METAGENOMIC ASSEMBLY USING METASPADES

DEVAN BULSARA AND ADITYA SHAH, ECES-650, SPRING 2021, DREXEL UNIVERSITY

REVIEW OF METAGENOMICS

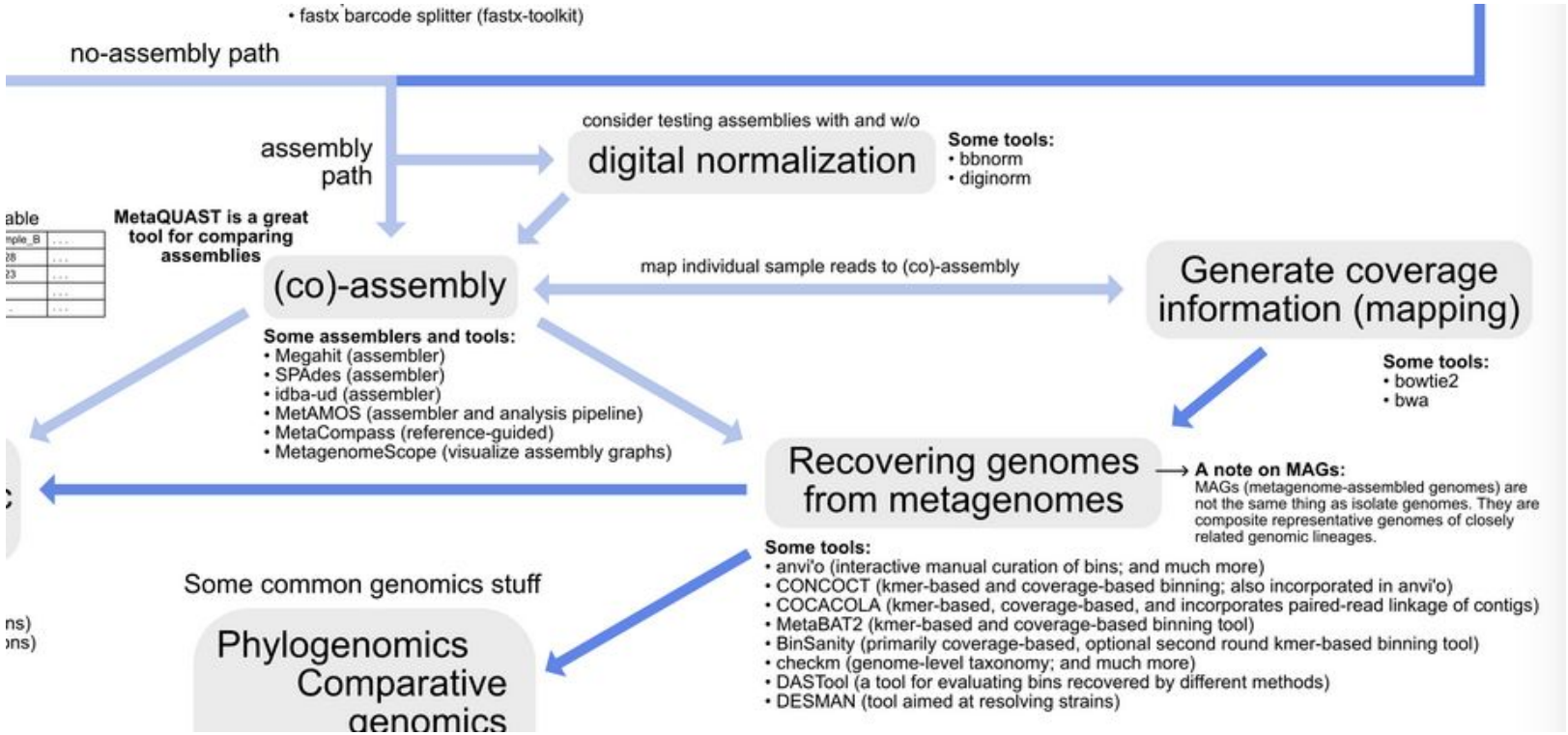
- Metagenomics:
  - Study of genetic materials from a mixed community of organisms.
  - Analyzing genomes of microorganisms present in environments where it may not be possible differentiate between organisms.
  - May not possible to obtain cultured samples in order to study the community of microorganisms present and find out how they are related.
- History:
  - 1985 – Norman R. Pace and colleagues used direct analysis of 5S and 16S rRNA sequences to described the diversity in an environment without culturing ,i.e. the first report of isolating and cloning DNA from environmental sample. [1]
  - 1988 - The term Metagenomics first appeared in a publication authored by Jo Handelsman's group describing the examination of uncultured microorganisms based on functional analysis in a mixed environment. [1]

REVIEW OF METAGENOMICS CONT'D

- Functional Analysis:
  - "…integrates molecular biology and cell biology studies, and deals with the whole structure, function and regulation of a gene in contrast to the gene-by-gene approach of classical molecular biology technique…This involves comprehensive analysis to understand genes, their functional roles and variable levels of protein expression." [2]
- Problem:
  - Technology to sequence the whole genome at the same time.
  - Sequencing generally provides ~150bp but many encoded genes are generally in the order of ~1000bp [3]
- Solution:
  - Get read sequences, then assemble those into contigs, scaffolds, and finally the gene.

