



# Genome Assembly Preliminary Results

Team 1: Vasanta, Qinwei, Tianze, Seonggeon, Kunal, Siarhei, Victoria,  
Nirav, Hunter

Presented by Nirav Shah, Hunter Seabolt, and SRR5666627

# Presentation Outline

- Trim/QC
  - Workflow
  - Examples
  - Comparison and selection of trimming software
  - Clean Data
- 
- Assembly and Preliminary Results
  - Workflow
  - Biological Considerations
  - Reference Based genome assembly
  - de novo genome assembly
  - Comparison of de novo assemblers
  - Post-processing
  - QC
  - Visualization

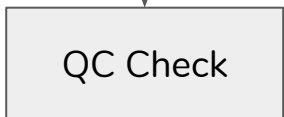
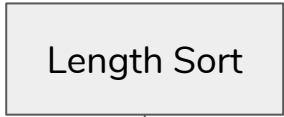
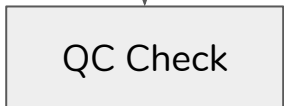
# Data and Background

- Problem:
  - Antibiotic resistance in *Klebsiella*
- Data:
  - 260 isolates of *Klebsiella* spp.
  - 2 x 250 Illumina short reads (MiSeq platform)
- Background:
  - Generic biological characteristics
    - 1 chromosome, likely some plasmids
    - Genome size: ~ 5.3-5.5 Mbp
    - GC content: ~57.1 % GC
- Objective:

Generate a pipeline to assemble and QC short read data, with respect to biological characteristics and downstream analyses.



<https://sciencesource.com/Doc/SCS/Media/TR7/f/3/6/f/SS2294165.jpg?d63641835809>



Yes

The 'Yes' path label, indicating successful completion.



No

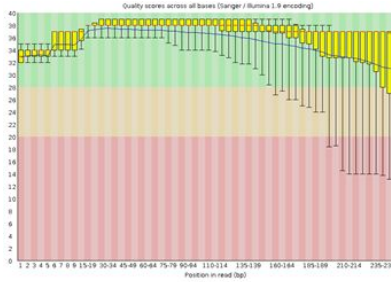
The 'No' path label, indicating a need for re-checking.



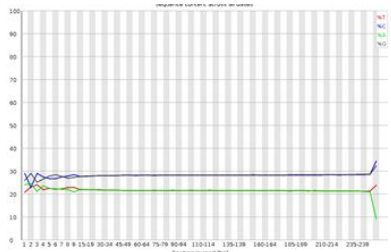
- Remove sequencing bias
- Remove bases with low Q-score
- Remove sequencing adapters
- Filter out short reads

# Raw, unloved sequence data

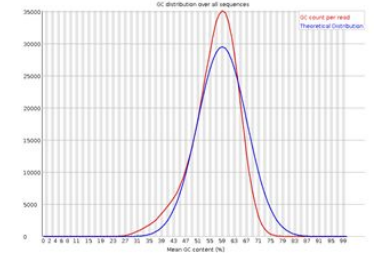
## Forward Reads



Avg. Phred  
Score

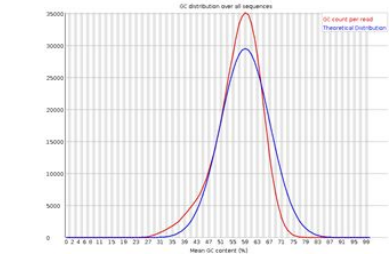
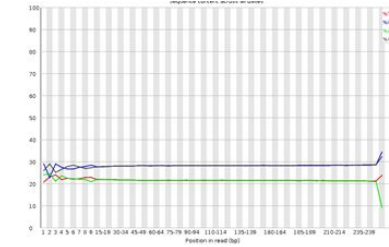
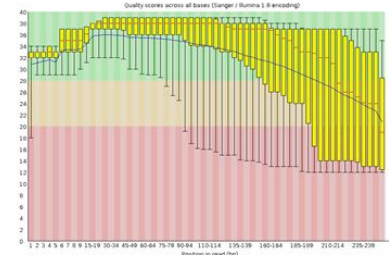


