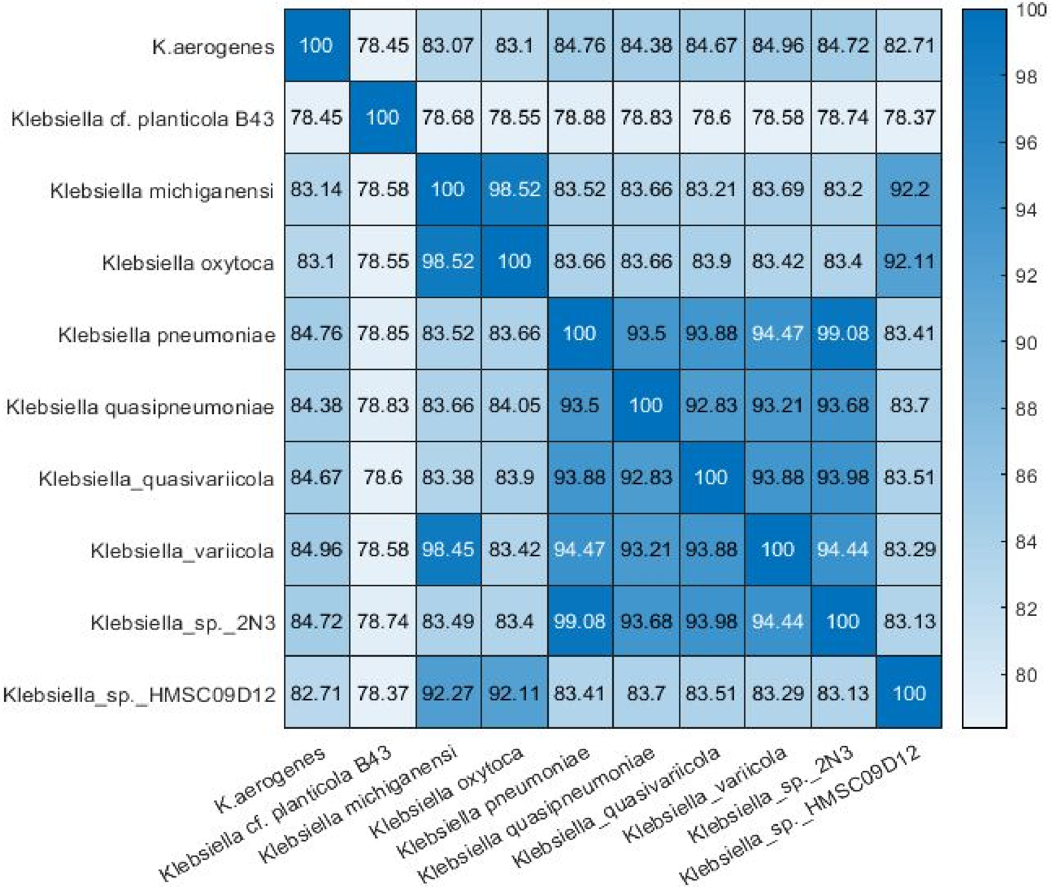
Test the level of conservation between different klebsiella species:

- selected the reference genomes for *K.aerogenes, K.oxytoca, K.pneumoniae, K. cf. planticola, K.quasipneumoniae, K.quasivariicola, K.variicola, K.sp. 2N3, K.sp. HMSC09D12*

- calculated the ANI percentages using OrthoANI tool

- whole genome alignment by Mauve

Pair-wise ANI: Results



	K.aerogenes	Klebsiella cf. planticola B43	Klebsiella michiganensi	Klebsiella oxytoca	Klebsiella pneumoniae	Klebsiella quasipneumoniae	Klebsiella_quasivariicola	Klebsiella_variicola	Klebsiella_sp._2N3	Klebsiella_sp._HMSC09D12
K.aerogenes	100	78.45	83.07	83.1	84.76	84.38	84.67	84.96	84.72	82.71
Klebsiella cf. planticola B43	78.45	100	78.68	78.55	78.88	78.83	78.6	78.58	78.74	78.37
Klebsiella michiganensi	83.14	78.58	100	98.52	83.52	83.66	83.21	83.69	83.2	92.2
Klebsiella oxytoca	83.1	78.55	98.52	100	83.66	83.66	83.9	83.42	83.4	92.11
Klebsiella pneumoniae	84.76	78.85	83.52	83.66	100	93.5	93.88	94.47	99.08	83.41
Klebsiella quasipneumoniae	84.38	78.83	83.66	84.05	93.5	100	92.83	93.21	93.68	83.7
Klebsiella_quasivariicola	84.67	78.6	83.38	83.9	93.88	92.83	100	93.88	93.98	83.51
Klebsiella_variicola	84.96	78.58	98.45	83.42	94.47	93.21	93.88	100	94.44	83.29
Klebsiella_sp._2N3	84.72	78.74	83.49	83.4	99.08	93.68	93.98	94.44	100	83.13
Klebsiella_sp._HMSC09D12	82.71	78.37	92.27	92.11	83.41	83.7	83.51	83.29	83.13	100

Reference Assembly: StrainSeeker

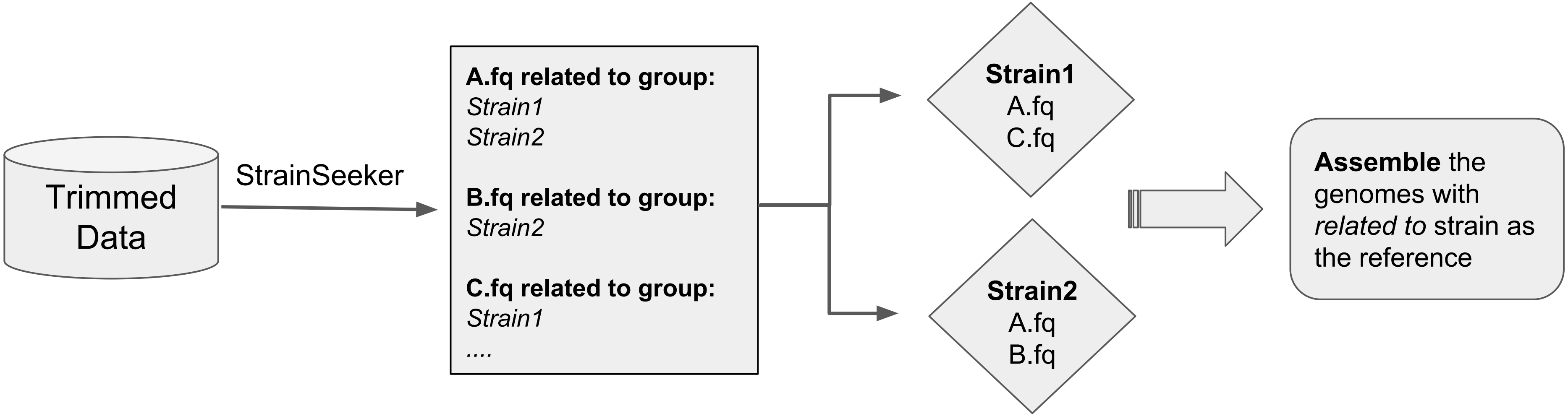
- Tool to classify/evaluate the species from each file after trimming

- Broad Goal:
  - Strains falling under same *related to* group can be binned together and reference assembly can be done

Welcome to *StrainSeeker*

sequencing read analyzer for detecting bacterial strains

Reference Assembly: StrainSeeker



Trimmed Data

StrainSeeker

A.fq related to group:
Strain1
Strain2

B.fq related to group:
Strain2

C.fq related to group:
Strain1
....