ASSEMBLY

- Computational step to reconstruct the genome from its reads, after sequencing.
- Stiches together individual DNA sequences into genes or organisms



Overview of generic* metagenomic workflow [4]

TOOLS FOR METAGENOMIC ASSEMBLY

- Referred to as Assemblers

- Different assemblers are optimized for different purposes [5]
  - Spades – Assemble small genomes such as from a bacteria.
  - Unicycler – Works as spade optimizer for short reads, long read (3rd gen: PacBio, Nanopore, etc.), create hybrid assembly
  - Plasmidspades – Plasmid data or whole genome sequence that includes plasmids.
  - Metaspades – Metagenomics assembly based on the spades toolkit.
  - Metavelvet – Based on velvet toolkit for short read metagenomics assembly.
  - Other tools: SOAPdenovo2, Omgea, etc.

- In this tutorial we are working with MetaSpades.

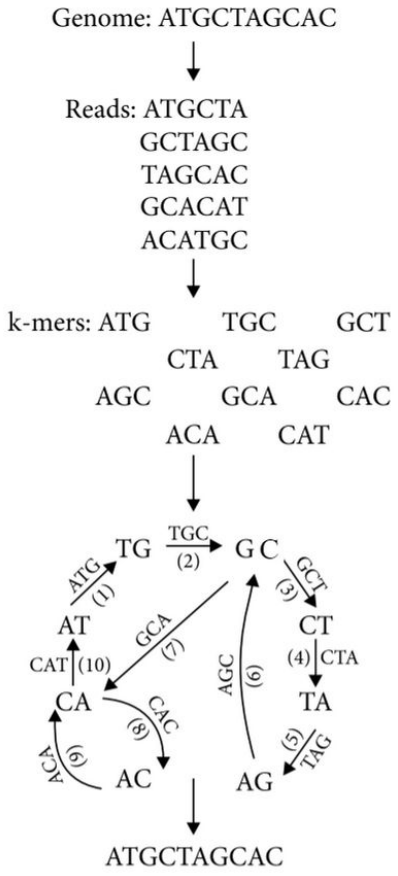
WHY METASPADES?

- **Metagenomics issues:**
  - Difference in the abundance levels of various species.
  - Sharing of conserved genomic regions in a microbial community.
  - Multiple related strains with varying abundance.

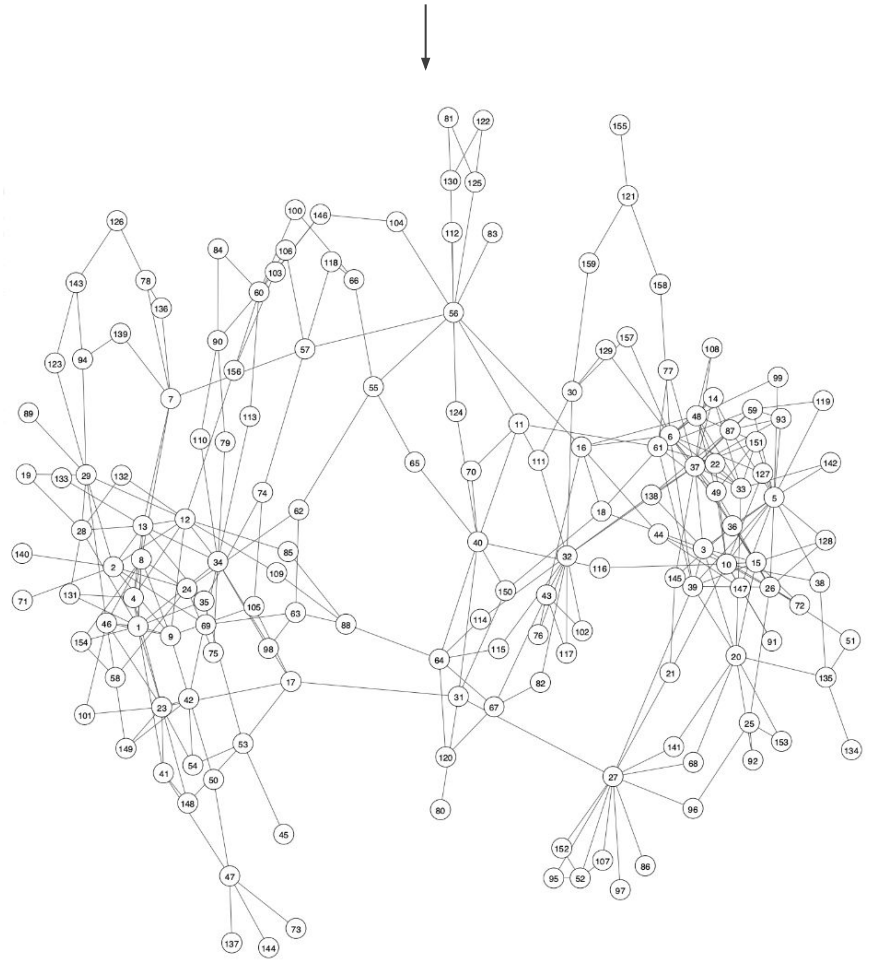
- Spades [6]
  - Works well for assembling low-complexity metagenomes but its performance deteriorates in the case of complex bacterial communities.