Avg Nucleotide %



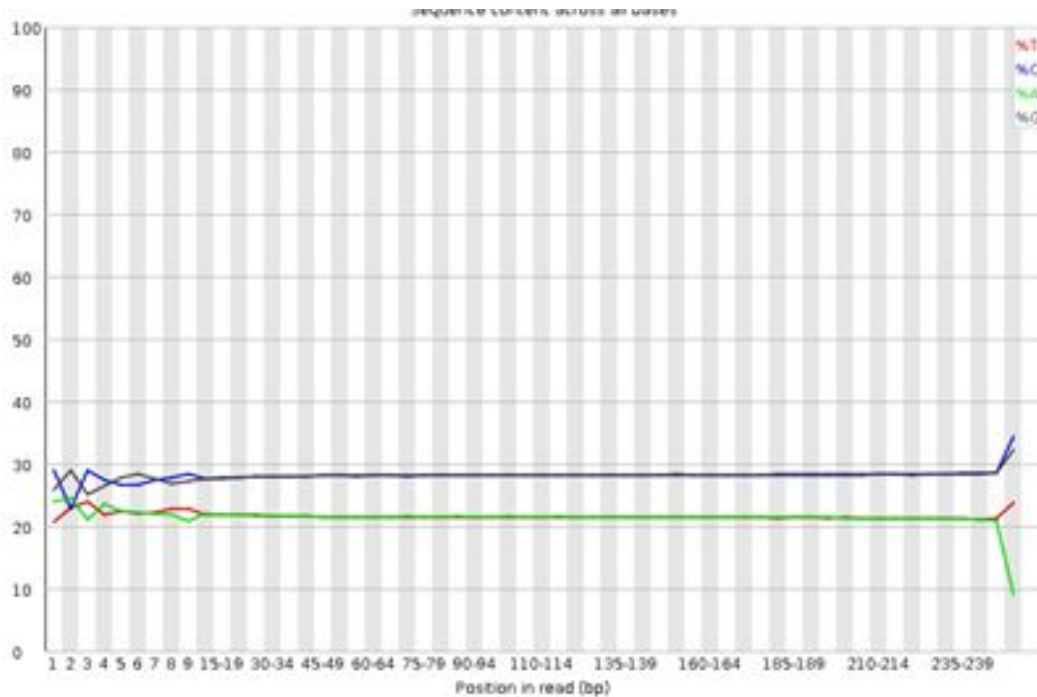
GC content

## Reverse Reads



Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

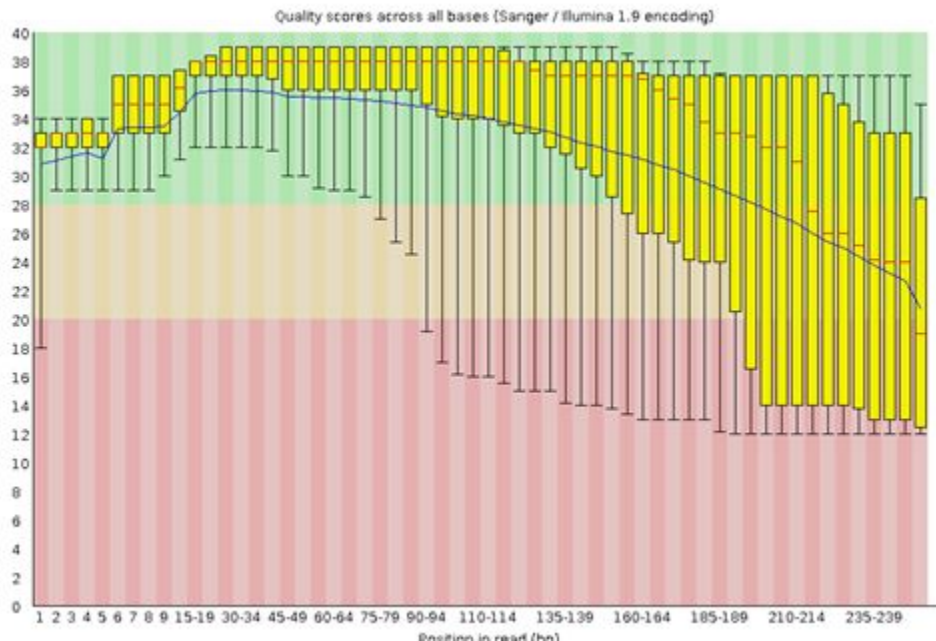
# Trimming



## Sequencing Bias

- Introduced during library preparation
- Non-random favoring of primer sequences during amplification
- To be removed or not?

# Trimming Tools



## Quality scores:

– Phred Q-score =  $-10\log_{10}P$

\* At 1x depth coverage:

Q-score	Incorrect Base Call (Probability)
10	1 in 10 (90%)
20	1 in 100 (99%)
30	1 in 1000 (99.9%)
40	1 in 10,000 (99.99%)

# Trimming Software:

BBDuk

Trimmomatic

SolexaQA++

Sickle

Seqtk

TrimGalore



# Trimming Software

BBDuk

Trimmomatic

SolexaQA++

Sickle

Seqtk

TrimGalore

# What to look for?

Average Phred Score

Per Base Sequence Content

Adapter Content

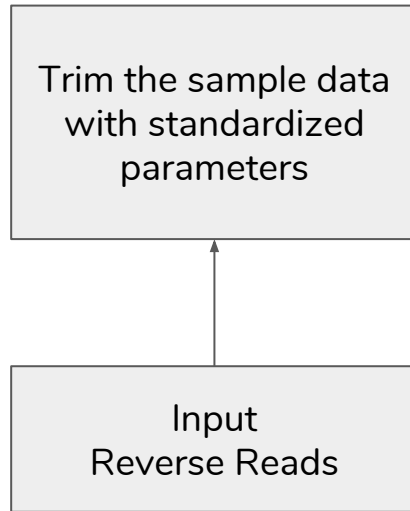
Sequence Length Distribution

Which trimming tool performs the best?

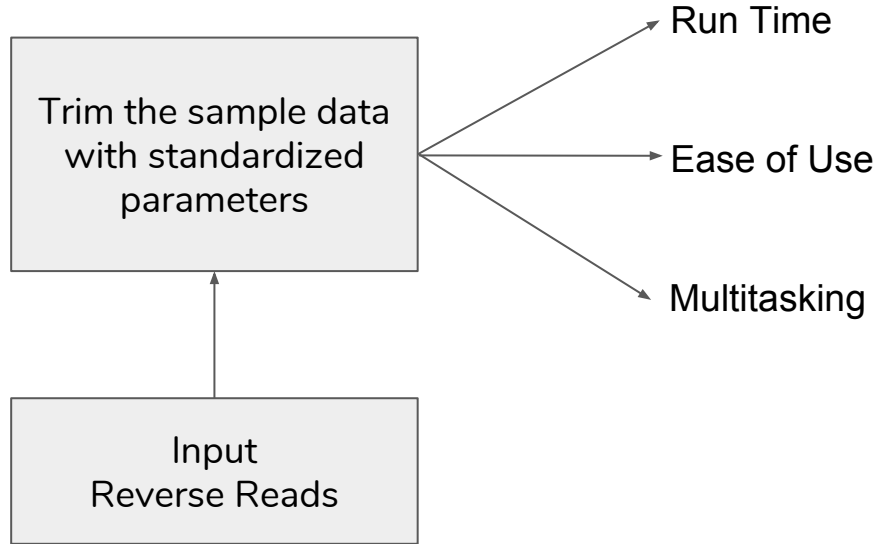
# Which trimming tool performs the best?

Trim the sample data  
with standardized  
parameters

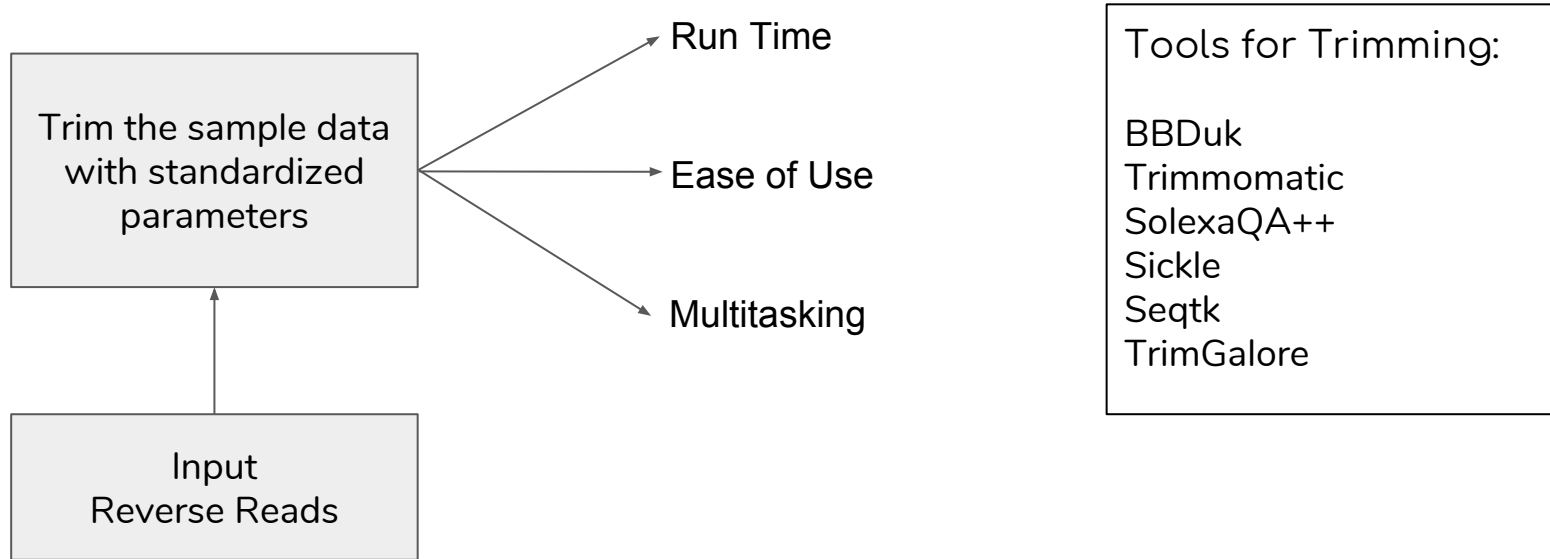
# Which trimming tool performs the best?



