Genotyping, Taxonomic and Quality Assessment: B3 Preliminary Results

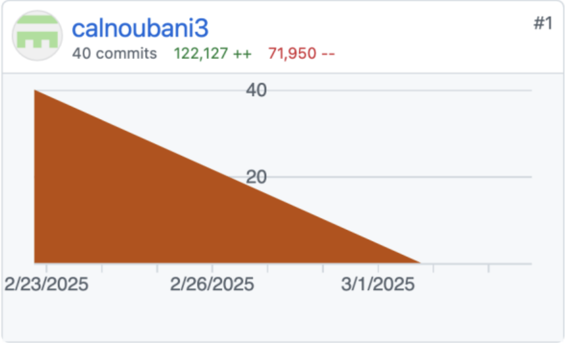
By: Lydia Keller, Celine Al-Noubani, Kyungbeom Kim, Eunsu Hwang, & Krisha Shetty

Georgia Tech

Breakdown & Student Roles

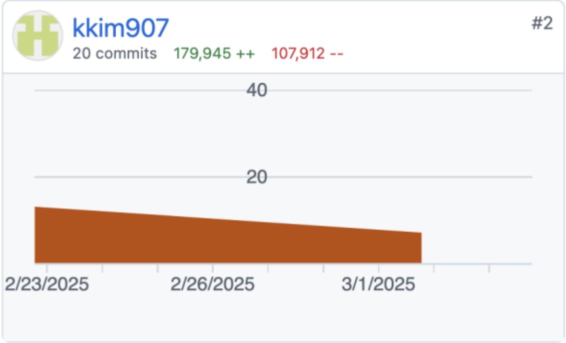
1. Use various taxonomic classification methods to determine identity of samples
   - Genus Level:
     - **Mash** - Lydia
     - **ANI Calculator** - Kyungbeom
   - Species Level:
     - **FastANI** - Krisha
     - **Skani** - Eunsu
2. Genotype our samples with **MLST** - Lydia
3. Quality Assessment Results
   - Contamination and completeness of assemblies (**CheckM**) - Celine
   - (Fine) Contig-by-contig (**Kraken2**) - Kyungbeom

Student Roles - Github



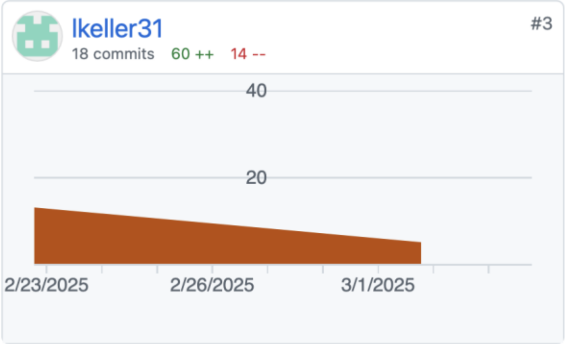
calnoubani3 #1

40 commits 122,127 ++ 71,950 --



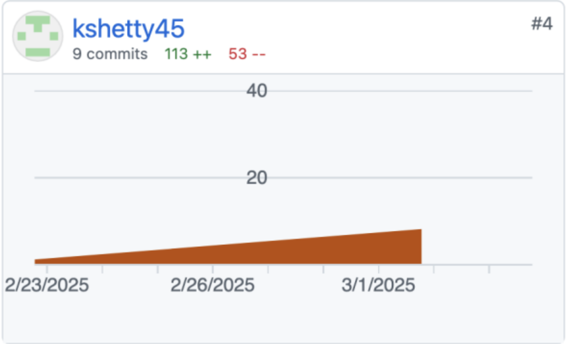
kkim907 #2

20 commits 179,945 ++ 107,912 --



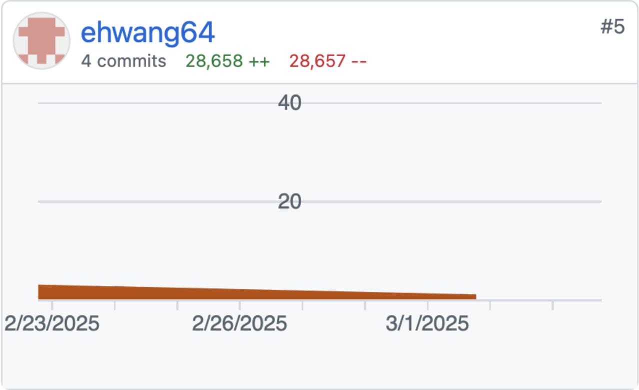
lkeller31 #3

18 commits 60 ++ 14 --



kshetty45 #4

9 commits 113 ++ 53 --

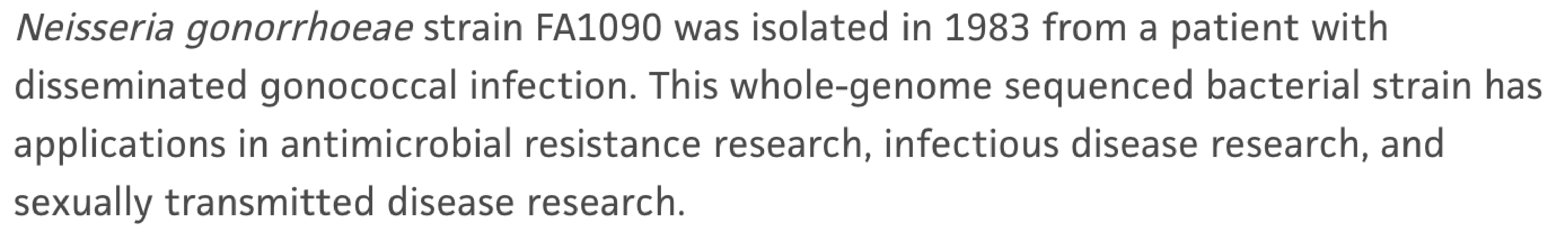


ehwang64 #5

4 commits 28,658 ++ 28,657 --

Preliminary Data Required

Largest Dataset: B1299860_S01_L001
Smallest Dataset: B1838859_S01_L001
Reference Assembly: Neisseria gonorrhoeae FA 1090

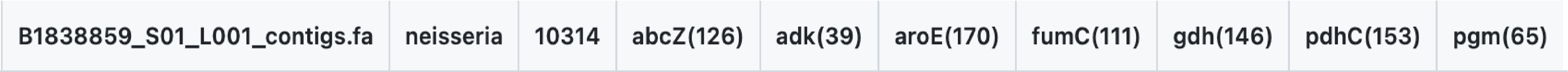


Neisseria gonorrhoeae strain FA1090 was isolated in 1983 from a patient with disseminated gonococcal infection. This whole-genome sequenced bacterial strain has applications in antimicrobial resistance research, infectious disease research, and sexually transmitted disease research.