METASPADES PIPELINE

- metaSPAdes Pipeline [6]
  - de Bruijn graph using SPAdes
  - Assembly graph through simplification procedures
  - Reconstruct paths for long genomic fragments.



de Bruijn Graph [8]



Assembly graph [10]

INTERACTING WITH METASPADES

- Input and Output Files:
  - Input: Fasta sequence read and corresponding quality scores, i.e. Fastq file.
  - Output: Fasta file with the assembled contigs.
- K-mers [7]
  - Increased connectivity, so more ambiguity and less clear "paths" through the graph.
  - Less connectivity but higher specificity, so can lead to breaking down.
- Read types: single-read and paired-end sequencing.
  - Single-read: useful for some applications (small sequences), fast and economical.
  - Paired-end: twice the number of sequencing reads, enables more accurate read alignment.
- Coverage Cutoff [7]
  - Low – more sequencing errors and misconnections
  - High - mis-assemblies in the contigs and destroys lots of useful data.
- PHRED quality score:
  - Dictates the probability that the decision is correct.

INTERACTING WITH METASPADES:

```
module load shared
module load spades/3.15.2
spades.py --meta --12 RL_S001__insert_270.fq -o Output2 -t $SLURM_CPUS_PER_TASK
```

metaSPAdes Pipeline Input

```
[db3265@picotte001 Output2]$ ls
assembly_graph_after_simplification.gfa  contigs.paths
assembly_graph.fastg                     corrected
assembly_graph_with_scaffolds.gfa        dataset.info
before_rr.fasta                          first_pe_contigs.fas
contigs.fasta                            input_dataset.yaml
[db3265@picotte001 Output2]$
```

metaSPAdes Pipeline Output

INTRODUCTION TO QUAST

- Quality Assessment Tools for Genomic Assemblies (QUAST), a tool to evaluate and compare genome assemblies.
- Input multiple assemblies for comparison.
- Multi-threaded operation
- Provide basics statistics on assembly data (i.e. information regarding mis-assemblies, unaligned contigs, genomic statistics, and alignment statistics)

```
module load shared
module load python/gcc
python3 ./quast/quast.py \
                        ./Output2/contigs.fasta \
                        -o ./QuastResults \
                        -t $SLURM_CPUS_PER_TASK
```

QUAST Pipeline Input

QUAST Results Definitions [9]