# Which trimming tool performs the best?



# Which trimming tool performs the best?



Parameter:

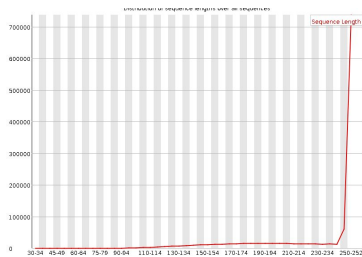
- Quality Trimming - Q20

# Which tool performs the best?

Software	Run time (in seconds)	Output File (MB)	Multitask capacity
SolexaQA++	80.24	393.6	no
Sickle	5.19	508.3	no
TrimGalore	17.77	486.4	yes
BBDuk	3.17	428.7	yes
Trimmomatic	3.09	383.9*	yes
Seqtk	3.41	433.5	yes

# Which tool performs the best?

Raw Reads

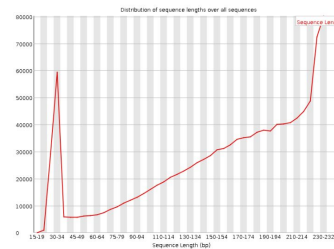
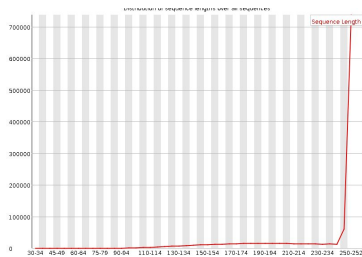


Software	Run time (in seconds)	Output File (MB)	Multitask capacity
SolexaQA++	80.24	393.6	no
Sickle	5.19	508.3	no
TrimGalore	17.77	486.4	yes
BBDuk	3.17	428.7	yes
Trimmomatic	3.09	383.9*	yes
Seqtk	3.41	433.5	yes



# Which tool performs the best?

Raw Reads

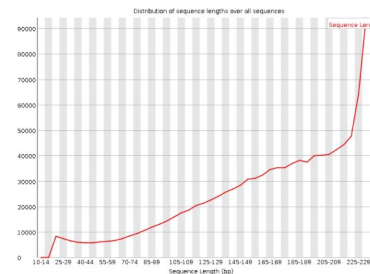


SeqTK

Software	Run time (in seconds)	Output File (MB)	Multitask capacity
SolexaQA++	80.24	393.6	no
Sickle	5.19	508.3	no
TrimGalore	17.77	486.4	yes
BBDuk	3.17	428.7	yes
Trimmomatic	3.09	383.9*	yes
Seqtk	3.41	433.5	yes



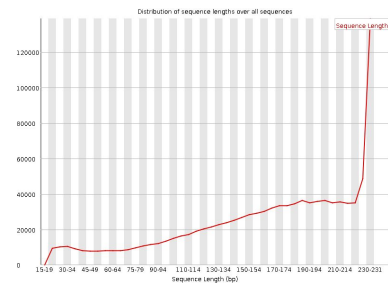
SolexaQA++



BBDuk

# Which tool performs the best?

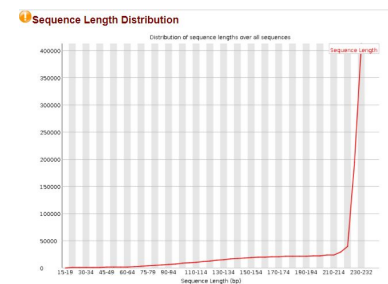
Software	Run time (s)	Output File (MB)	Multitask capacity
SolexaQA++	80.24	393.6	no
Sickle	5.19	508.3	no
TrimGalore	17.77	589.9	yes
BBDuk	3.17	428.7	yes
Trimmomatic	3.09	383.9*	yes
Seqtk	3.41	433.5	yes



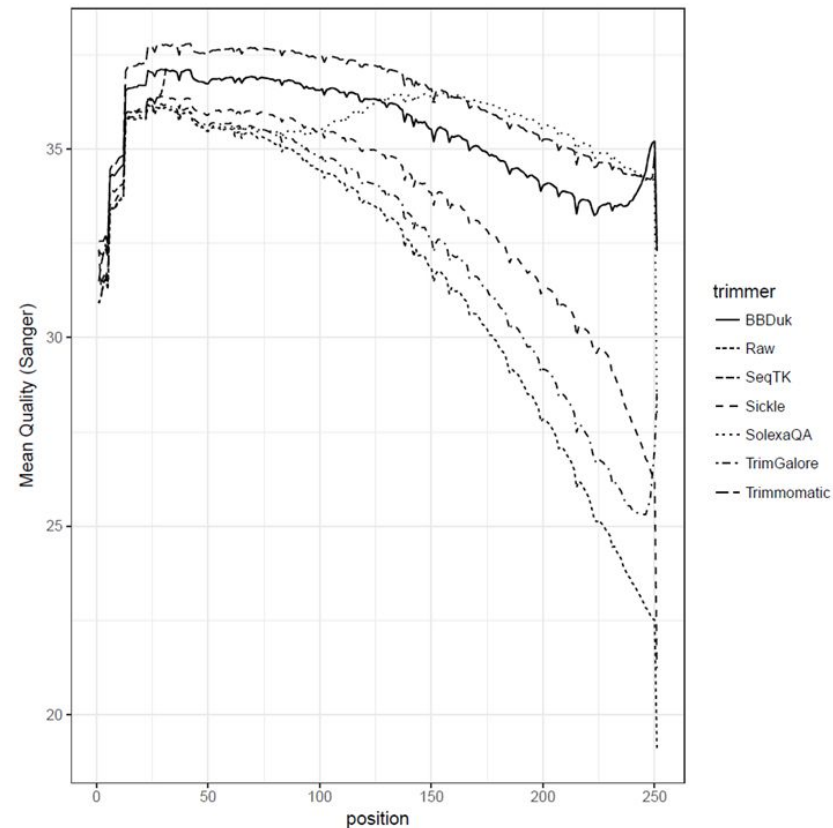
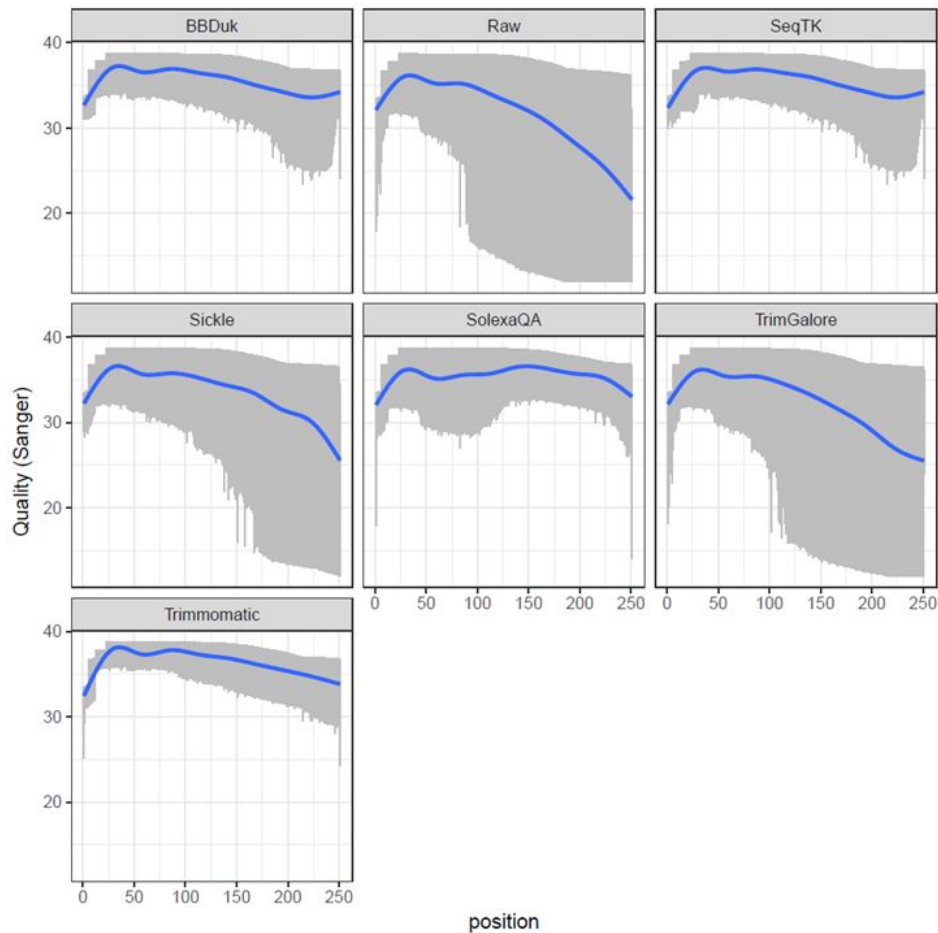
Trimmomatic