Strain1
A.fq
C.fq

Strain2
A.fq
B.fq

Assemble the genomes with *related to* strain as the reference

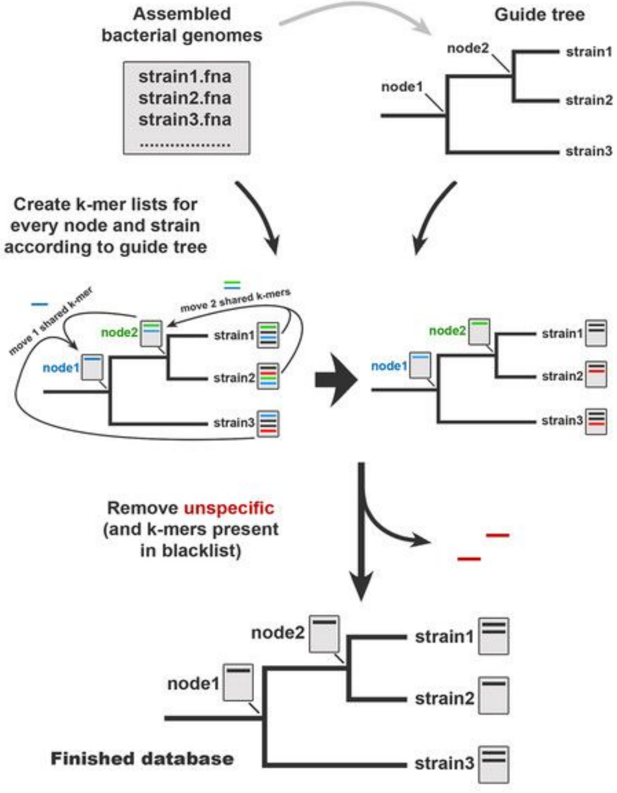
StrainSeeker groups unassembled reads under closely related species

Reference Assembly: StrainSeeker Methodology

1. Build a reference database
   a. build guide tree of this database
   b. assembled genome converted to k-mer list
   c. Guide tree of k-mer list is created by moving shared k-mers to the root.
   d. Cleaning up of database by removing non-specific k-mers

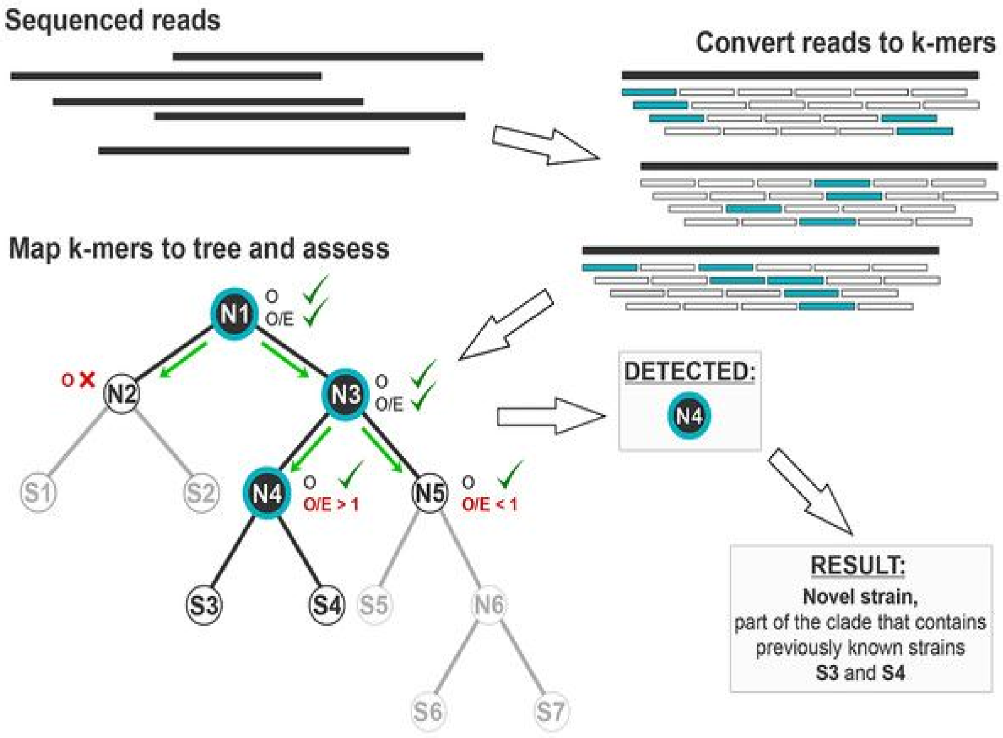
2. Strain Identification
   a. Reads are converted to k-mers
   b. Mapping to database tree

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)



RESULT:
Novel strain,
part of the clade that contains
previously known strains
S3 and S4

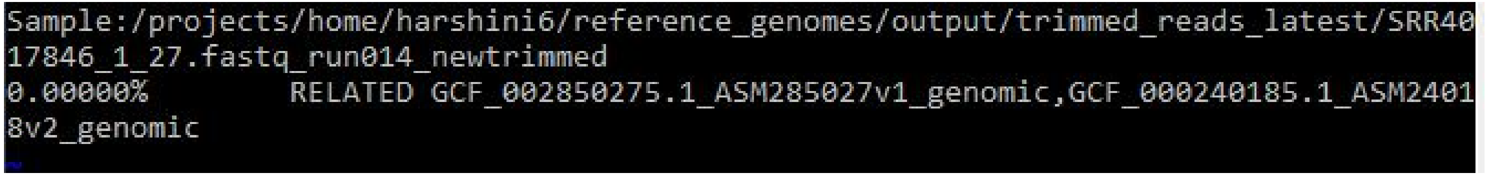
Roosaare M et al. StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ. 2017;5:e3353.

Reference Assembly: StrainSeeker Results

- 11 reference genomes of *Klebsiella* were used to build the database
- ran StrainSeeker for 10 selected files after trimming
- observed that K.pneumoniae and K. sp N13 appear most common in the *related to* group