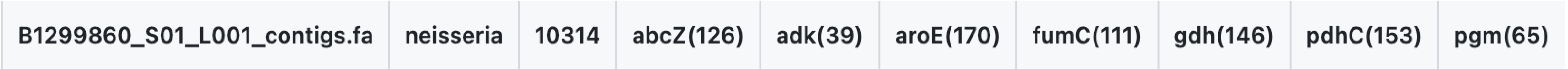
ATCC

MLST: Genotyping Tool

- Identifies isolates of bacterial species using the sequences of internal fragments of (usually) seven house-keeping genes (https://pubmlst.org/multilocus-sequence-typing)
- Ignores exact sequence differences in favor of giving sequences "allele numbers"
- Seven genes of interest can be identified from PCR products if culturing is not available



B1838859_S01_L001_contigs.fa	neisseria	10314	abcZ(126)	adk(39)	aroE(170)	fumC(111)	gdh(146)	pdhC(153)	pgm(65)

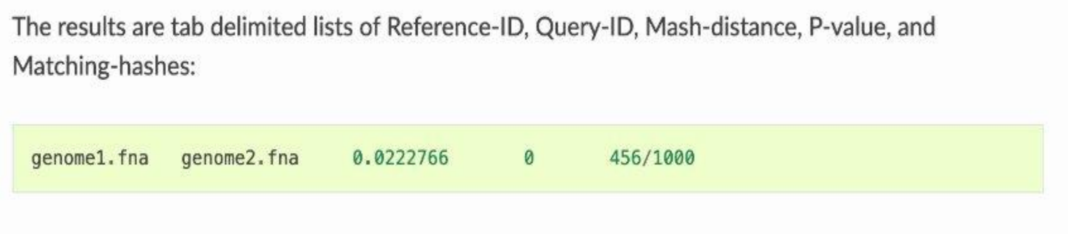


B1299860_S01_L001_contigs.fa	neisseria	10314	abcZ(126)	adk(39)	aroE(170)	fumC(111)	gdh(146)	pdhC(153)	pgm(65)

Taxonomic Classifiers – Genus Level

- MASH
  - Mash dist <genome1> <genome2>
    - "Mash distances correlate well with ANI (a common measure of genome similarity), with $D \approx 1 - ANI$ "
    - Mash distance ≤0.05 = ANI of ≥95 %
      - "This threshold roughly corresponds to a 70 % DNA-DNA reassociation value"

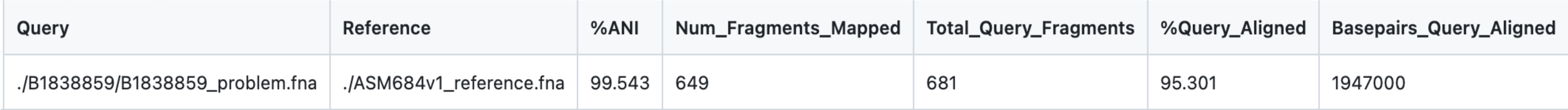
Mash distance	ANI Score
0.00396608	99.6%



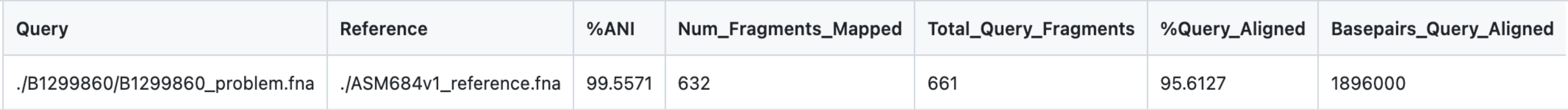
The results are tab delimited lists of Reference-ID, Query-ID, Mash-distance, P-value, and Matching-hashes:

```
genome1.fna   genome2.fna     0.0222766      0        456/1000
```

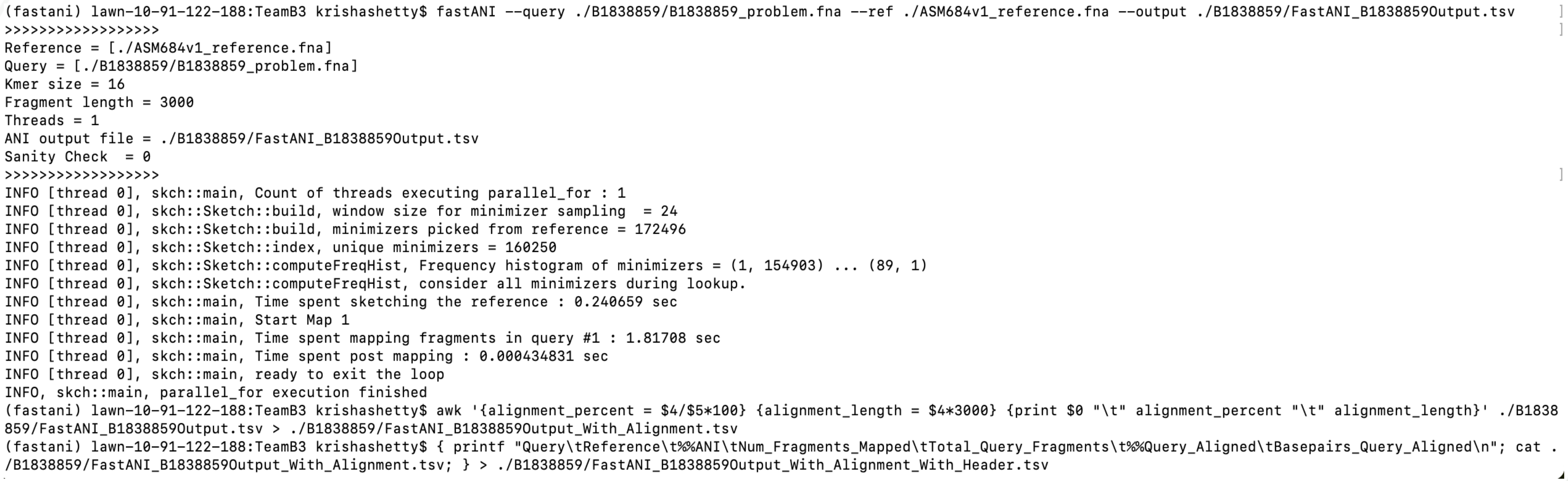
Taxonomic Classifiers – FastANI (Species Level)



Query	Reference	%ANI	Num_Fragments_Mapped	Total_Query_Fragments	%Query_Aligned	Basepairs_Query_Aligned
./B1838859/B1838859_problem.fna	./ASM684v1_reference.fna	99.543	649	681	95.301	1947000