Sickle



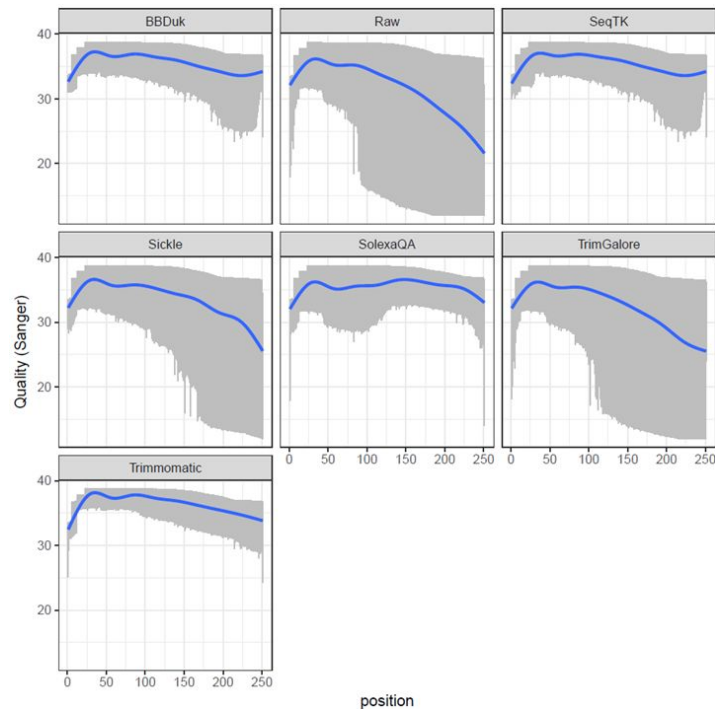
TrimGalore



Graphs generated using QRQC( R)  
 Vince Buffalo (2012). qrqc: Quick Read Quality Control. R package version 1.32.0.  
<http://github.com/vsbuffalo/qrc>

# Which tool performs the best?

Software	Run time (in seconds)	Output File (MB)	Multitask capacity
SolexaQA++	80.24	393.6	no
Sickle	5.19	508.3 ✓	no
TrimGalore	17.77	589.9	yes
BBDuk	3.17	428.7	yes
Trimmomatic	3.09 ✓	383.9*	yes ✓
Seqtk	3.41	433.5	yes



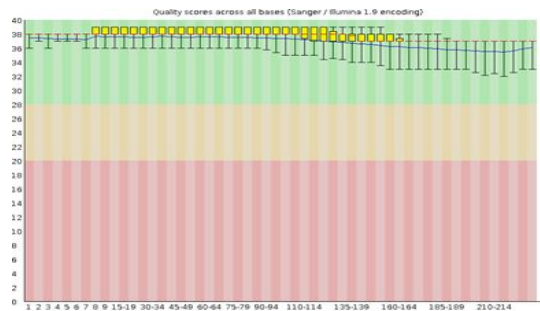
**Winner:**  
**Trimmomatic**

\* Trimmomatic uses a sliding window trimming algorithm – was set to 1 here in order to be most comparable.

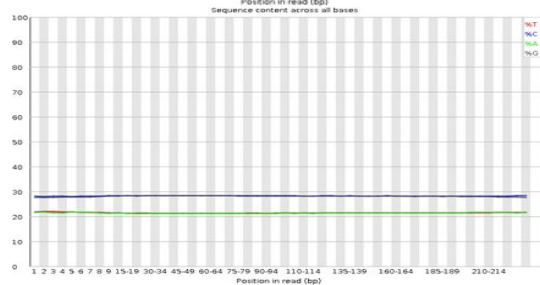
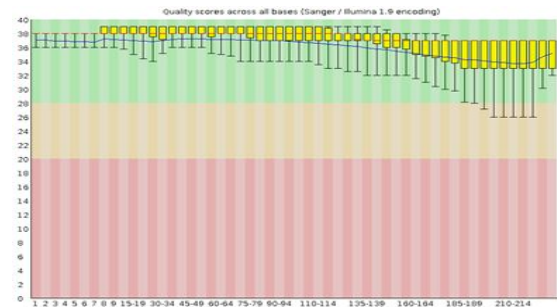
# Clean Data, courtesy of Trimmomatic

**Forward Reads**

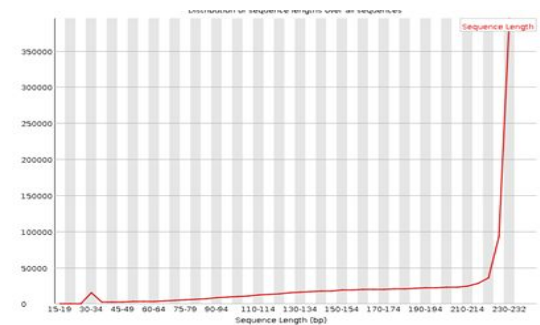
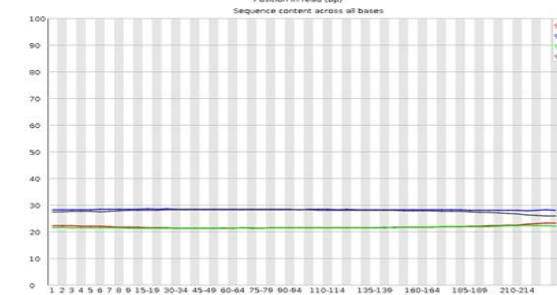
**Reverse Reads**