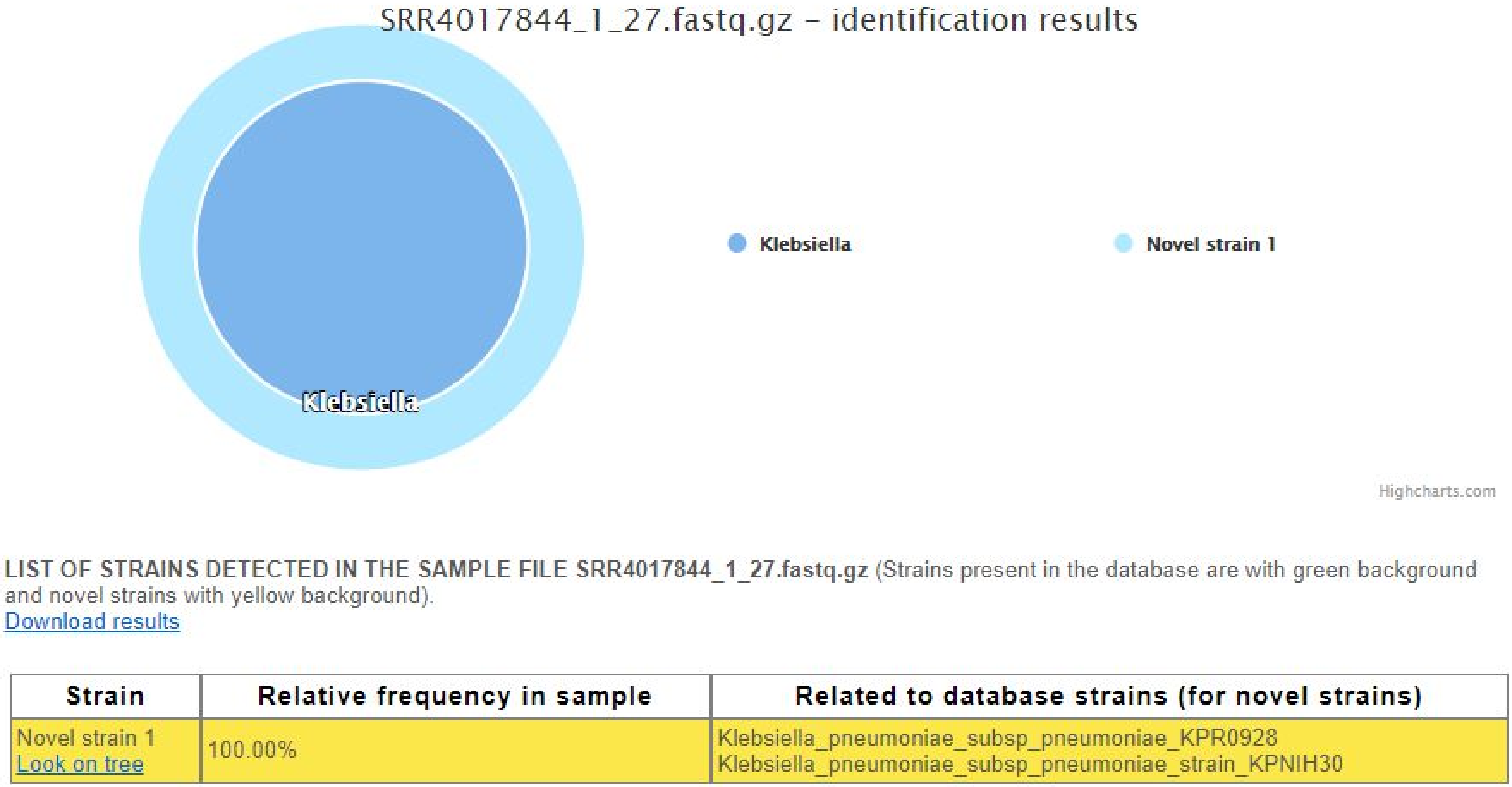
Reference Assembly: StrainSeeker Issues

- Need to build a comprehensive database covering maximum *Klebsiella* reference genomes (we only used 10)
- Need to figure out the 0.000% in the results (read more about interpreting the results)



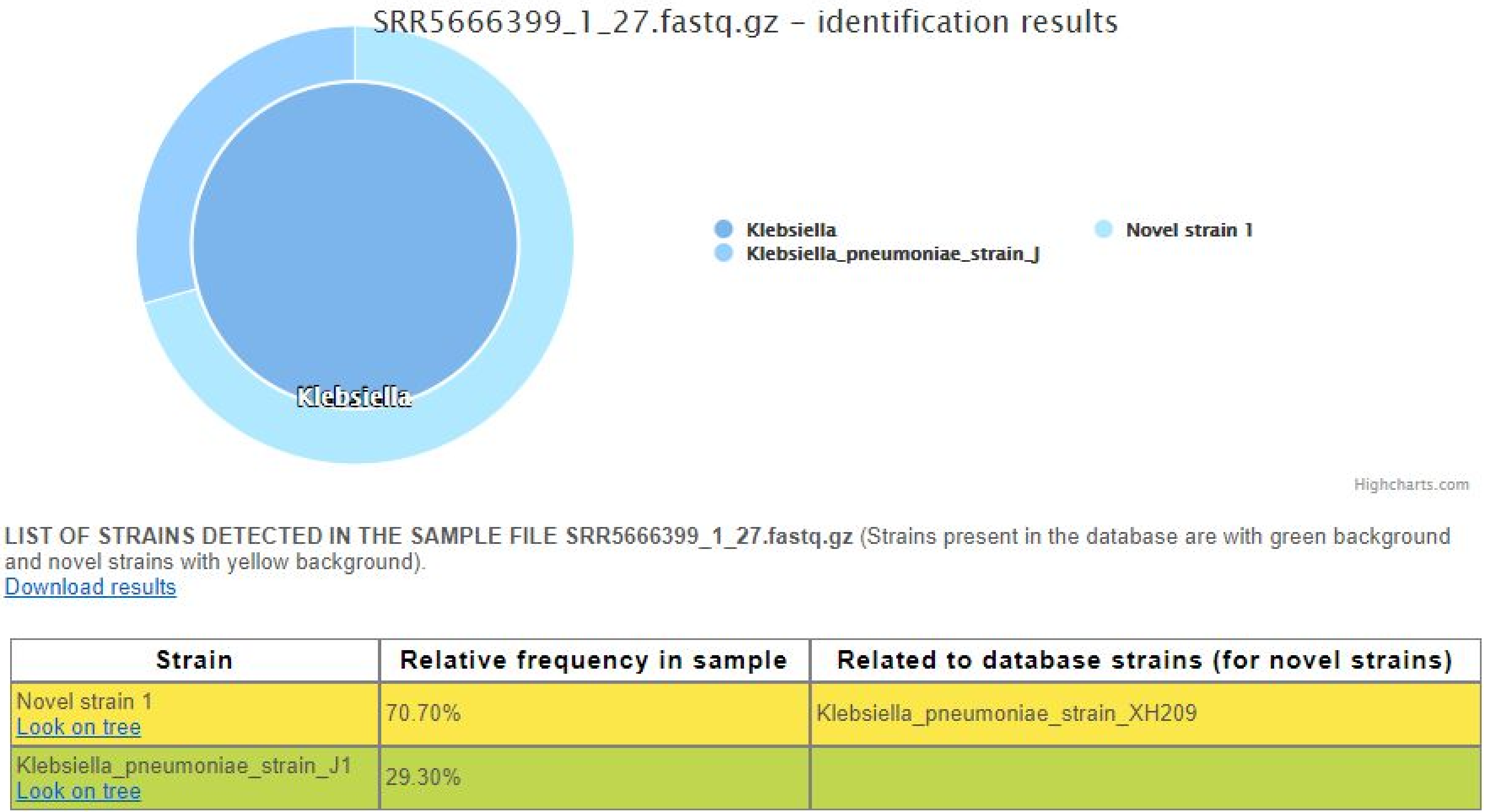
```
Sample:/projects/home/harshini6/reference_genomes/output/trimmed_reads_latest/SRR40
17846_1_27.fastq_run014_newtrimmed
0.00000%        RELATED GCF_002850275.1_ASM285027v1_genomic,GCF_000240185.1_ASM2401
8v2_genomic
~
```

- StrainSeeker Web tool is not suitable for our number of files
    - Only takes in 5 input files at a time
    - Our results for 5 input read files: http://bioinfo.ut.ee/strainseeker/index.php?r=site/results&id=66nrnjqc4a9fb22p1bia9i6km034387&additional=2a3a4a5a



LIST OF STRAINS DETECTED IN THE SAMPLE FILE SRR4017844_1_27.fastq.gz (Strains present in the database are with green background and novel strains with yellow background).
Download results

Strain	Relative frequency in sample	Related to database strains (for novel strains)
Novel strain 1 Look on tree	100.00%	Klebsiella_pneumoniae_subsp_pneumoniae_KPR0928 Klebsiella_pneumoniae_subsp_pneumoniae_strain_KPNIH30


SRR5666399_1_27.fastq.gz – identification results
Klebsiella
Klebsiella
Klebsiella_pneumoniae_strain_J
Novel strain 1
Highcharts.com

LIST OF STRAINS DETECTED IN THE SAMPLE FILE SRR5666399_1_27.fastq.gz (Strains present in the database are with green background and novel strains with yellow background).

Download results

Strain	Relative frequency in sample	Related to database strains (for novel strains)
Novel strain 1 Look on tree	70.70%	Klebsiella_pneumoniae_strain_XH209
Klebsiella_pneumoniae_strain_J1 Look on tree	29.30%	

Reference Assembly: Assembly Tools

Perform reference assembly on a subset of trimmed data using different reference genomes and assembly tools