Query	Reference	%ANI	Num_Fragments_Mapped	Total_Query_Fragments	%Query_Aligned	Basepairs_Query_Aligned
./B1299860/B1299860_problem.fna	./ASM684v1_reference.fna	99.5571	632	661	95.6127	1896000



```
(fastani) lawn-10-91-122-188:TeamB3 krishashetty$ fastANI --query ./B1838859/B1838859_problem.fna --ref ./ASM684v1_reference.fna --output ./B1838859/FastANI_B1838859Output.tsv
>>>>>>>>>>>>>>>>>>
Reference = [./ASM684v1_reference.fna]
Query = [./B1838859/B1838859_problem.fna]
Kmer size = 16
Fragment length = 3000
Threads = 1
ANI output file = ./B1838859/FastANI_B1838859Output.tsv
Sanity Check  = 0
>>>>>>>>>>>>>>>>>>
INFO [thread 0], skch::main, Count of threads executing parallel_for : 1
INFO [thread 0], skch::Sketch::build, window size for minimizer sampling  = 24
INFO [thread 0], skch::Sketch::build, minimizers picked from reference = 172496
INFO [thread 0], skch::Sketch::index, unique minimizers = 160250
INFO [thread 0], skch::Sketch::computeFreqHist, Frequency histogram of minimizers = (1, 154903) ... (89, 1)
INFO [thread 0], skch::Sketch::computeFreqHist, consider all minimizers during lookup.
INFO [thread 0], skch::main, Time spent sketching the reference : 0.240659 sec
INFO [thread 0], skch::main, Start Map 1
INFO [thread 0], skch::main, Time spent mapping fragments in query #1 : 1.81708 sec
INFO [thread 0], skch::main, Time spent post mapping : 0.000434831 sec
INFO [thread 0], skch::main, ready to exit the loop
INFO, skch::main, parallel_for execution finished
(fastani) lawn-10-91-122-188:TeamB3 krishashetty$ awk '{alignment_percent = $4/$5*100} {alignment_length = $4*3000} {print $0 "\t" alignment_percent "\t" alignment_length}' ./B1838
859/FastANI_B1838859Output.tsv > ./B1838859/FastANI_B1838859Output_With_Alignment.tsv
(fastani) lawn-10-91-122-188:TeamB3 krishashetty$ { printf "Query\tReference\t%%ANI\tNum_Fragments_Mapped\tTotal_Query_Fragments\t%%Query_Aligned\tBasepairs_Query_Aligned\n"; cat .
/B1838859/FastANI_B1838859Output_With_Alignment.tsv; } > ./B1838859/FastANI_B1838859Output_With_Alignment_With_Header.tsv
```

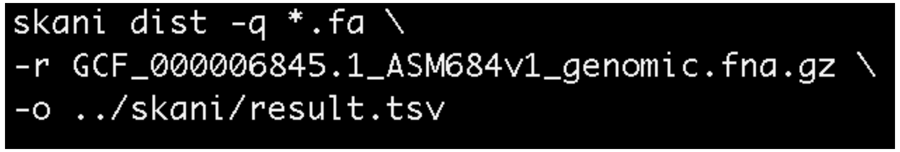
Taxonomic Classifiers – skani

Skani

Whole genome sequence comparison tool designed for a taxonomic classification and genome distance estimation

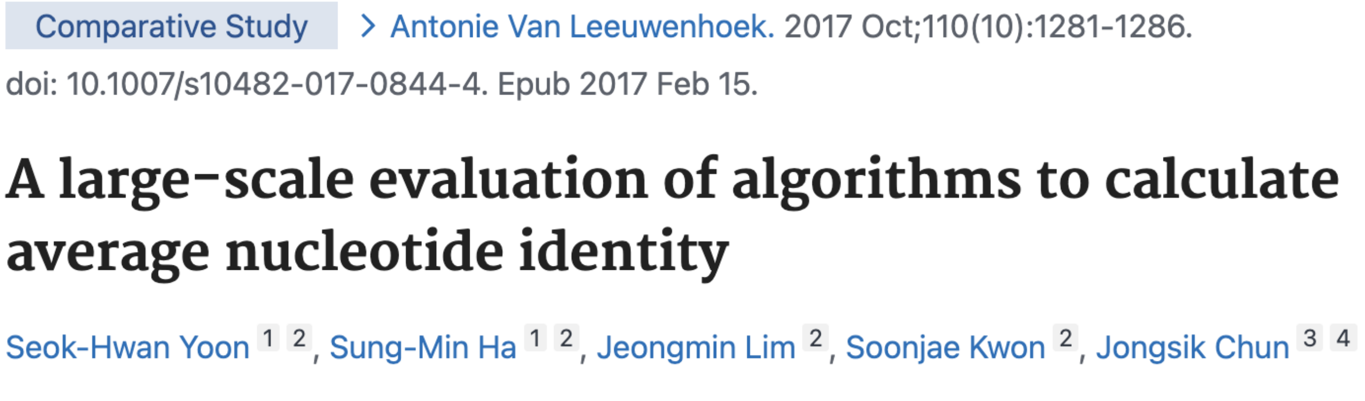
- Calculate ANI for MAGs
- Aligned fraction result : fraction of genome aligned
- Fast computation

	ANI	Align_fraction_ref	Align_fraction_query
LARGE	**99.57**	**94.01**	**94.33**
SMALL	**99.55**	**94.69**	**94.76**



```
skani dist -q *.fa \
-r GCF_000006845.1_ASM684v1_genomic.fna.gz \
-o ../skani/result.tsv
```

ANI Calculator (online Average Nucleotide Identity (ANI) calculator)