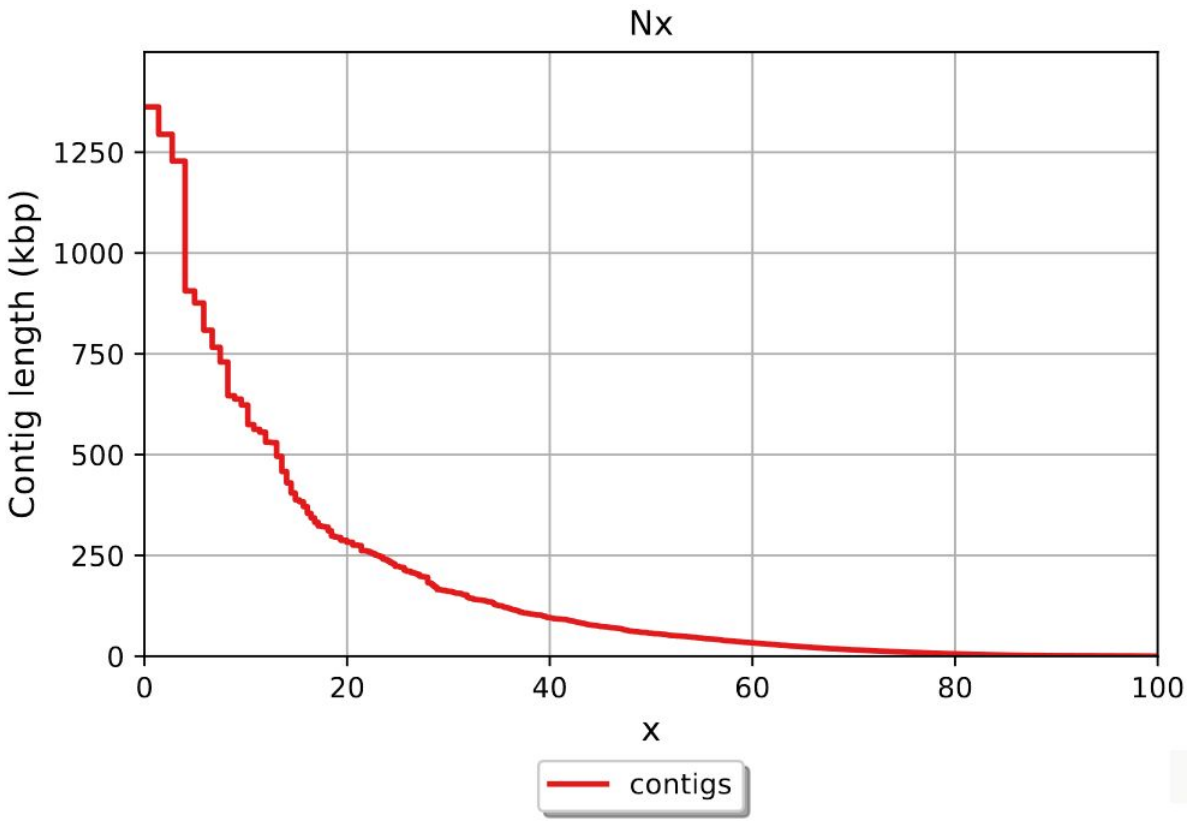
- GC (%): Total number of Guanine and Cytosine nucleotides in the assembly divided by the total length of the assembly
- N50: Length for which the collection of all contigs of that length or longer covers at least half the assembly
  - 50th percentile of base pairs
- N90: Length for which the collection of all contigs of that length of longer covers at least 90% of the assembly
  - 90th percentile of base pairs
- L50: number of contigs equal to or longer than N50
  - minimum number of contigs to cover half the assembly
- L90: number of contigs equal to or longer then N90
  - minimum number of contigs to cover 90% of the assembly
- N's per 100 kbp: average number of uncalled bases per 100,000 assembly bases