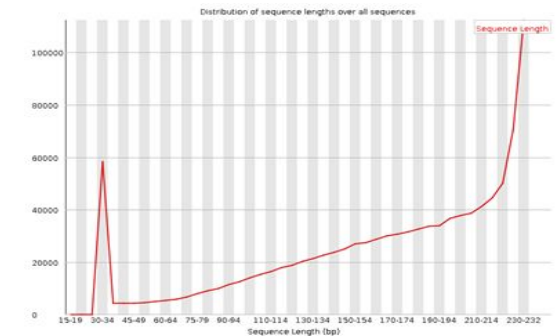
Avg. Phred Q-scores



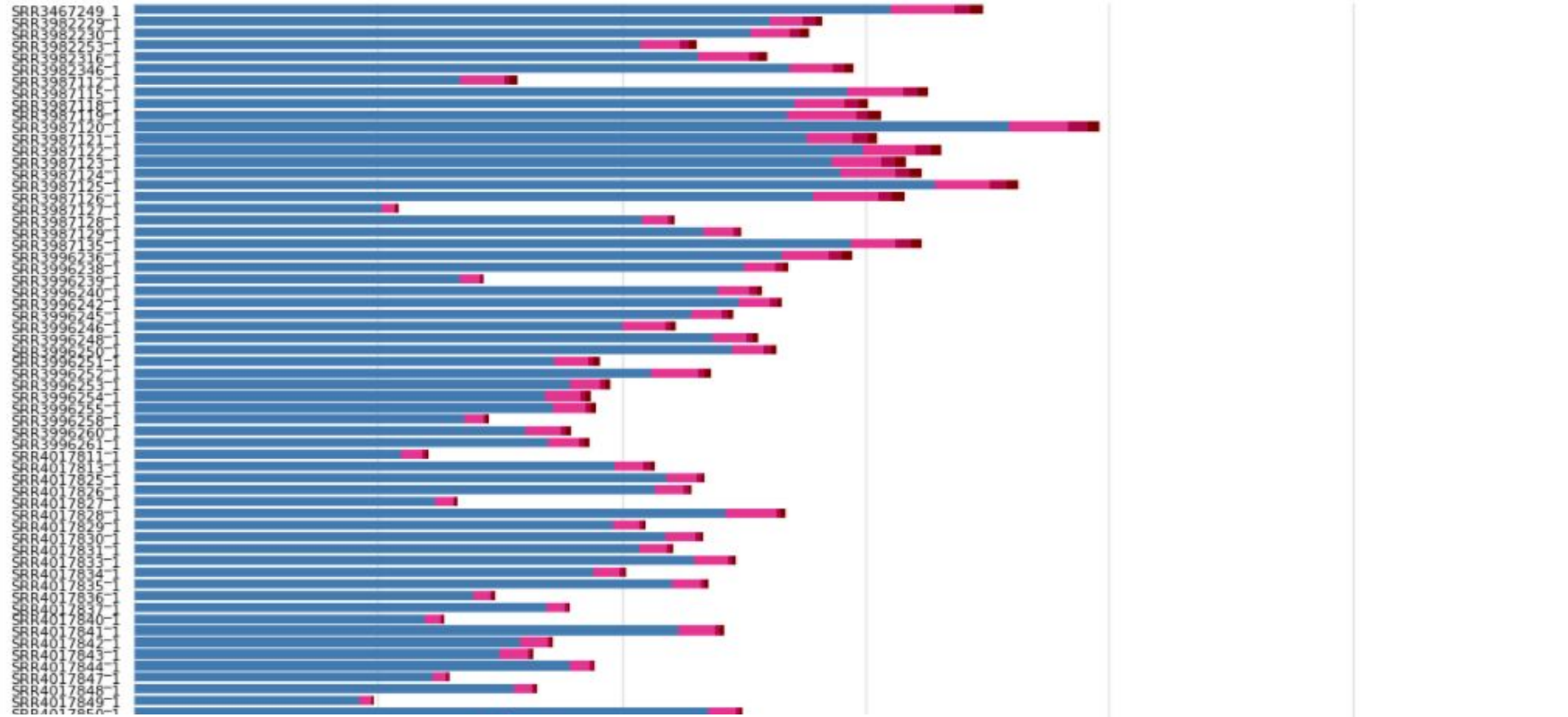
Avg nucleotide %



Length Distribution



# Trimmomatic

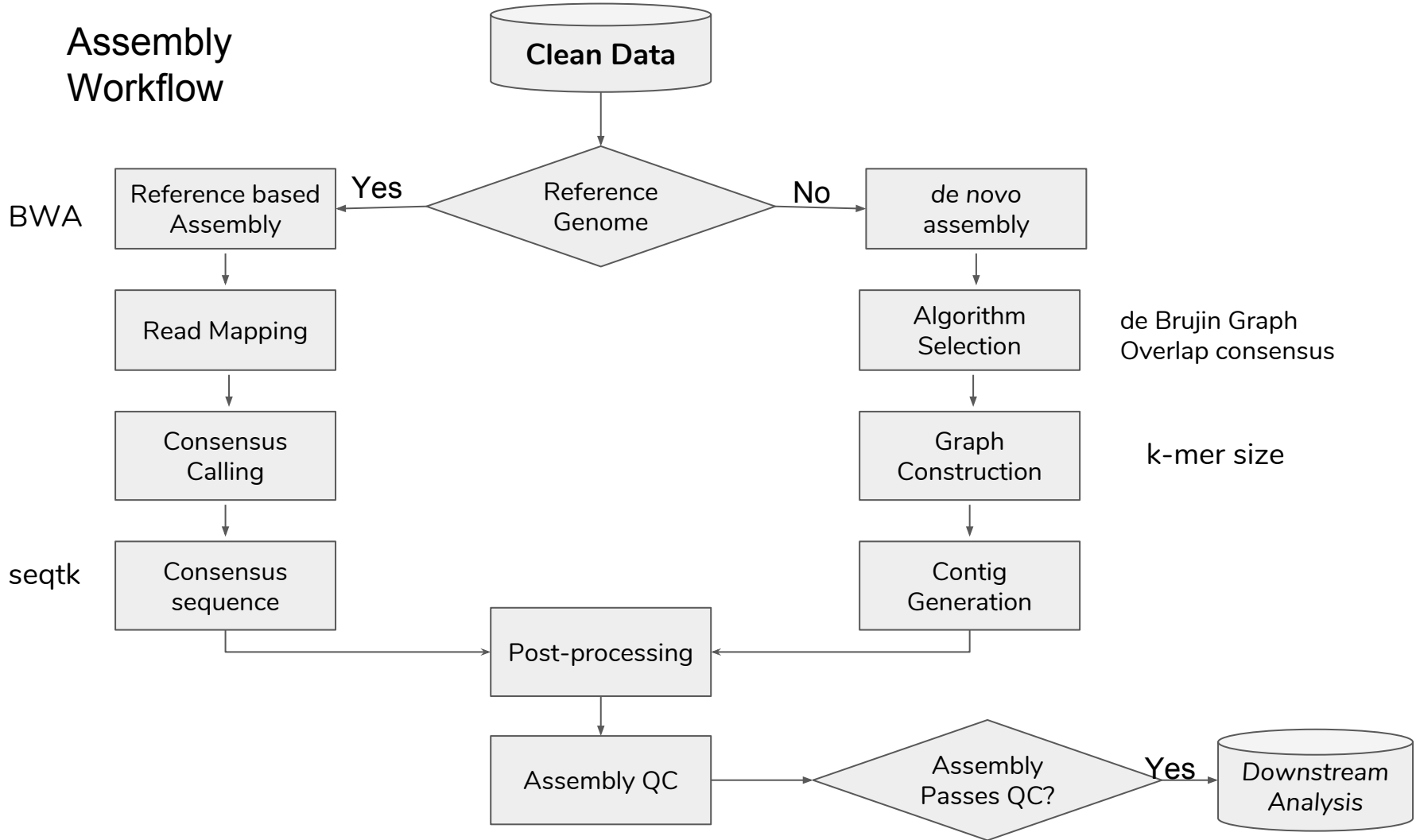


MultiQC: Summarize analysis results for multiple tools and samples in a single report

*Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller*

Bioinformatics (2016)

# Assembly Workflow



# Biological Considerations

Bacterial genomes are  
single, circular  
chromosomes



Ideal End Goal:  
An assembly containing only 1  
contig  
(extra credit if it's circular)

How do  
we tell?