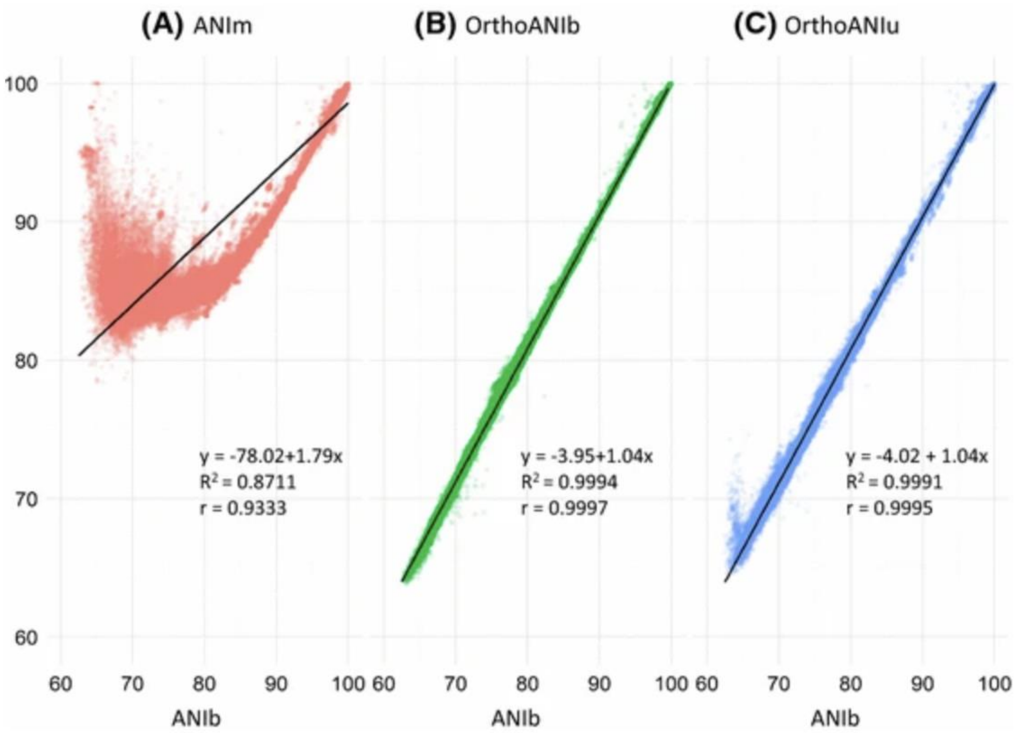
Comparative Study > Antonie Van Leeuwenhoek. 2017 Oct;110(10):1281-1286. doi: 10.1007/s10482-017-0844-4. Epub 2017 Feb 15.

A large-scale evaluation of algorithms to calculate average nucleotide identity

Seok-Hwan Yoon [1] [2], Sung-Min Ha [1] [2], Jeongmin Lim [2], Soonjae Kwon [2], Jongsik Chun [3] [4]

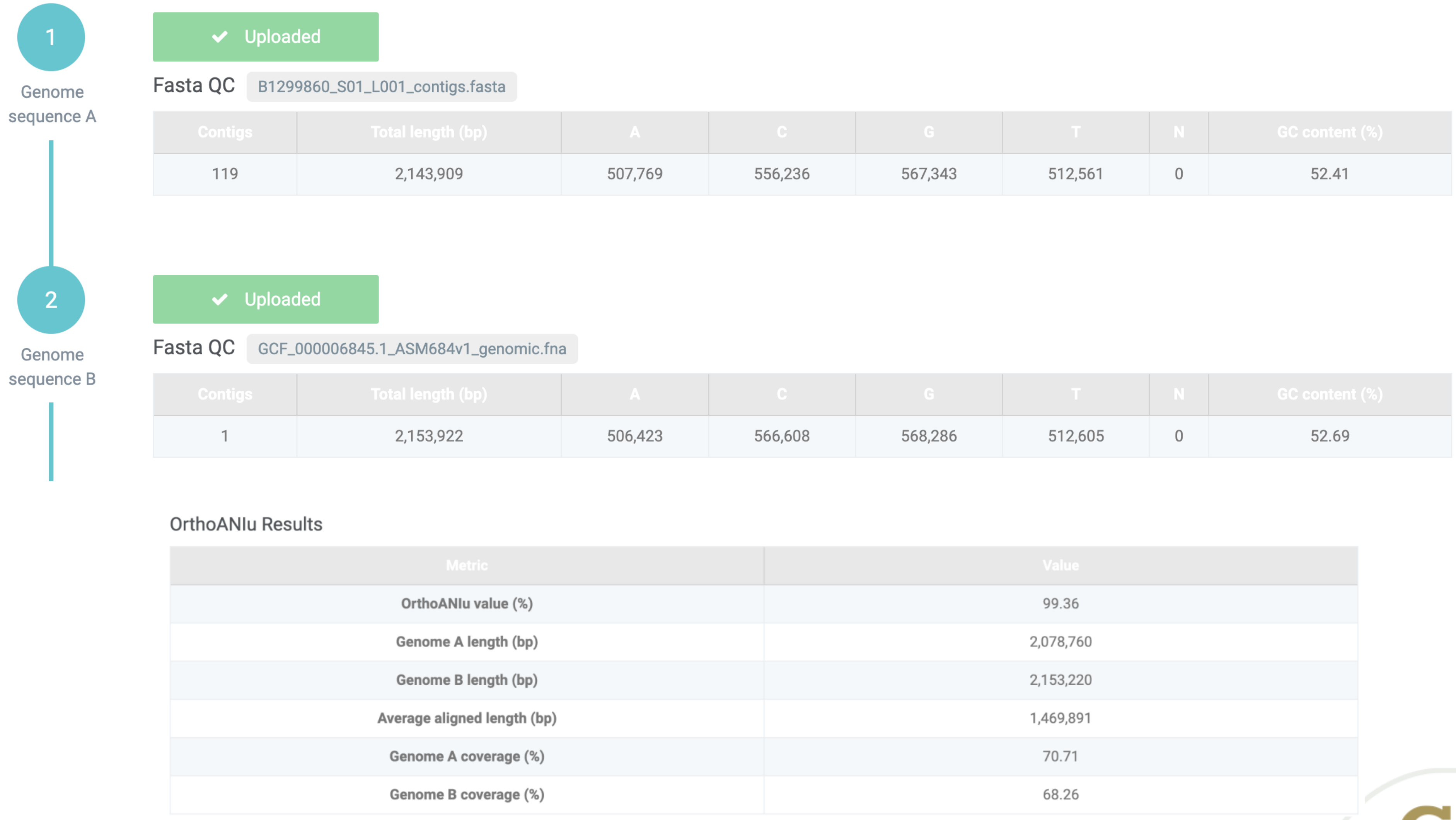
Affiliations + expand

PMID: 28204908 DOI: 10.1007/s10482-017-0844-4



ANI Calculator (online Average Nucleotide Identity (ANI) calculator)

- Largest



Fasta QC B1299860_S01_L001_contigs.fasta

Contigs	Total length (bp)	A	C	G	T	N	GC content (%)
119	2,143,909	507,769	556,236	567,343	512,561	0	52.41