Reference Genomes
K.pneumoniae
K.oxytoca

Reference Assembly Tools
Bowtie2
SMALT

Chose the 10 largest read files for assembly

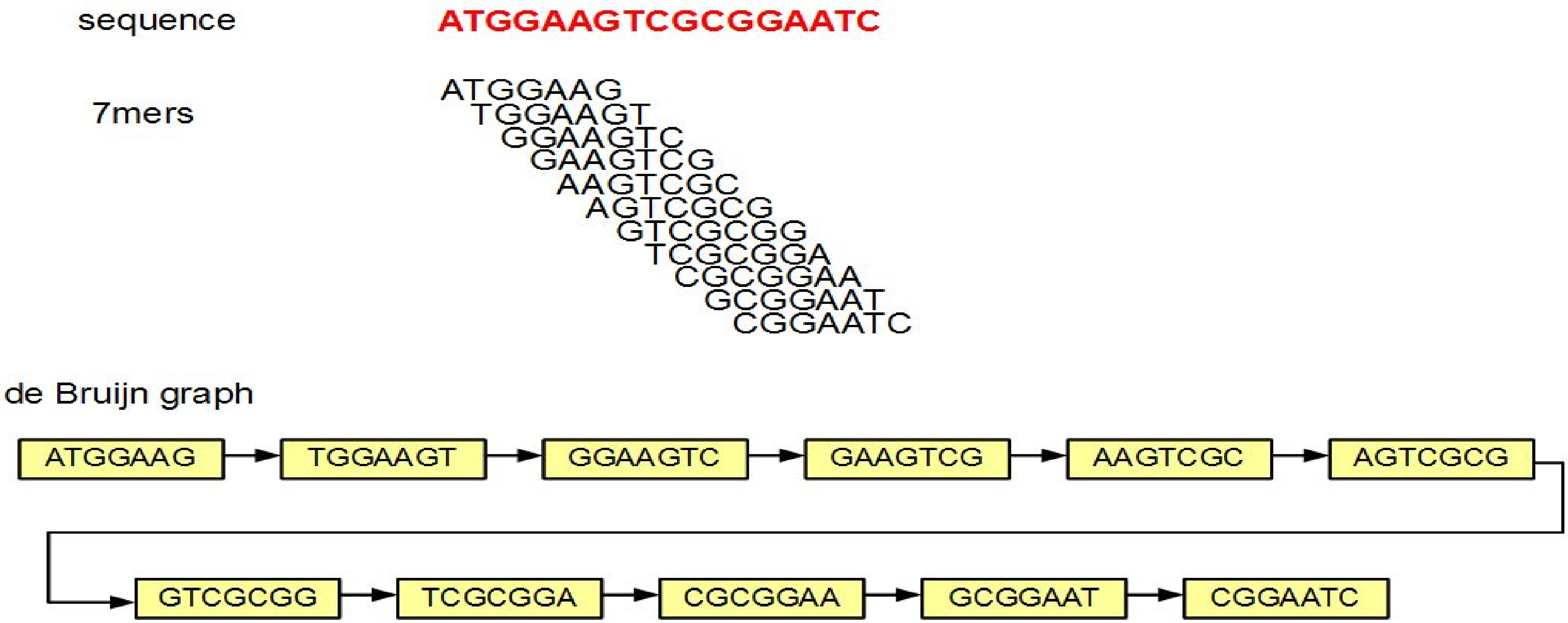
Reference Assembly: Results

BOWTIE2	*K.pneumoniae*	*K.oxytoca*
Time for 10 alignments	17 minutes	17 minutes
Coverage	97%	No contig larger than 7%

SMALT	*K.pneumoniae*	*K.oxytoca*
Approx time for one alignment	17 minutes	26 minutes
Coverage	97%	No contig larger than 58%

De Novo Assembly

De bruijn graphs



sequence ATGGAAGTCGCGGAATC

7mers

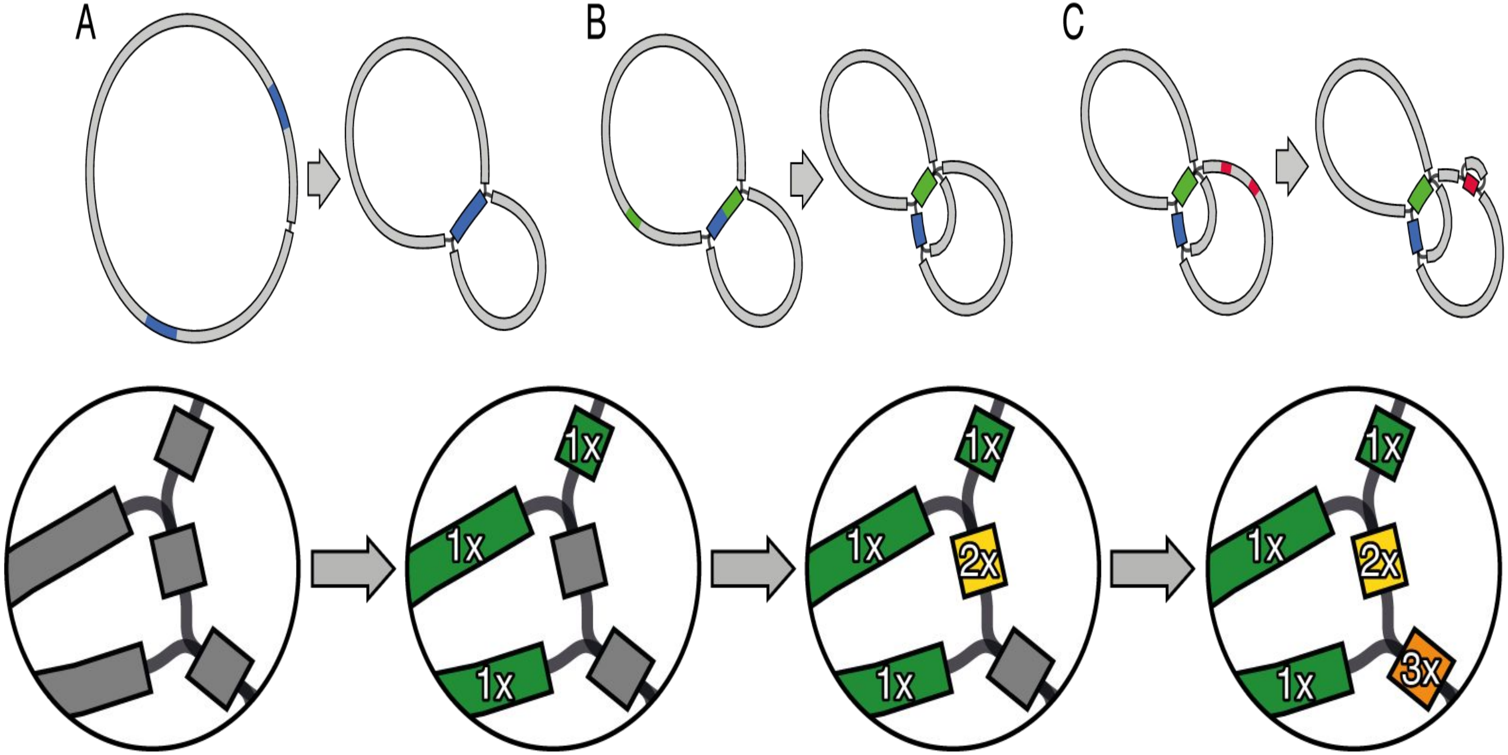
```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGC
     AGTCGCG
      GTCGCGG
       TCGCGGA
        CGCGGAA
         GCGGAAT
          CGGAATC
```