Which species  
of *Klebsiella*  
do we have?

What is a  
good reference  
genome to use?

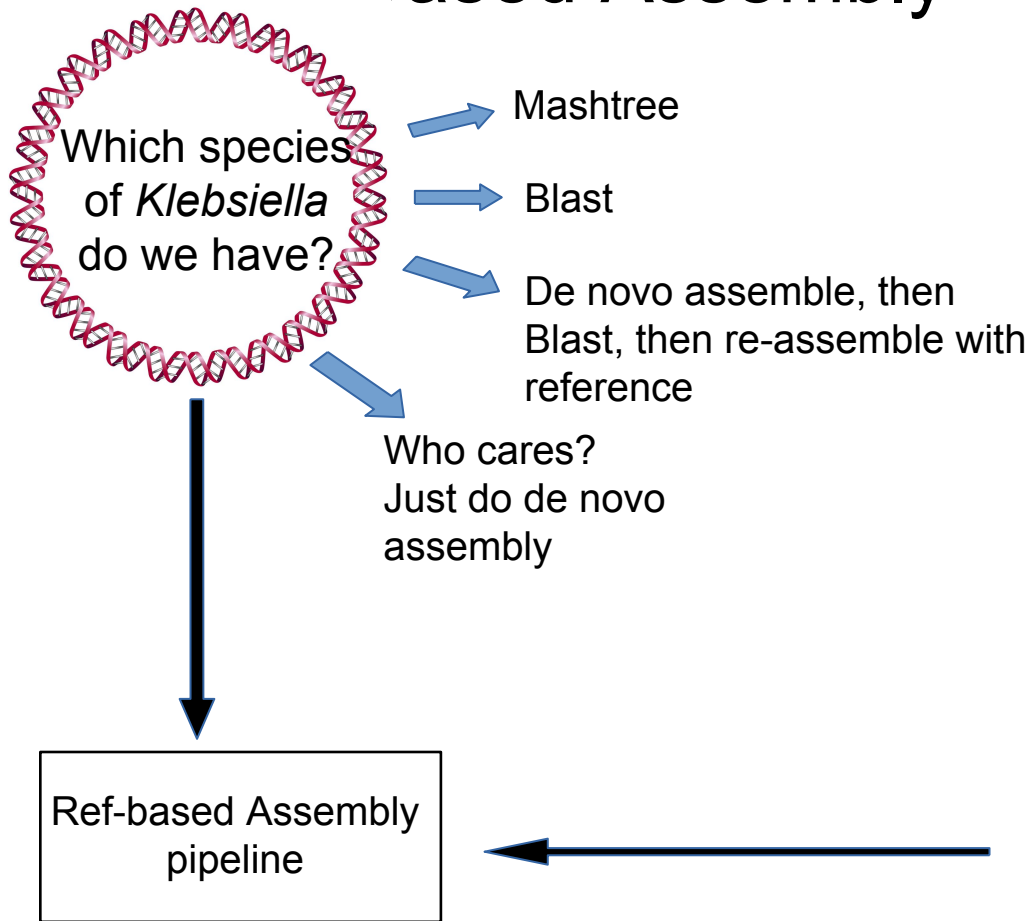
When/how to  
check for  
contamination in  
our sequences?

What is  
contamination  
and how did it  
get there?

How to detect  
and remove it?



# Reference Based Assembly



What is a good reference genome to use?

Ask NCBI!

NCBI Resources How To

Assembly Assembly Advanced Browse by organism

Full Report

**ASM988v1**

Organism name: [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044 \(enterobacteria\)](#)

Intraspecific name: Strain: NTUH-K2044

BioSample: [SAM00060934](#)

Submitter: National Health Research Institutes

Date: 2005/01/05

**Assembly level: Complete Genome**

**Genome representation: full**

GenBank assembly accession: GCA\_000009885.1 (latest)

RefSeq assembly accession: GCF\_000009885.1 (latest)

RefSeq assembly and GenBank assembly identical: yes

IDs: 31388 [UID] 10688 [GenBank] 31388 [RefSeq]

**History** (Show revision history)

**Comment**

This sequence was determined by the K. pneumoniae Genome Project at the Yang-Ming University VYM G study were from NRPGM of R.O.C.

**Global statistics**

Total sequence length	5,472,672
Total assembly gap length	0
Total number of chromosomes and plasmids	2

The right side of the image features a cartoon of a person at a computer asking 'What is a good reference genome to use?' and 'I'M ON THE INTERNET!' with 'TIKA TIKA TIKA' written on the screen. Below this is a screenshot of the NCBI website showing the assembly page for ASM988v1. Red boxes highlight the 'Assembly level: Complete Genome' and 'Genome representation: full' fields, and another red box highlights the 'Global statistics' table, which shows a total sequence length of 5,472,672, zero assembly gaps, and two chromosomes/plasmids.

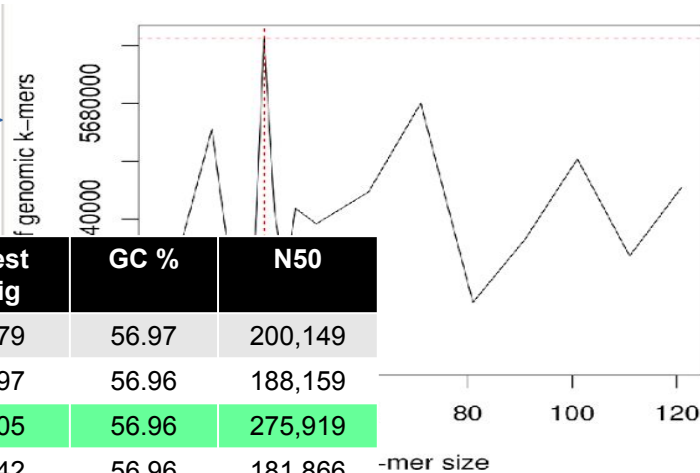
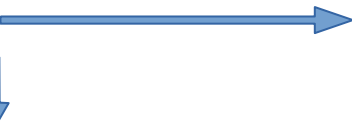
# De novo assembly: Kmer selection

- De Bruijn graph-based assemblers split reads into kmers for graph construction.
- Assembly outcome is heavily influenced by the choice of kmer values.
  - Problematic palindromes
  - Sweet spot between sensitivity and specificity
- **Where is this magical sweet spot?**
  - Short answer: it's different for every sample you assemble due to quality of seq data, genome complexity, etc. Garbage in, garbage out.