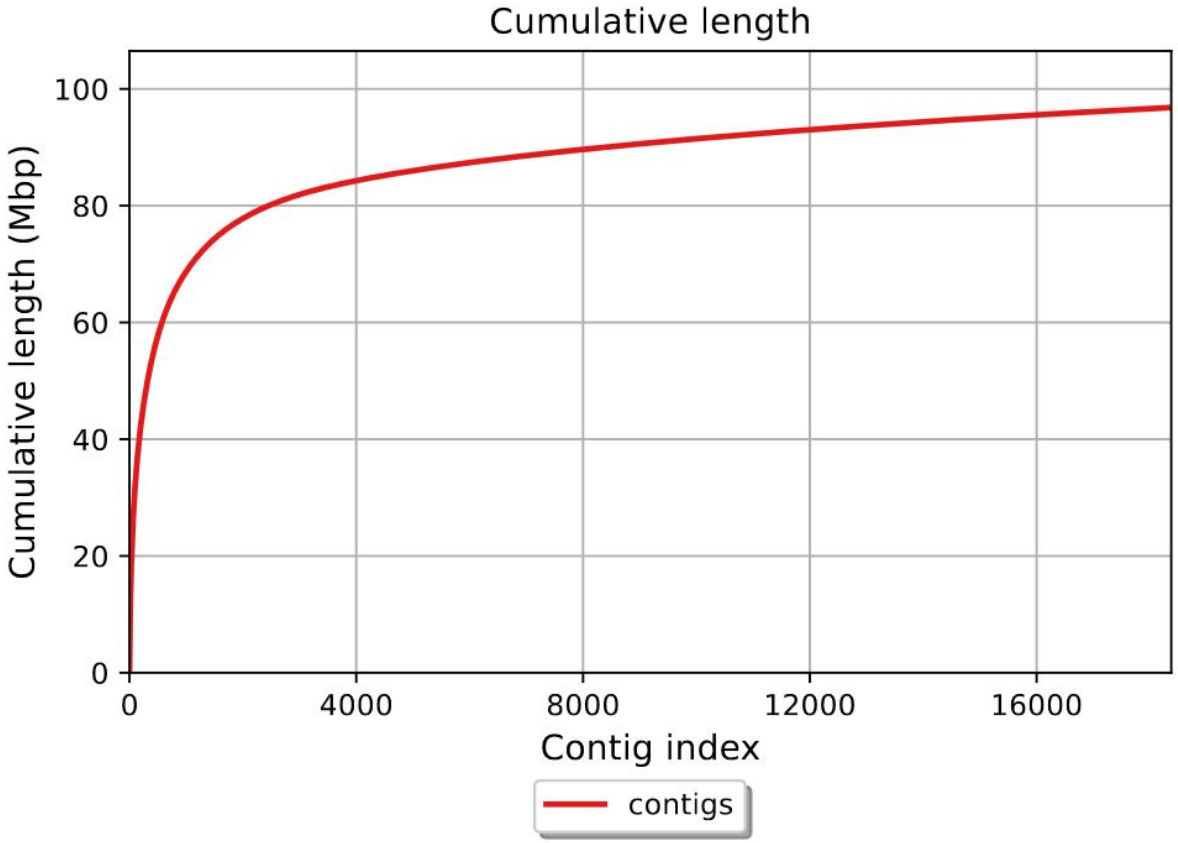
RESULTS

	contigs
# contigs (>= 0 bp)	47915
# contigs (>= 1000 bp)	8043
# contigs (>= 5000 bp)	2173
# contigs (>= 10000 bp)	1352
# contigs (>= 25000 bp)	650
# contigs (>= 50000 bp)	342
Total length (>= 0 bp)	105684805
Total length (>= 1000 bp)	89684411