de Bruijn graph

ATGGAAG → TGGAAGT → GGAAGTC → GAAGTCG → AAGTCGC → AGTCGCG → GTCGCGG → TCGCGGA → CGCGGAA → GCGGAAT → CGGAATC

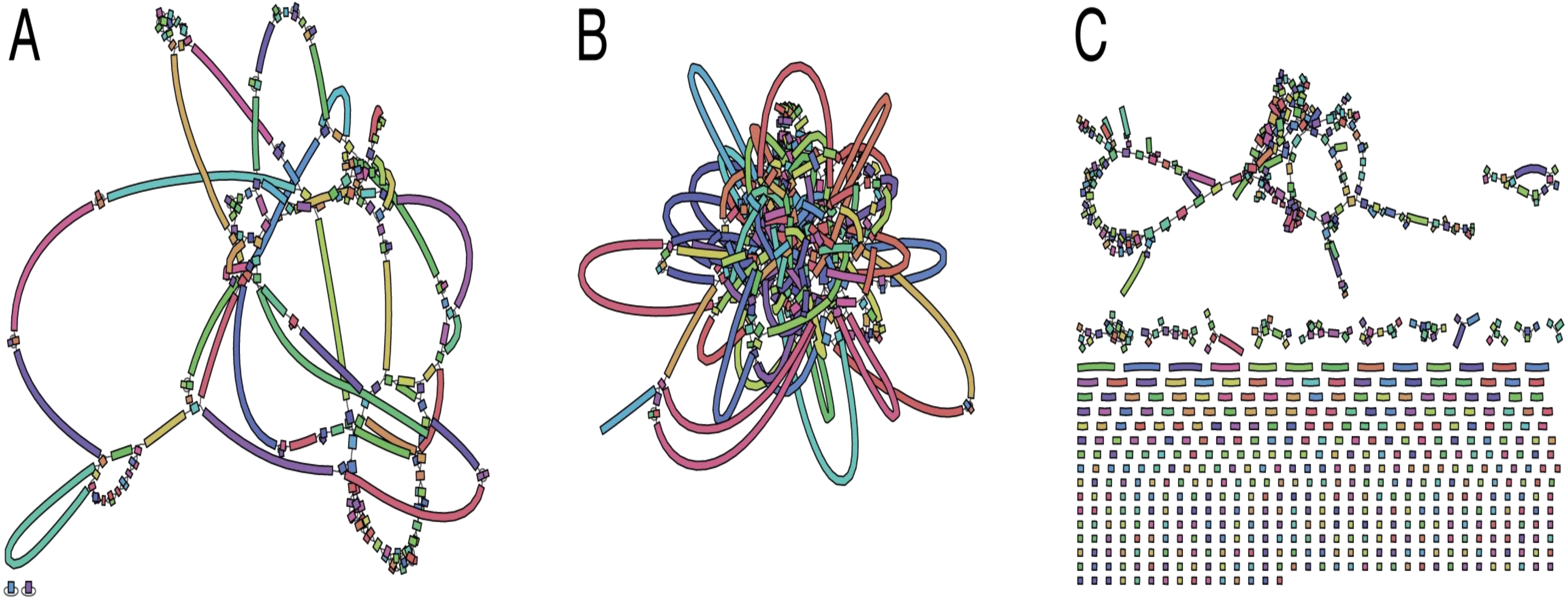
Source: http://homolog.us/blogs/

De bruijn graphs - repeat regions



Source: https://github.com/rrwick/Unicycler

De bruijn graphs - whole genome

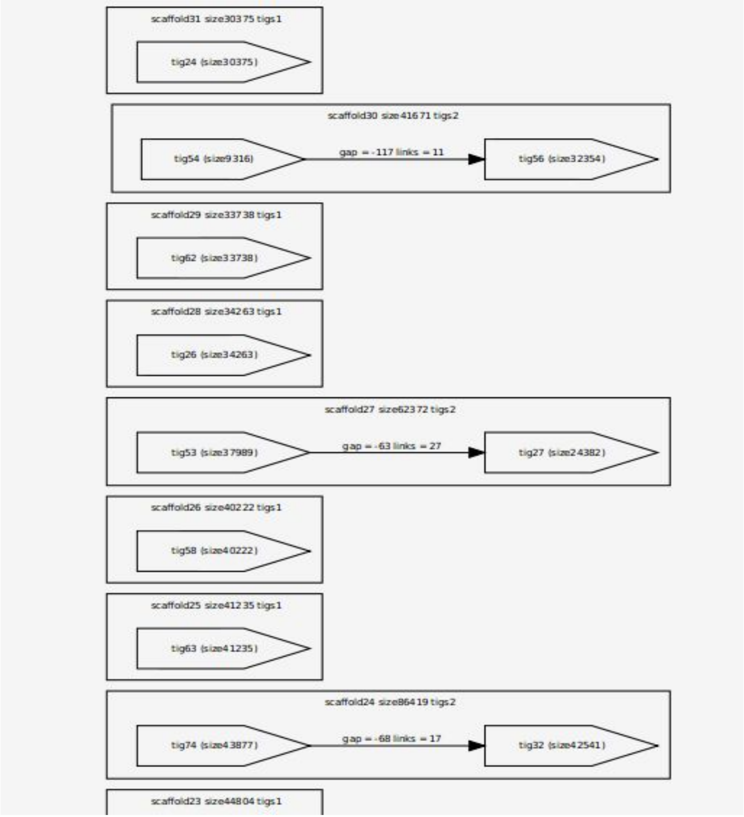


Source: https://github.com/rrwick/Unicycler

De Novo Assembly: Tools

1. SPAdes
2. MaSuRCA
3. Unicycler
4. Skesa

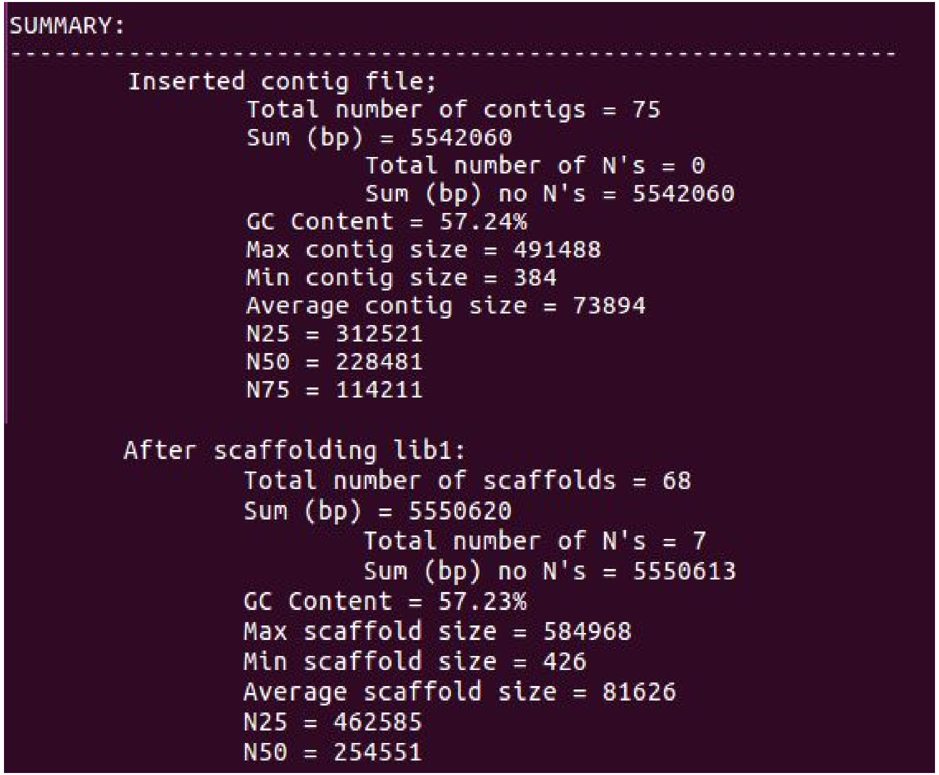
De Novo Assembly: Scaffolding



- A dotfile showing representation of a section of scaffolds in order

- In the case of contig extensions, the gap length is shown and the contigs are placed together