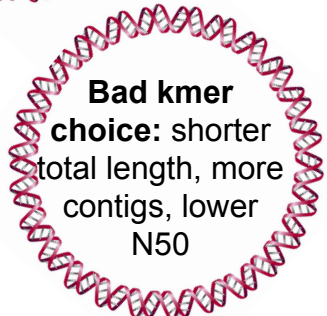
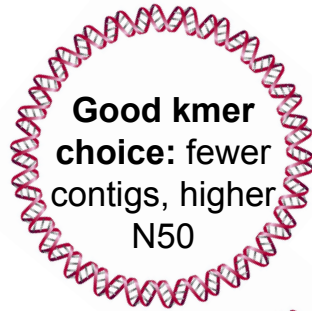
## Tools for kmer estimation:

- Kmergenie

- Spades



kmer	Total length	# contigs	Largest contig	GC %	N50
41	5,743,700	82	469779	56.97	200,149
77	5,748,634	76	469897	56.96	188,159
99	5,745,519	68	471705	56.96	275,919
127	5,751,083	76	409842	56.96	181,866



# Preliminary results: comparison of assemblies

## Using k=41, determined by kmergenie

Assembler	Run Time (s)	Kmer	# contigs	N50 (kbp)	Total length (Mbp)	GC %	# N's
Spades	403	41	82	200.1	5.74	56.97	370
Skesa	87	41	111	120.4	5.67	56.98	0
IDBA-UD	56	41	192	66.5	5.75	56.98	0
Tadpole	13.3	41	343	56.0	5.70	56.96	0
IDBA-Hybrid	81	41	190	62.5	5.63	56.98	0
Ref-based (Samtools)	395	--	2	5,248.5	5.47	58.08	521,755

### Caveats to comparison:

- all assemblers compared here support multi-threading. This parameter left as default.
- Only Spades and Tadpole allow for add'n single end read input (not used here)
- Skesa does not include a built-in scaffolder.

### Which assembly is best?

Overall, Spades and Skesa are pretty comparable with this kmer value.

# Preliminary results: comparison of assemblies

## Using k=99, determined by us using Spades

Assembler	Run Time (s)	Kmer	# contigs	N50 (kbp)	Total length (Mbp)	GC %	# N's
Spades	421	99	68	275.9	5.745	56.96	170
Skesa	89.5	99	195	60.7	5.668	56.96	0
IDBA-UD	50.8	99	121	119.7	5.748	56.96	0
Tadpole	15.3	99	244	44.5	5.745	56.97	0
IDBA-Hybrid	82.8	99	161	83.5	5.751	56.96	0
Ref-based (Samtools)	395	--	2	5,248.5	5.47	58.08	521,755

### Caveats to comparison:

- all assemblers compared here support multi-threading. This parameter left as default.
- Only Spades and Tadpole allow for add'n single end read input (not used here)
- Skesa does not include a built-in scaffolder.

### Which assembly is best?

Spades (k=99) has lowest # contigs, highest N50 of all de novo assemblies attempted.

Drawback: also takes the longest to run, has some N's

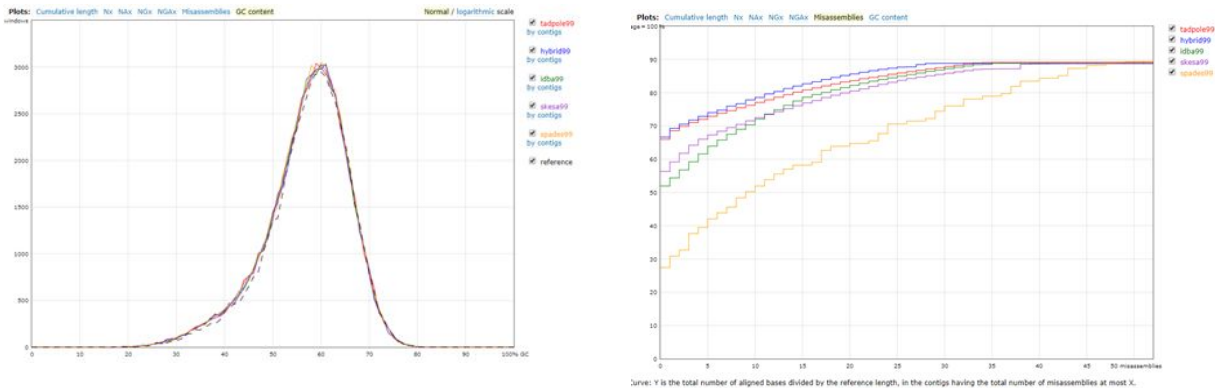
# Preliminary results: Draft QC (Quast)

For draft assemblies where k=99

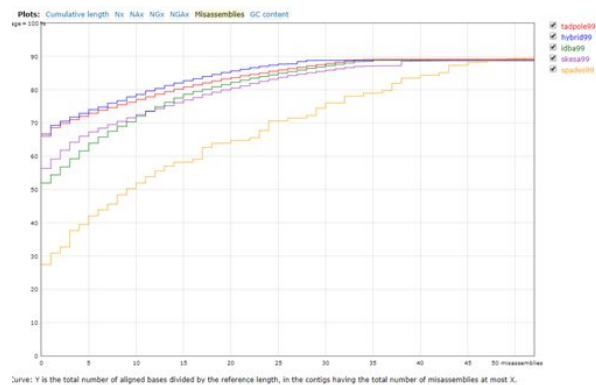
## Genome statistics heatmap (m=500)



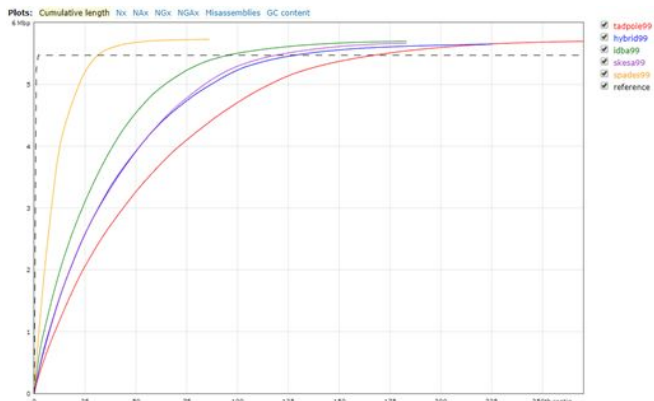
## GC content



## Misassemblies compared to reference



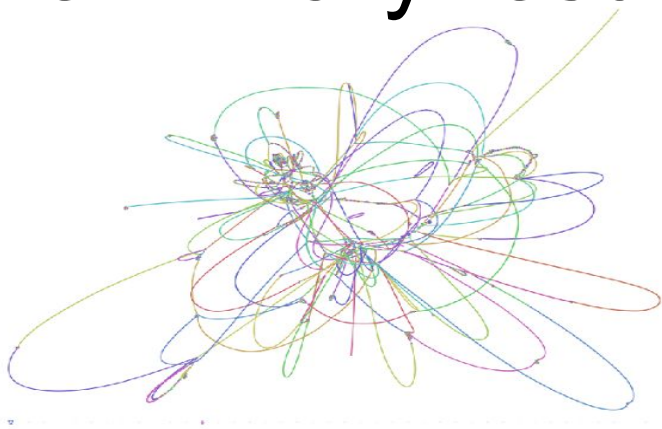
## Cumulative length of draft assemblies