Total length (>= 5000 bp)	78727664
Total length (>= 10000 bp)	72972882
Total length (>= 25000 bp)	62116476

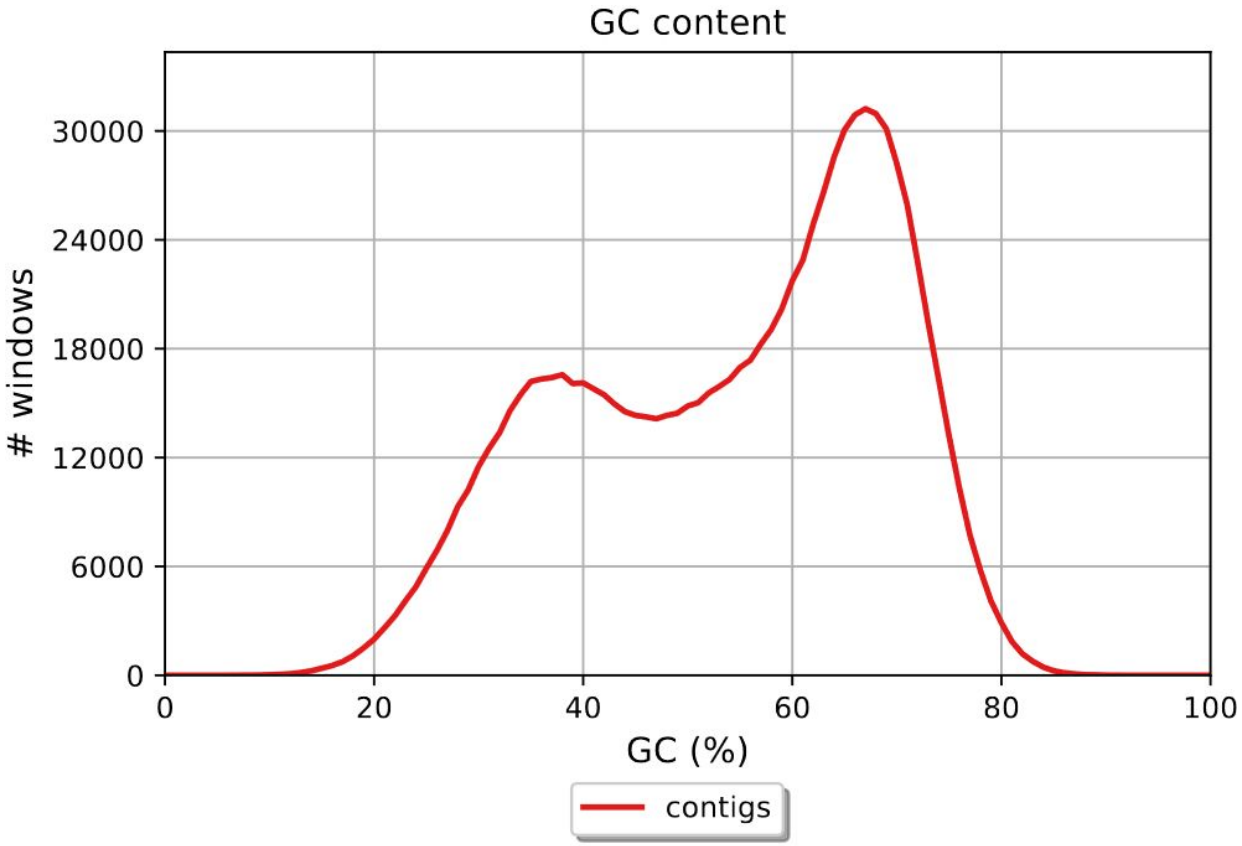
Total length (>= 50000 bp)	51213430
# contigs	18375
Largest contig	1362261
Total length	96821116
GC (%)	54.20
N50	56980
N90	1321
L50	290
L90	5808
# N's per 100 kbp	0.00