De Novo Assembly: Scaffolding Results

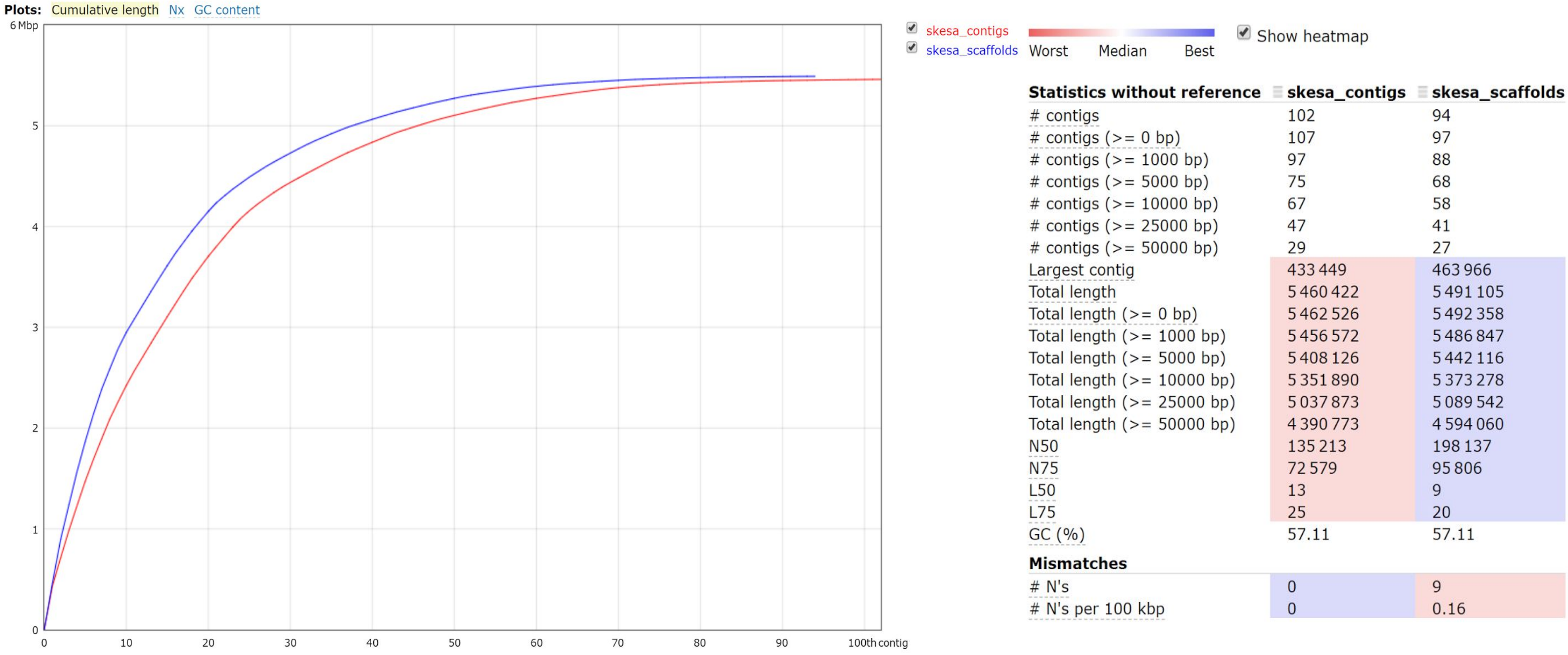


```
SUMMARY:
------------------------------------------------------------
        Inserted contig file;
                Total number of contigs = 75
                Sum (bp) = 5542060
                        Total number of N's = 0
                        Sum (bp) no N's = 5542060
                GC Content = 57.24%
                Max contig size = 491488
                Min contig size = 384
                Average contig size = 73894
                N25 = 312521
                N50 = 228481
                N75 = 114211

        After scaffolding lib1:
                Total number of scaffolds = 68
                Sum (bp) = 5550620
                        Total number of N's = 7
                        Sum (bp) no N's = 5550613
                GC Content = 57.23%
                Max scaffold size = 584968
                Min scaffold size = 426
                Average scaffold size = 81626
                N25 = 462585
                N50 = 254551
```

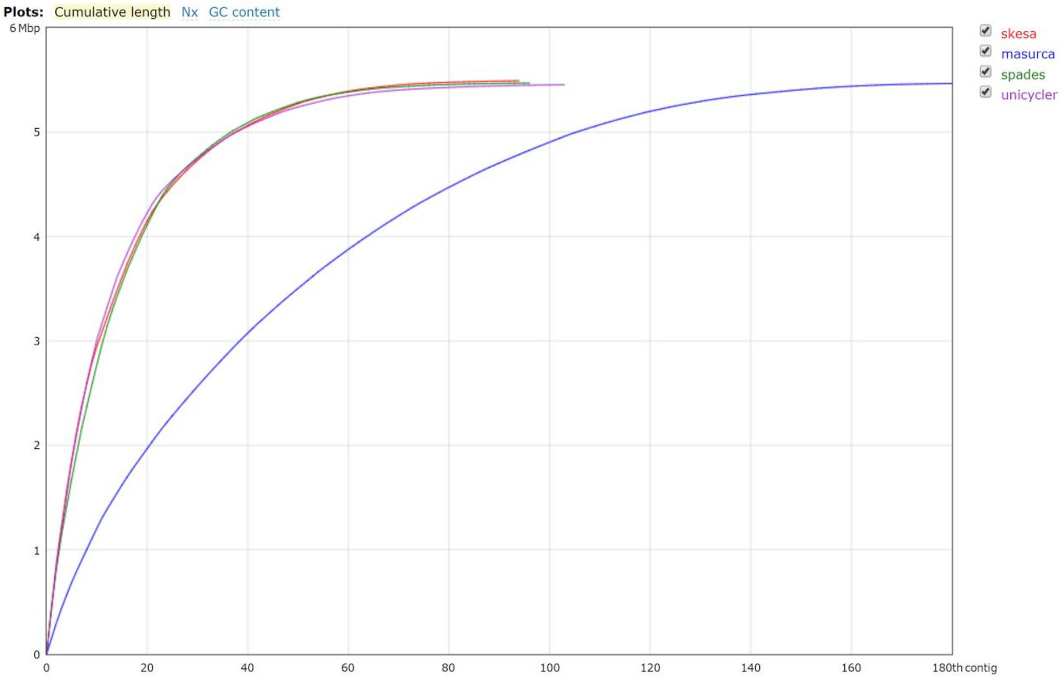
- Used a minimum of 5 links (read pairs) between contigs to compute contigs

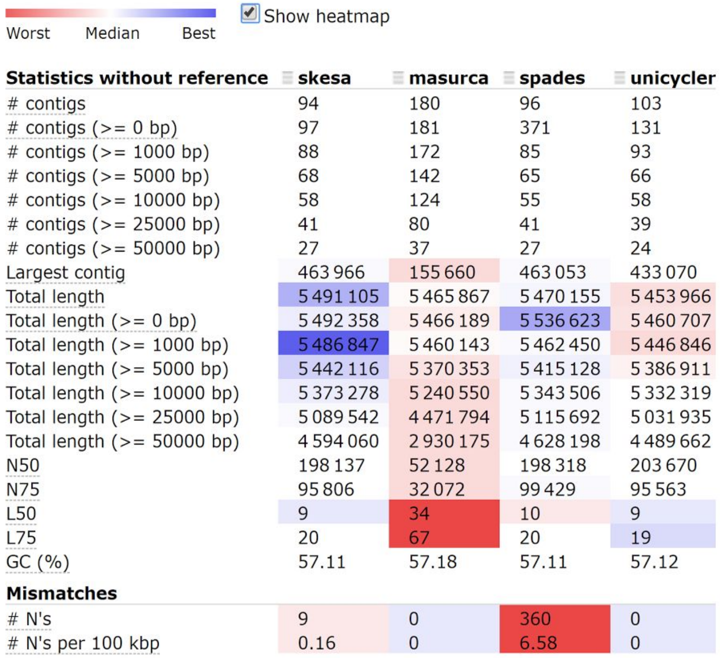
De Novo Assembly: Scaffolding Improvement



Statistics without reference	skesa_contigs	skesa_scaffolds
# contigs	102	94
# contigs (>= 0 bp)	107	97
# contigs (>= 1000 bp)	97	88
# contigs (>= 5000 bp)	75	68
# contigs (>= 10000 bp)	67	58
# contigs (>= 25000 bp)	47	41
# contigs (>= 50000 bp)	29	27
Largest contig	433 449	463 966
Total length	5 460 422	5 491 105
Total length (>= 0 bp)	5 462 526	5 492 358
Total length (>= 1000 bp)	5 456 572	5 486 847
Total length (>= 5000 bp)	5 408 126	5 442 116
Total length (>= 10000 bp)	5 351 890	5 373 278
Total length (>= 25000 bp)	5 037 873	5 089 542
Total length (>= 50000 bp)	4 390 773	4 594 060
N50	135 213	198 137
N75	72 579	95 806
L50	13	9
L75	25	20
GC (%)	57.11	57.11
Mismatches		
# N's	0	9
# N's per 100 kbp	0	0.16