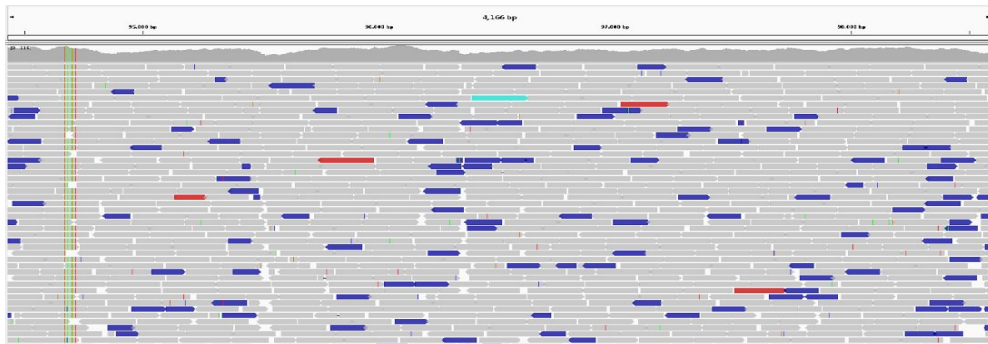
## Indications of contamination

- GC content
- Many misassemblies compared to reference genome\*
- Depth coverage anomalies
- Highly fragmented assemblies

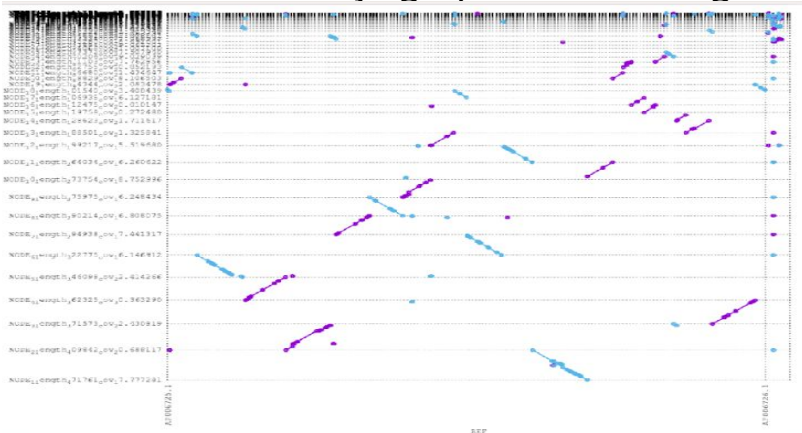
# Preliminary results: Visualization



Visualize the de Bruijn graph with Bandage



Visually inspect alignment quality with IGV

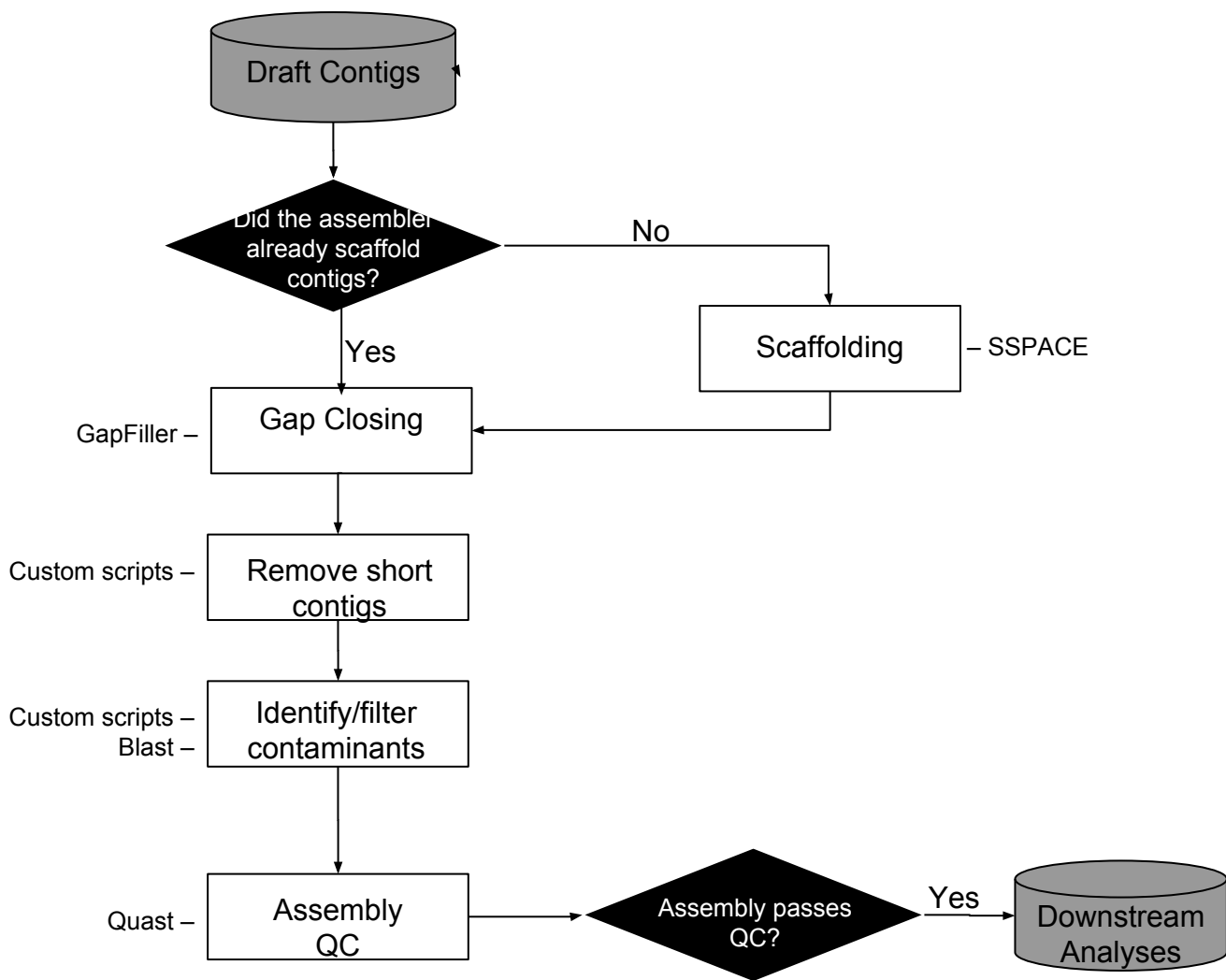


- Check for circular (ie. closed) chromosome



Identify indels, duplications, reversals, etc. using Mummer

# Post-Assembly Finishing



# Questions?

## Additional References

- Bankevich A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 2012
- Peng, Y., et al. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, *Bioinformatics*, 28, 1420-1428.
- Cox, M.P., D.A. Peterson, and P.J. Biggs. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
- <https://github.com/lh3/seqtk>
- <https://github.com/FelixKrueger/TrimGalore>
- Li H, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009 Aug 15;25(16):2078-9
- Chikhi R., Medvedev P. Informed and Automated k-Mer Size Selection for Genome Assembly, HiTSeq 2013
- Kurtz S, et al. Versatile and open software for comparing large genomes. *Genome Biology* (2004), 5:R12.
- James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. *Nature Biotechnology* 29, 24–26 (2011)
- Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015). Bandage: interactive visualisation of de novogenome assemblies. *Bioinformatics*, 31(20), 3350-3352.
- <https://jgi.doe.gov/data-and-tools/bbtools/>
- Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

## Special Thanks:

- Dr. David Weiss (Emory)
- Richa Agarwala (NCBI)
- Team 1 Genome Assembly Group

Look for a homework  
announcement shortly!