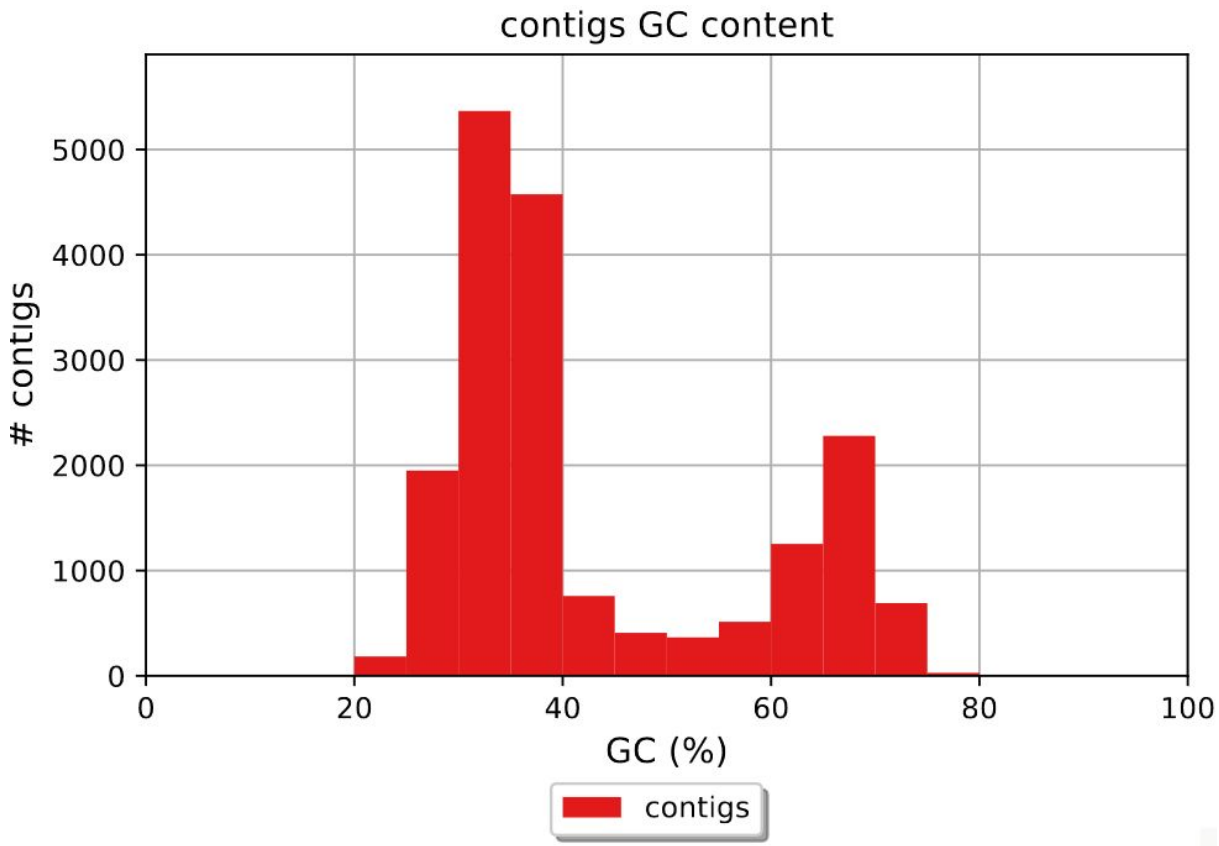
RESULTS


Nx
Contig length (kbp)
1250
1000
750
500
250
0
0
20
40
60
80
100
x
contigs

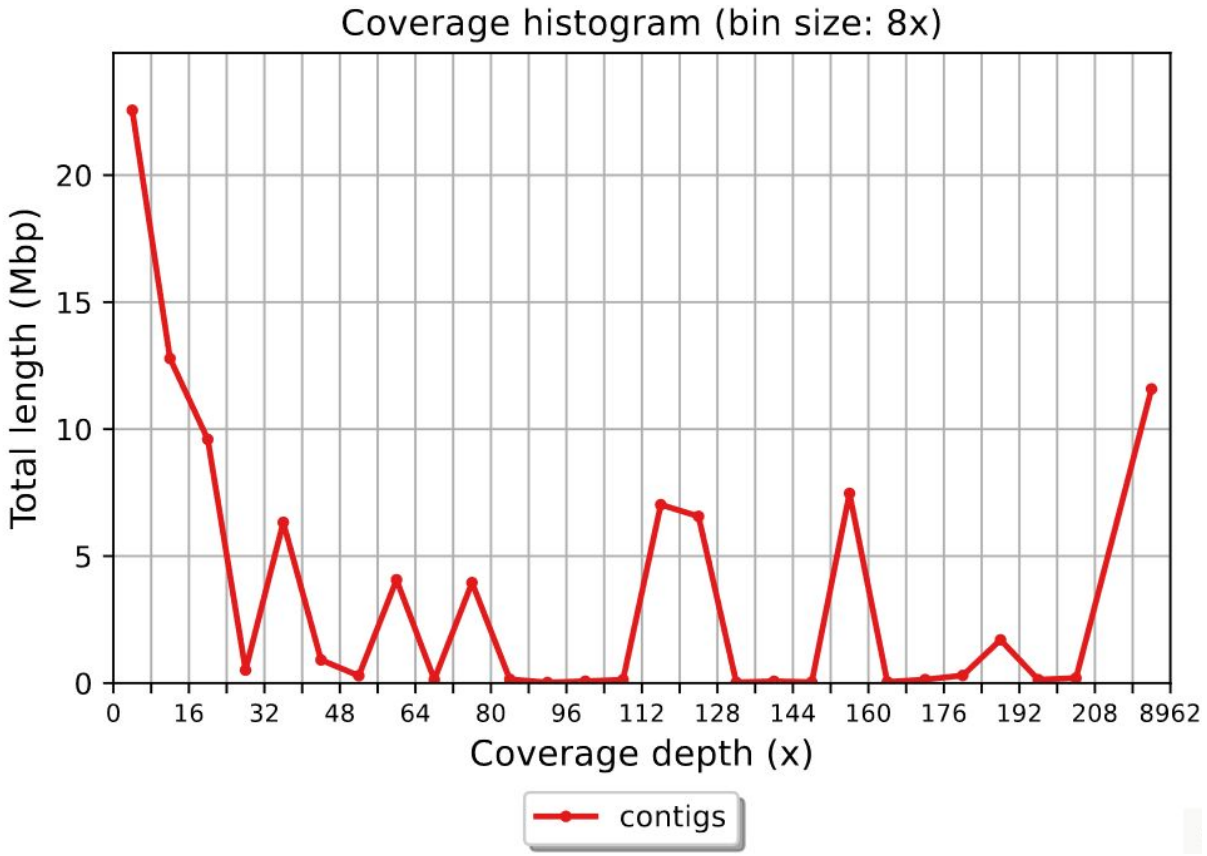

Cumulative length
Cumulative length (Mbp)
100
80
60
40
20
0
0
4000
8000
12000
16000
Contig index
contigs