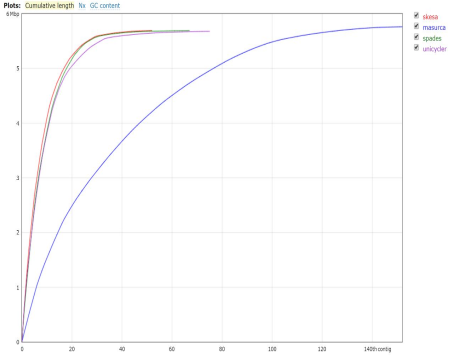
De Novo Assembly: Results

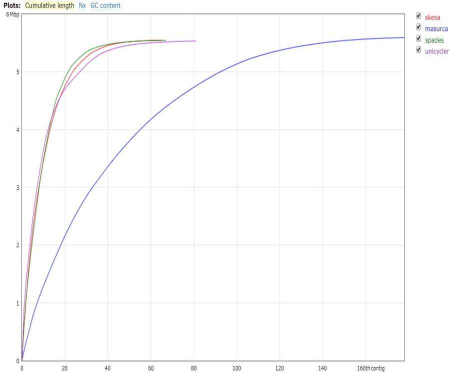


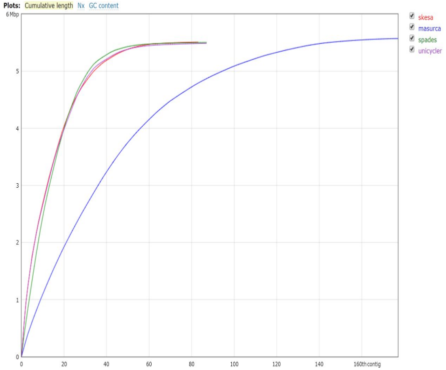


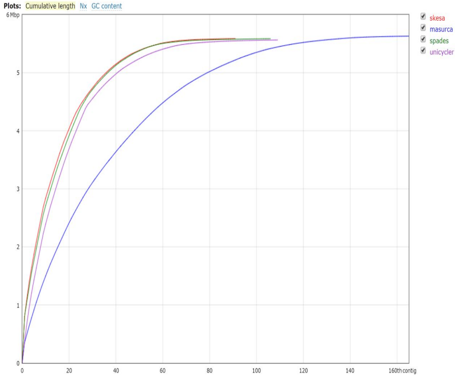
Worst
Median
Best
☑ Show heatmap

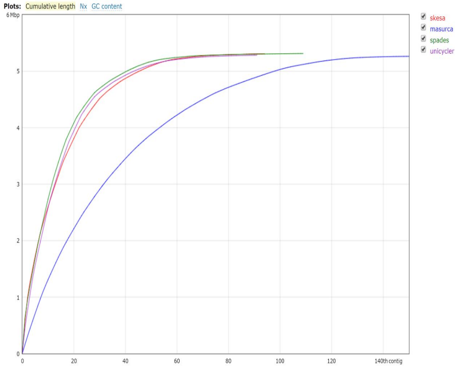
Statistics without reference	skesa	masurca	spades	unicycler
# contigs	94	180	96	103
# contigs (>= 0 bp)	97	181	371	131
# contigs (>= 1000 bp)	88	172	85	93
# contigs (>= 5000 bp)	68	142	65	66
# contigs (>= 10000 bp)	58	124	55	58
# contigs (>= 25000 bp)	41	80	41	39
# contigs (>= 50000 bp)	27	37	27	24
Largest contig	463 966	155 660	463 053	433 070
Total length	5 491 105	5 465 867	5 470 155	5 453 966
Total length (>= 0 bp)	5 492 358	5 466 189	5 536 623	5 460 707
Total length (>= 1000 bp)	5 486 847	5 460 143	5 462 450	5 446 846
Total length (>= 5000 bp)	5 442 116	5 370 353	5 415 128	5 386 911
Total length (>= 10000 bp)	5 373 278	5 240 550	5 343 506	5 332 319
Total length (>= 25000 bp)	5 089 542	4 471 794	5 115 692	5 031 935
Total length (>= 50000 bp)	4 594 060	2 930 175	4 628 198	4 489 662
N50	198 137	52 128	198 318	203 670
N75	95 806	32 072	99 429	95 563
L50	9	34	10	9
L75	20	67	20	19
GC (%)	57.11	57.18	57.11	57.12
Mismatches				
# N's	9	0	360	0
# N's per 100 kbp	0.16	0	6.58	0

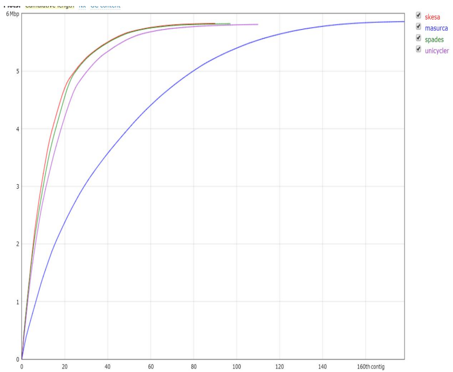

Plots: Cumulative length Nx GC content
6 Mbp
5
4
3
2
1
0
0
20
40
60
80
100
120
140th contig
☑ skesa
☑ masurca
☑ spades
☑ unicycler

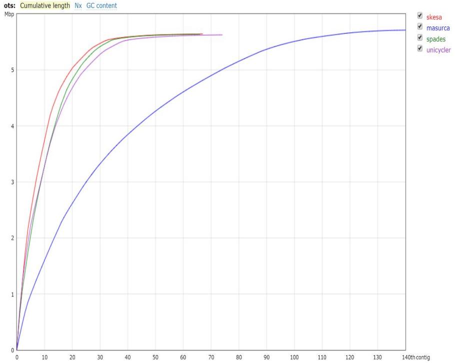

Plots: Cumulative length Nx GC content
6 Mbp
5
4
3
2
1
0
0
20
40
60
80
100
120
140
160th contig
☑ skesa
☑ masurca
☑ spades
☑ unicycler

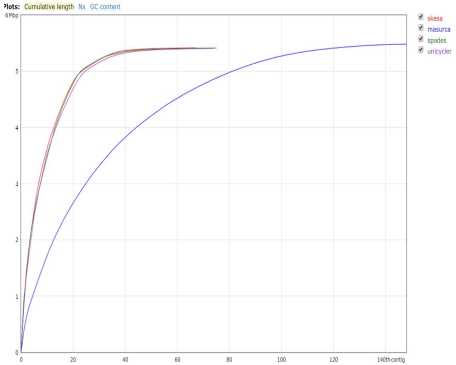

Plots: Cumulative length Nx GC content
6 Mbp
5
4
3
2
1
0
0
20
40
60
80
100
120
140
160th contig
☑ skesa
☑ masurca
☑ spades
☑ unicycler


Plots: Cumulative length Nx GC content
6 Mbp
5
4
3
2
1
0
0
20
40
60
80
100
120
140
160th contig
☑ skesa
☑ masurca
☑ spades
☑ unicycler


Plots: Cumulative length Nx GC content
6 Mbp
5
4
3
2
1
0
0
20
40
60
80
100
120
140th contig
☑ skesa
☑ masurca
☑ spades
☑ unicycler