Fasta QC GCF_000006845.1_ASM684v1_genomic.fna

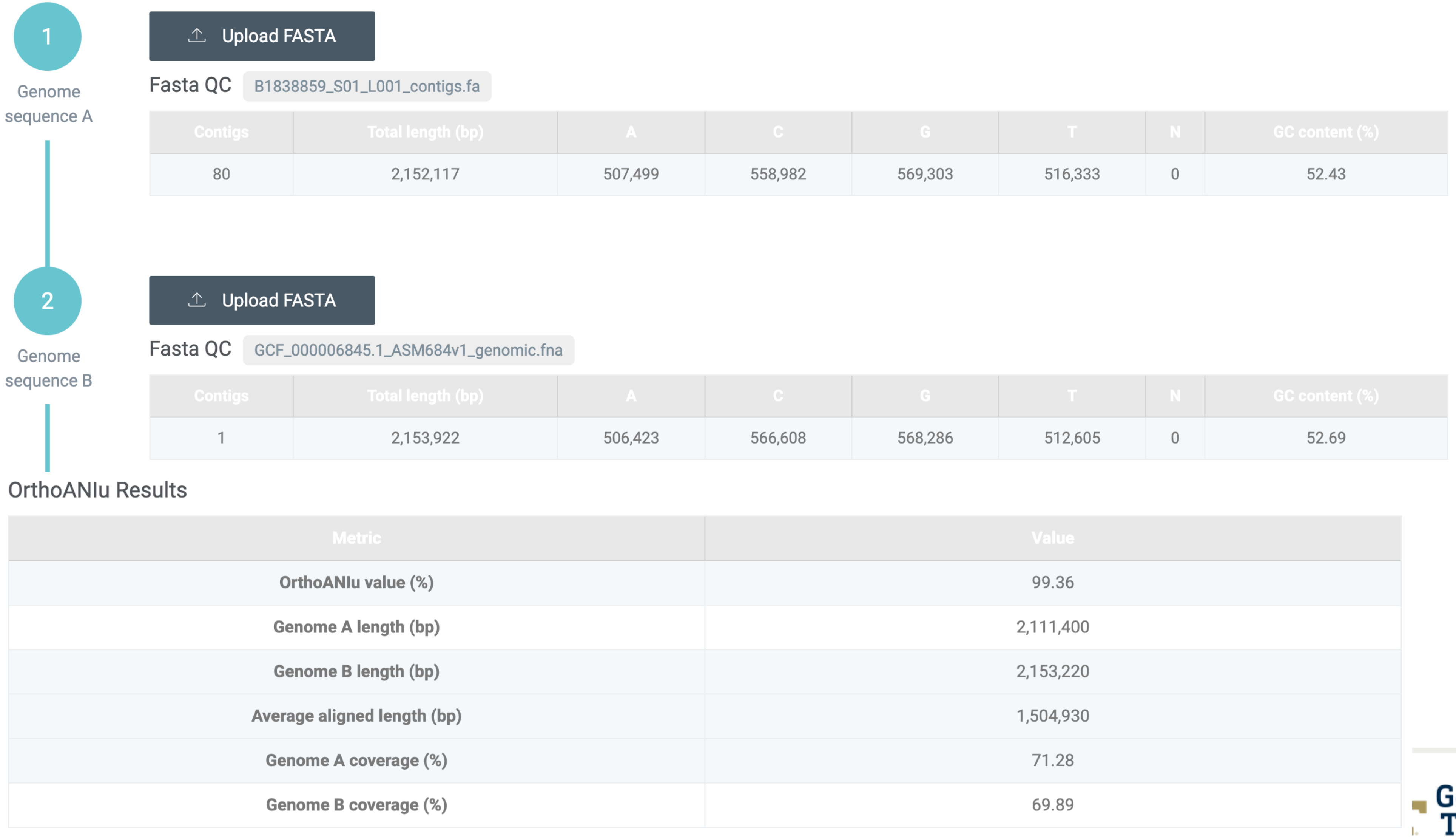
Contigs	Total length (bp)	A	C	G	T	N	GC content (%)
1	2,153,922	506,423	566,608	568,286	512,605	0	52.69

OrthoANIu Results

Metric	Value
OrthoANIu value (%)	99.36
Genome A length (bp)	2,078,760
Genome B length (bp)	2,153,220
Average aligned length (bp)	1,469,891
Genome A coverage (%)	70.71
Genome B coverage (%)	68.26

ANI Calculator (online Average Nucleotide Identity (ANI) calculator)

- Smallest



Upload FASTA

Fasta QC B1838859_S01_L001_contigs.fa

Contigs	Total length (bp)	A	C	G	T	N	GC content (%)
80	2,152,117	507,499	558,982	569,303	516,333	0	52.43

Upload FASTA

Fasta QC GCF_000006845.1_ASM684v1_genomic.fna

Contigs	Total length (bp)	A	C	G	T	N	GC content (%)
1	2,153,922	506,423	566,608	568,286	512,605	0	52.69

OrthoANIu Results

Metric	Value
OrthoANIu value (%)	99.36
Genome A length (bp)	2,111,400
Genome B length (bp)	2,153,220
Average aligned length (bp)	1,504,930
Genome A coverage (%)	71.28
Genome B coverage (%)	69.89

Quality Assessment

CheckM: Whole Assembly

- CheckM uses lineage-specific marker genes derived from a curated reference database of thousands of genomes to evaluate whole genome assembly quality and detect contamination

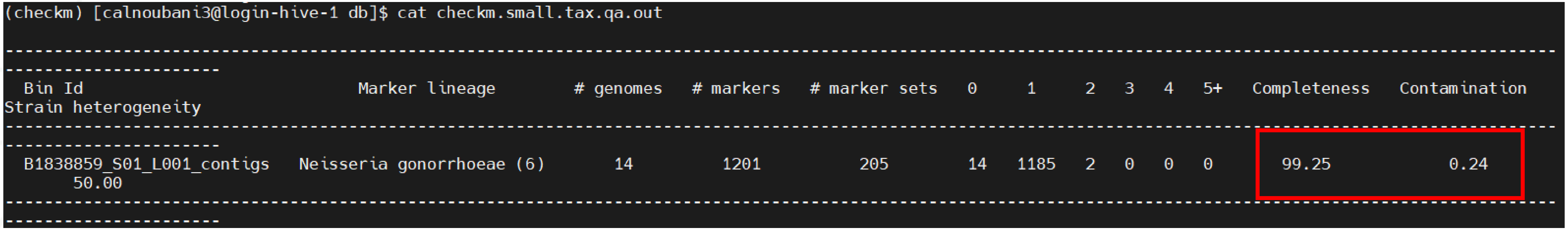
- Parameters

```
checkm \
  analyze \
  --threads 8 -x fa \
  Ng.markers \
  /storage/home/hhive1/calnoubani3/data/checkm/asm/small/ \
  analyze_small_output

- Use 8 threads
- Look for bin files with the ".fa" extension (-x fa)
- Used the generated "Ng.markers" marker set
- Input directory is the "small" assemblies folder
- Output results are saved in "analyze_small_output"
```

```
checkm \
  qa \
  --file checkm.small.tax.qa.out \
  --out_format 1 \
  --threads 8 \
  Ng.markers \
  analyze_small_output


Ran CheckM's quality assessment (QA)
```