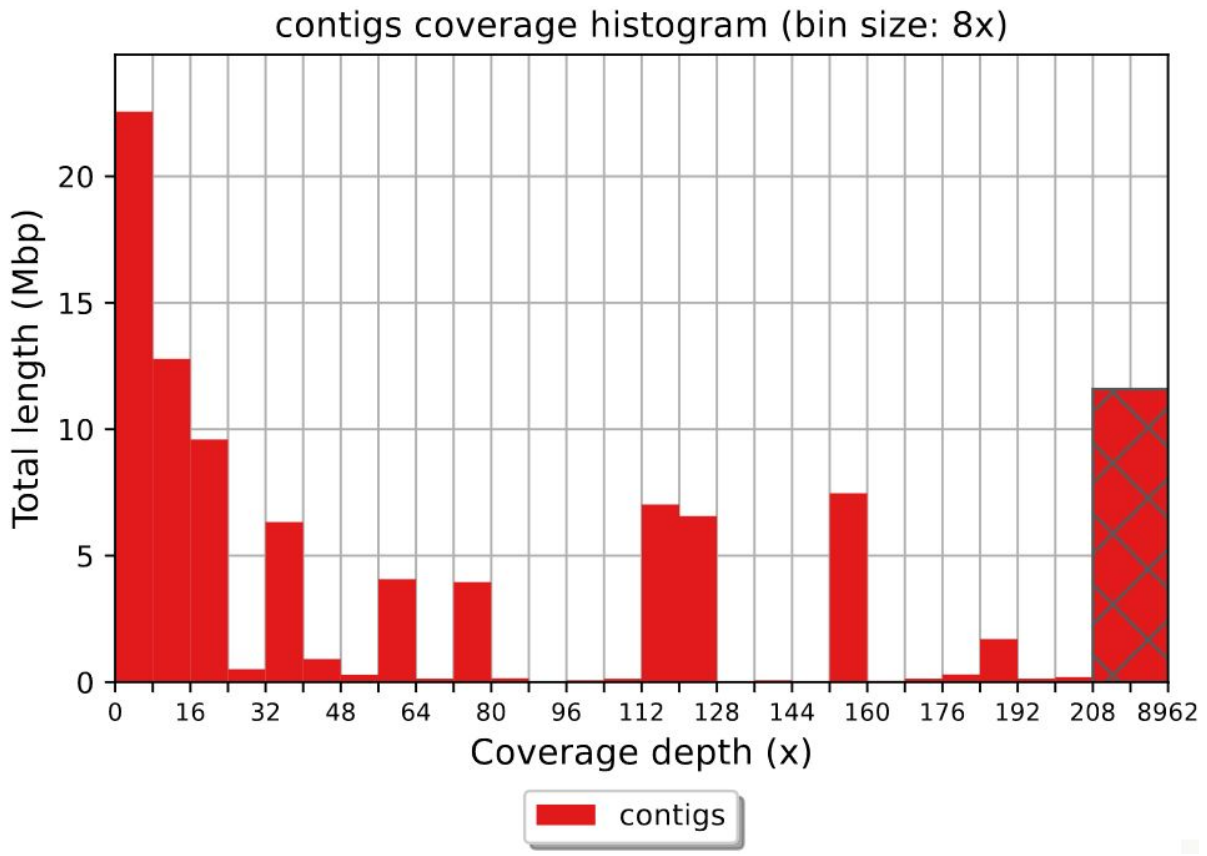
RESULTS


GC content
30000
24000
18000
12000
6000
0
windows
0
20
40
60
80
100
GC (%)
contigs


contigs GC content
5000
4000
3000
2000
1000
0
contigs
0
20
40
60
80
100
GC (%)
contigs

RESULTS


Coverage histogram (bin size: 8x)
Total length (Mbp)
0
5
10
15
20
0
16
32
48
64
80
96
112
128
144
160
176
192
208
8962
Coverage depth (x)
contigs


contigs coverage histogram (bin size: 8x)
Total length (Mbp)
0
5
10
15
20
0
16
32
48
64
80
96
112
128
144
160
176
192
208
8962
Coverage depth (x)
contigs

REFERENCES

[1] University of South Florida. Metagenomics. 2014. url: http://hulab.ucf.edu/research/projects/metagenomics/introduction.html.

[2] Kaushik, Shalini, et al. "Functional Genomics". In: Encyclopedia of Bioinformatics and Computational Biology. Oxford: Academic Press, 2019, pp. 118–133. isbn: 978-0-12-811432-2. doi: https://doi.org/10.1016/B978-0-12-809633-8.20222-7.

[3] van der Walt, A., et al. Assembling metagenomes, one community at a time. BMC Genomics **18,** 521 (2017). https://doi.org/10.1186/s12864-017-3918-9

[4] Lee, Michael D., Assembling a metagenome and recovering "genomes" with Anvi'o, Happy Belly Informatics (2019). url.: https://astrobiomike.github.io/genomics/metagen_anvio#building-up-our-contigs-database

[5] The PATRIC Team, Tutorial: Genome Assembly Service, Video 3: Selecting assembly strategies and submitting the assembly job (2020), url: https://docs.patricbrc.org/videos/genome_assembly_service.html#assembly3

[6] Nurk, Sergey et al. "metaSPAdes: a new versatile metagenomic assembler." *Genome research* vol. 27,5 (2017): 824-834. doi:10.1101/gr.213959.116

[7] Melbourne Informatics, De novo genome assembly using Velvet, url: https://www.melbournebioinformatics.org.au/tutorials/tutorials/assembly/assembly-background/#k-mer-size-and-coverage-cutoff-values

[8] Uditha Maduranga, Genome Assembly using de Bruijn Graphs, Towards Data Science, (2020). url:

https://towardsdatascience.com/genome-assembly-using-de-bruijn-graphs-69570efcc270

[9] Quast 5.0.2 Manual. url: http://quast.sourceforge.net/docs/manual.html#sec3

[10] Vijini Mallawaarachchi, Visualising Assembly Graphs, Towards Data Science, url: https://towardsdatascience.com/visualising-assembly-graphs-fb631f46bbd1