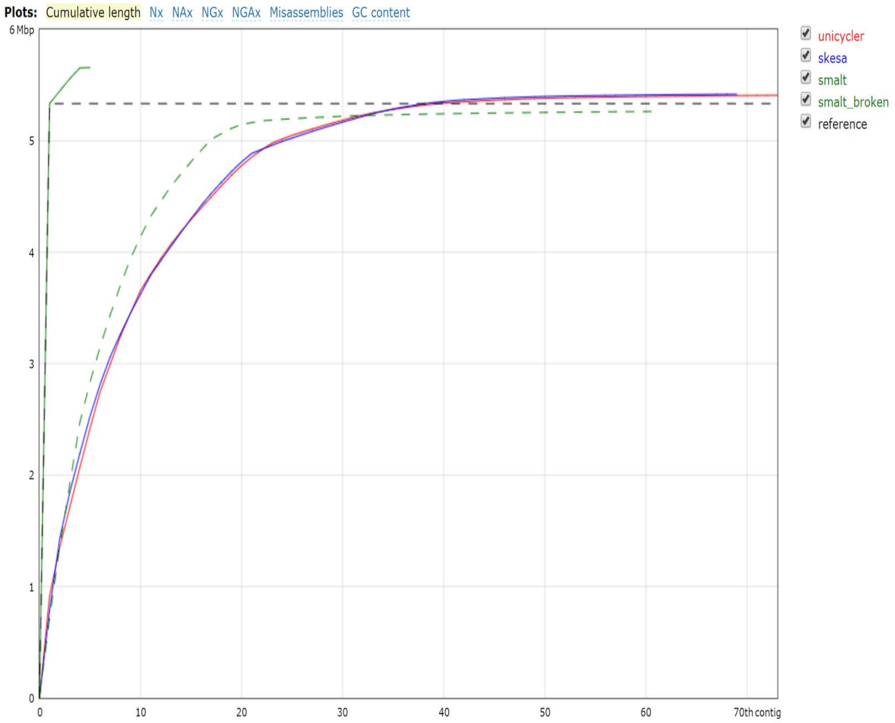

6 Mbp
5
4
3
2
1
0
0
20
40
60
80
100
120
140
160th contig
☑ skesa
☑ masurca
☑ spades
☑ unicycler

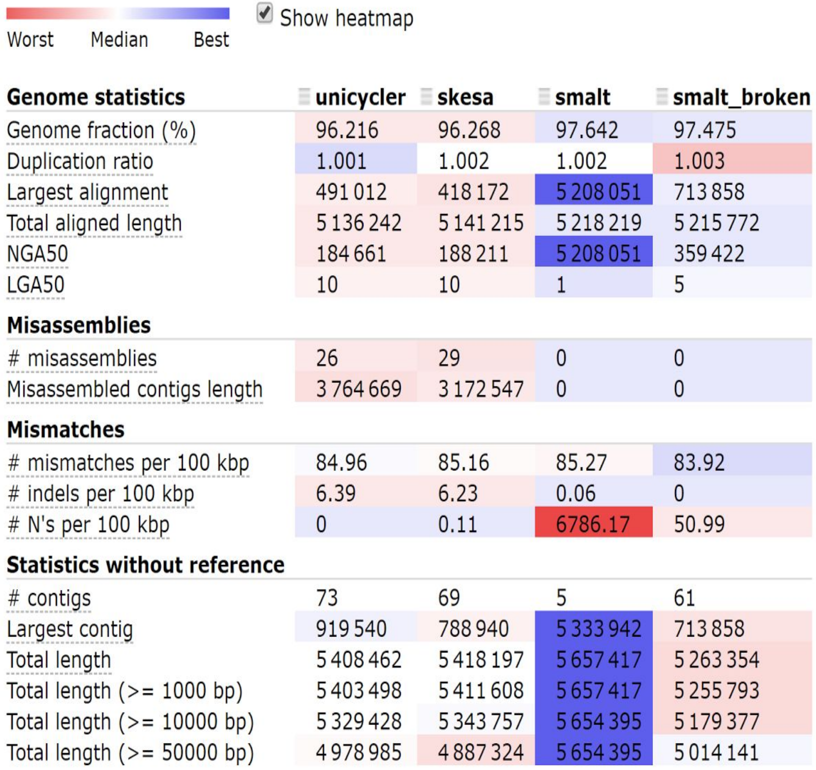

ots: Cumulative length Nx GC content
Mbp
5
4
3
2
1
0
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140th contig
☑ skesa
☑ masurca
☑ spades
☑ unicycler


lots: Cumulative length Nx GC content
6 Mbp
5
4
3
2
1
0
0
20
40
60
80
100
120
140th contig
☑ skesa
☑ masurca
☑ spades
☑ unicycler

De Novo vs. Reference

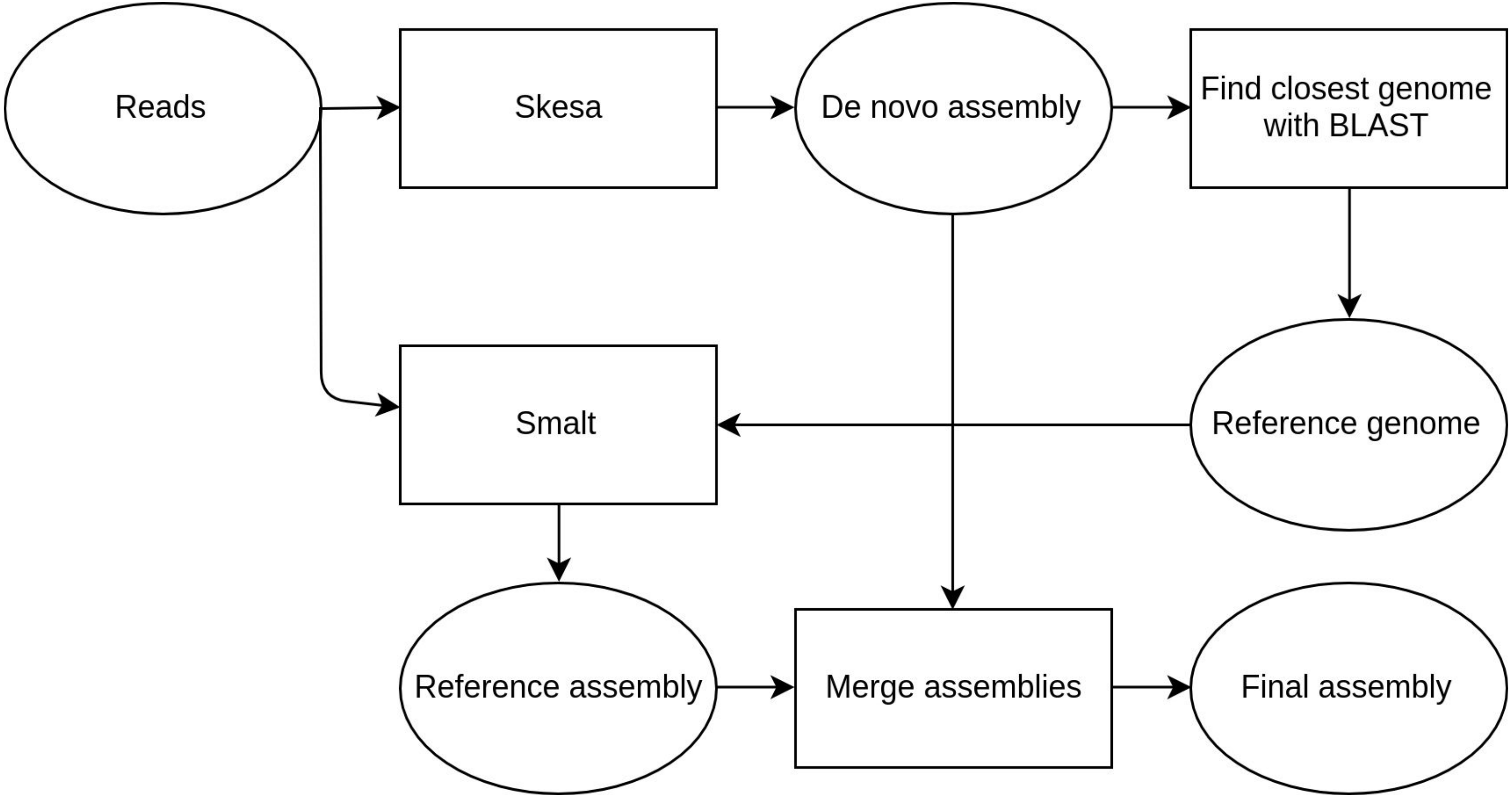
**Klebsiella pneumoniae* as reference





Genome statistics	unicycler	skesa	smalt	smalt_broken
Genome fraction (%)	96.216	96.268	97.642	97.475
Duplication ratio	1.001	1.002	1.002	1.003
Largest alignment	491 012	418 172	5 208 051	713 858
Total aligned length	5 136 242	5 141 215	5 218 219	5 215 772
NGA50	184 661	188 211	5 208 051	359 422
LGA50	10	10	1	5
Misassemblies				
# misassemblies	26	29	0	0
Misassembled contigs length	3 764 669	3 172 547	0	0
Mismatches				
# mismatches per 100 kbp	84.96	85.16	85.27	83.92
# indels per 100 kbp	6.39	6.23	0.06	0
# N's per 100 kbp	0	0.11	6786.17	50.99
Statistics without reference				
# contigs	73	69	5	61
Largest contig	919 540	788 940	5 333 942	713 858
Total length	5 408 462	5 418 197	5 657 417	5 263 354
Total length (>= 1000 bp)	5 403 498	5 411 608	5 657 417	5 255 793
Total length (>= 10000 bp)	5 329 428	5 343 757	5 654 395	5 179 377
Total length (>= 50000 bp)	4 978 985	4 887 324	5 654 395	5 014 141

Next Steps: Combined Pipeline



Next Steps

- Gap filling the scaffolds, error checking and contamination assessment
- Test more data
- Make a pipeline for assembly
- Deliver quality scaffolds the gene prediction team

Questions?