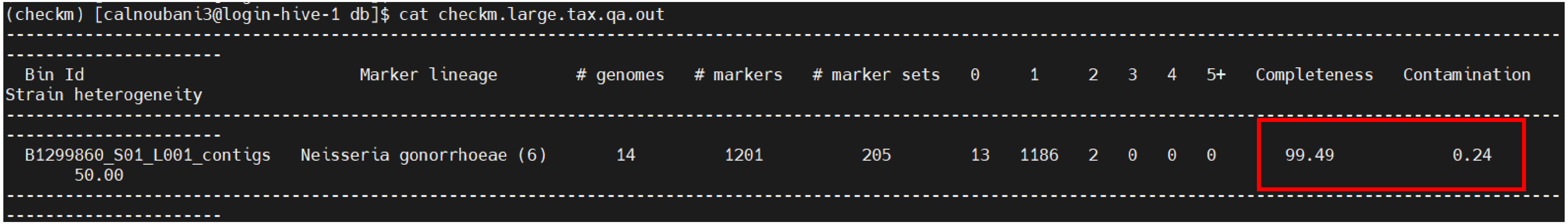
CheckM Output

- Small



```
(checkm) [calnoubani3@login-hive-1 db]$ cat checkm.small.tax.qa.out

--------------------------------------------------------------------------------------------------------------------------------
---------------------
  Bin Id                            Marker lineage         # genomes   # markers   # marker sets   0     1     2   3   4   5+   Completeness   Contamination
Strain heterogeneity
--------------------------------------------------------------------------------------------------------------------------------
---------------------
  B1838859_S01_L001_contigs   Neisseria gonorrhoeae (6)       14          1201          205         14   1185   2   0   0   0      99.25             0.24
       50.00
--------------------------------------------------------------------------------------------------------------------------------
---------------------
```

- Large

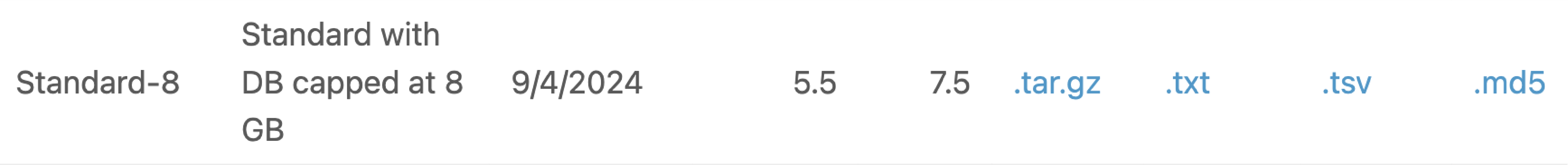


```
(checkm) [calnoubani3@login-hive-1 db]$ cat checkm.large.tax.qa.out
--------------------------------------------------------------------------------------------------------------------------------
---------------------
  Bin Id                            Marker lineage         # genomes   # markers   # marker sets   0     1     2   3   4   5+   Completeness   Contamination
Strain heterogeneity
--------------------------------------------------------------------------------------------------------------------------------
---------------------
  B1299860_S01_L001_contigs   Neisseria gonorrhoeae (6)       14          1201          205         13   1186   2   0   0   0      99.49             0.24
       50.00
--------------------------------------------------------------------------------------------------------------------------------
---------------------
```

kraken2: Contig-by-Contig

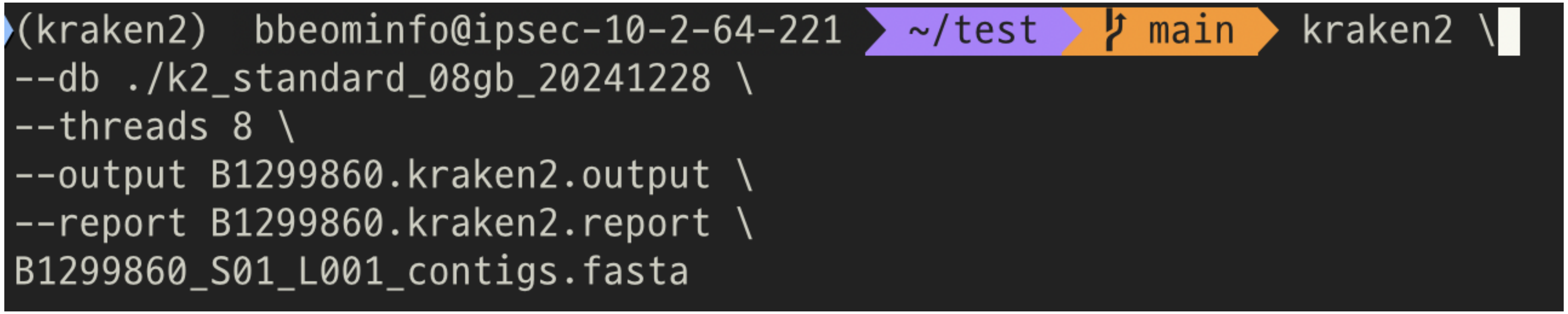
Kraken2 offers a good balance between speed and accuracy, making it widely used in metagenomic studies for contig-level classification.

- Database



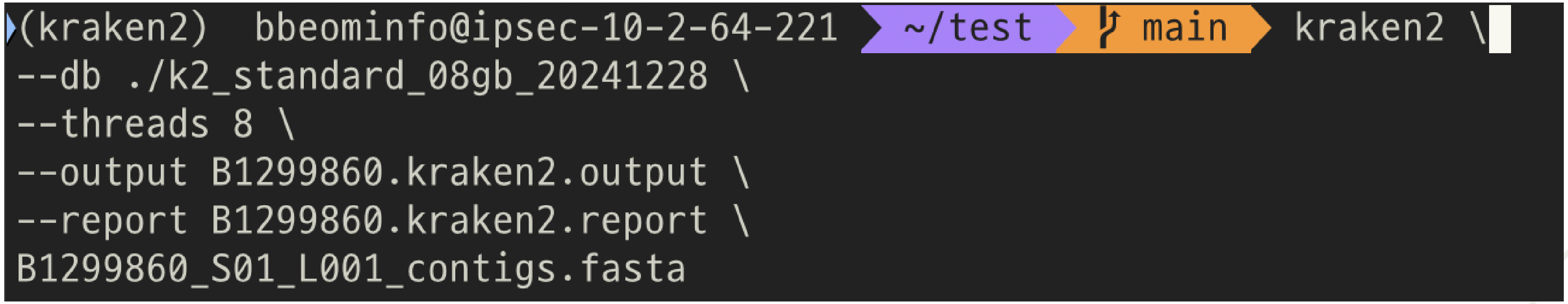
Standard-8	Standard with DB capped at 8 GB	9/4/2024	5.5	7.5	.tar.gz	.txt	.tsv	.md5

- Largest



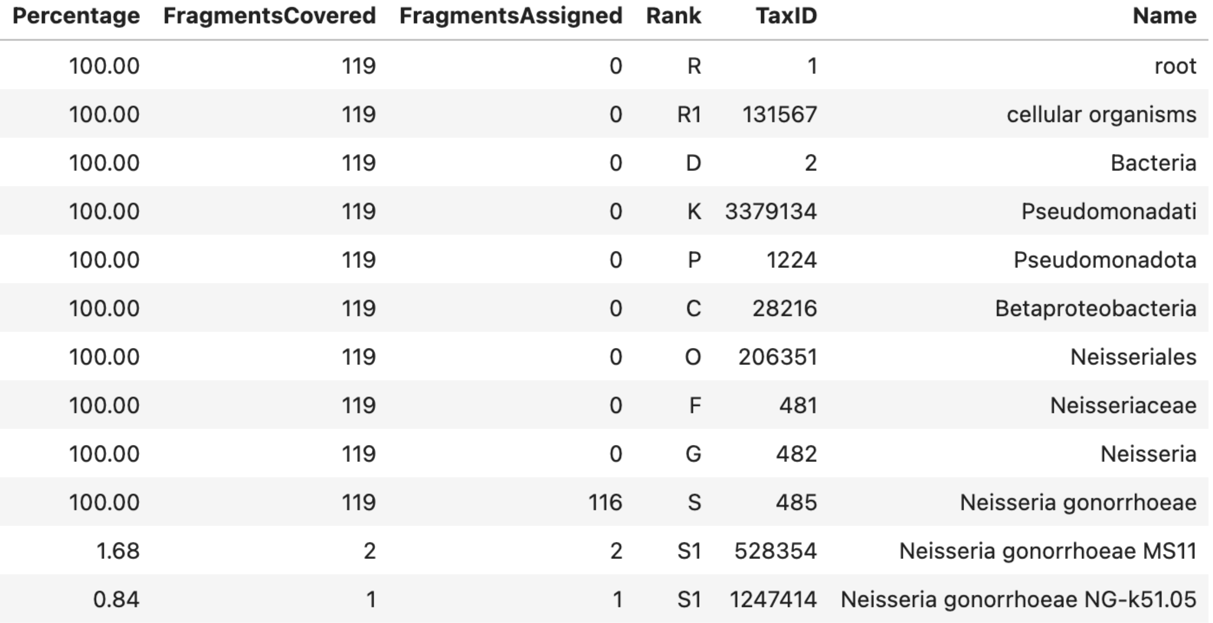
```
(kraken2)  bbeominfo@ipsec-10-2-64-221  ~/test  main  kraken2 \
--db ./k2_standard_08gb_20241228 \
--threads 8 \
--output B1299860.kraken2.output \
--report B1299860.kraken2.report \
B1299860_S01_L001_contigs.fasta
```

- Smallest

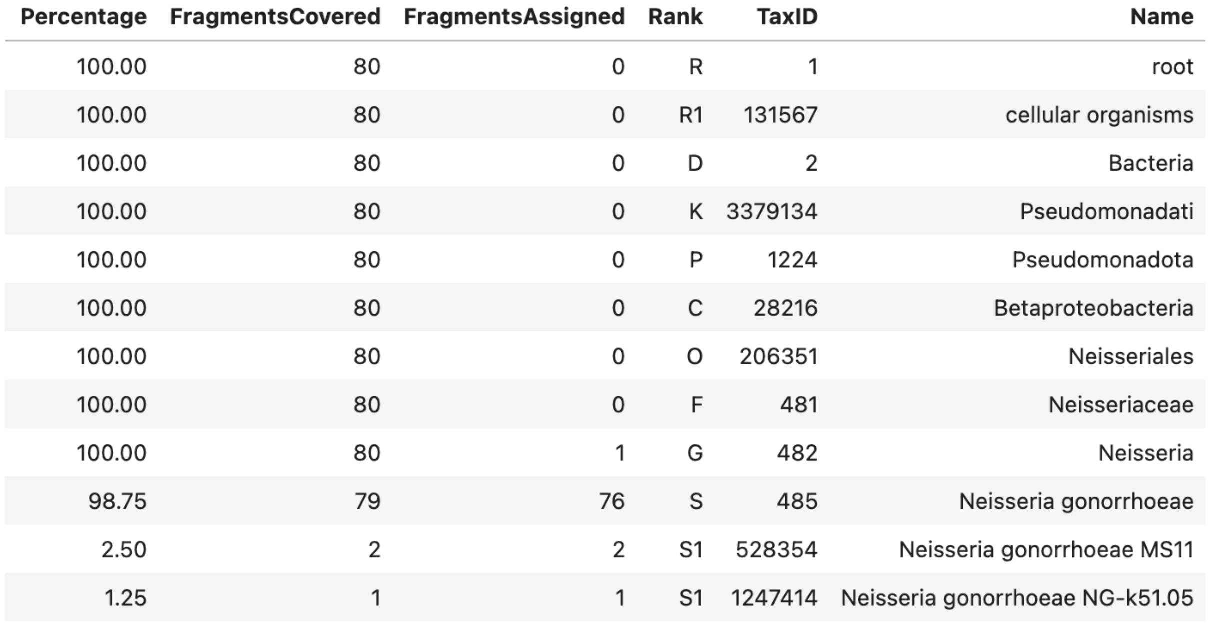


```
(kraken2)  bbeominfo@ipsec-10-2-64-221  ~/test  main  kraken2 \
--db ./k2_standard_08gb_20241228 \
--threads 8 \
--output B1299860.kraken2.output \
--report B1299860.kraken2.report \
B1299860_S01_L001_contigs.fasta
```

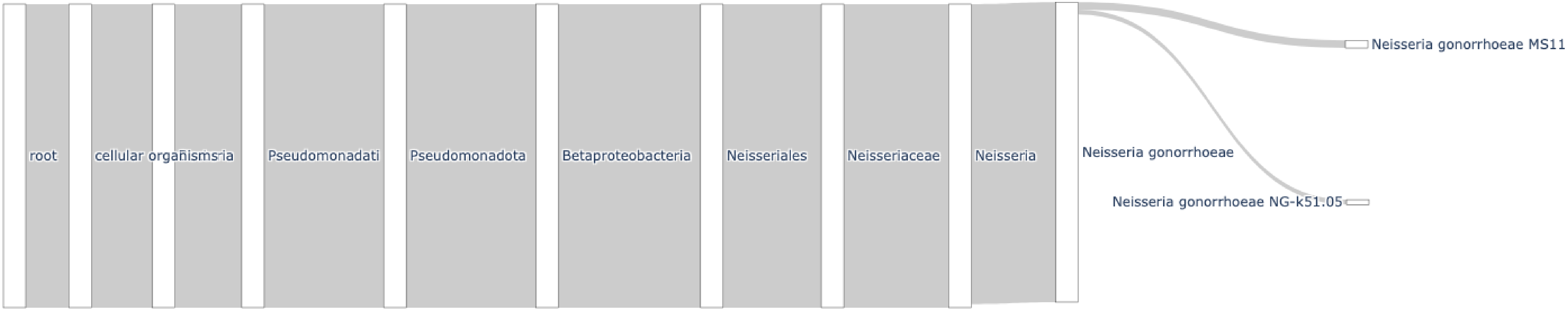
kraken2: Contig-by-Contig



Percentage	FragmentsCovered	FragmentsAssigned	Rank	TaxID	Name
100.00	119	0	R	1	root
100.00	119	0	R1	131567	cellular organisms
100.00	119	0	D	2	Bacteria
100.00	119	0	K	3379134	Pseudomonadati
100.00	119	0	P	1224	Pseudomonadota
100.00	119	0	C	28216	Betaproteobacteria
100.00	119	0	O	206351	Neisseriales
100.00	119	0	F	481	Neisseriaceae
100.00	119	0	G	482	Neisseria
100.00	119	116	S	485	Neisseria gonorrhoeae
1.68	2	2	S1	528354	Neisseria gonorrhoeae MS11
0.84	1	1	S1	1247414	Neisseria gonorrhoeae NG-k51.05



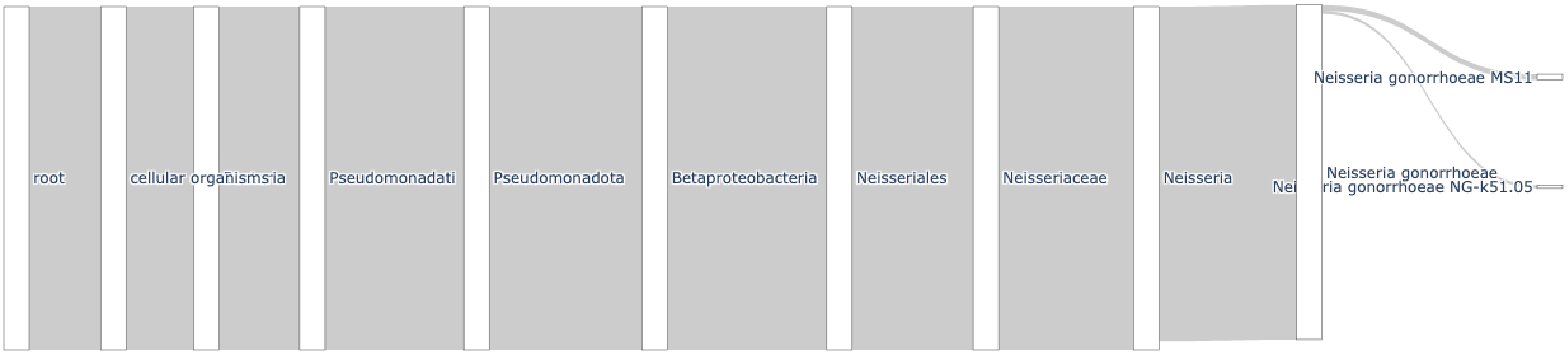
Percentage	FragmentsCovered	FragmentsAssigned	Rank	TaxID	Name
100.00	80	0	R	1	root
100.00	80	0	R1	131567	cellular organisms
100.00	80	0	D	2	Bacteria
100.00	80	0	K	3379134	Pseudomonadati
100.00	80	0	P	1224	Pseudomonadota
100.00	80	0	C	28216	Betaproteobacteria
100.00	80	0	O	206351	Neisseriales
100.00	80	0	F	481	Neisseriaceae
100.00	80	1	G	482	Neisseria
98.75	79	76	S	485	Neisseria gonorrhoeae
2.50	2	2	S1	528354	Neisseria gonorrhoeae MS11
1.25	1	1	S1	1247414	Neisseria gonorrhoeae NG-k51.05



kraken2: Contig-by-Contig

Percentage	FragmentsCovered	FragmentsAssigned	Rank	TaxID	Name
100.00	119	0	R	1	root
100.00	119	0	R1	131567	cellular organisms
100.00	119	0	D	2	Bacteria
100.00	119	0	K	3379134	Pseudomonadati
100.00	119	0	P	1224	Pseudomonadota
100.00	119	0	C	28216	Betaproteobacteria
100.00	119	0	O	206351	Neisseriales
100.00	119	0	F	481	Neisseriaceae
100.00	119	0	G	482	Neisseria
100.00	119	116	S	485	Neisseria gonorrhoeae
1.68	2	2	S1	528354	Neisseria gonorrhoeae MS11
0.84	1	1	S1	1247414	Neisseria gonorrhoeae NG-k51.05

Percentage	FragmentsCovered	FragmentsAssigned	Rank	TaxID	Name
100.00	80	0	R	1	root
100.00	80	0	R1	131567	cellular organisms
100.00	80	0	D	2	Bacteria
100.00	80	0	K	3379134	Pseudomonadati
100.00	80	0	P	1224	Pseudomonadota
100.00	80	0	C	28216	Betaproteobacteria
100.00	80	0	O	206351	Neisseriales
100.00	80	0	F	481	Neisseriaceae
100.00	80	1	G	482	Neisseria
98.75	79	76	S	485	Neisseria gonorrhoeae
2.50	2	2	S1	528354	Neisseria gonorrhoeae MS11
1.25	1	1	S1	1247414	Neisseria gonorrhoeae NG-k51.05



Conclusion

- MLST genotyping identified Neisseria
- ANI calculated baed on different tools with ANI score > 99%
- CheckM resulted in 99.25% completeness in the small sample, 99.49% completeness in the large sample, and found 0.24% contamination in both
- Kraken2: The result confirms that the assemblies are highly pure and taxonomically consistent, showing no significant contamination from unrelated species. This supports the reliability of downstream analyses using these assemblies

Future plan

- Run taxonomic classification and assessment tools on the rest of the files
- Explore other tools for finer level assessment (ex. Intra-contig assessment)

Citations

Jolley, K.A., Maiden, M.C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010). https://doi.org/10.1186/1471-2105-11-595

Ondov, B.D., Treangen, T.J., Melsted, P. *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016). https://doi.org/10.1186/s13059-016-0997-x

Ondov, B., Starrett, G., Sappington, A. *et al*. Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol* **20**, 232 (2019). https://doi.org/10.1186/s13059-019-1841-x

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015 Jul;25(7):1043-55. doi: 10.1101/gr.186072.114. Epub 2015 May 14. PMID: 25977477; PMCID: PMC4484387.

https://pubmlst.org/multilocus-sequence-typing

https://bisonnet.bucknell.edu/files/2021/05/Kraken2-Help-Sheet.pdf

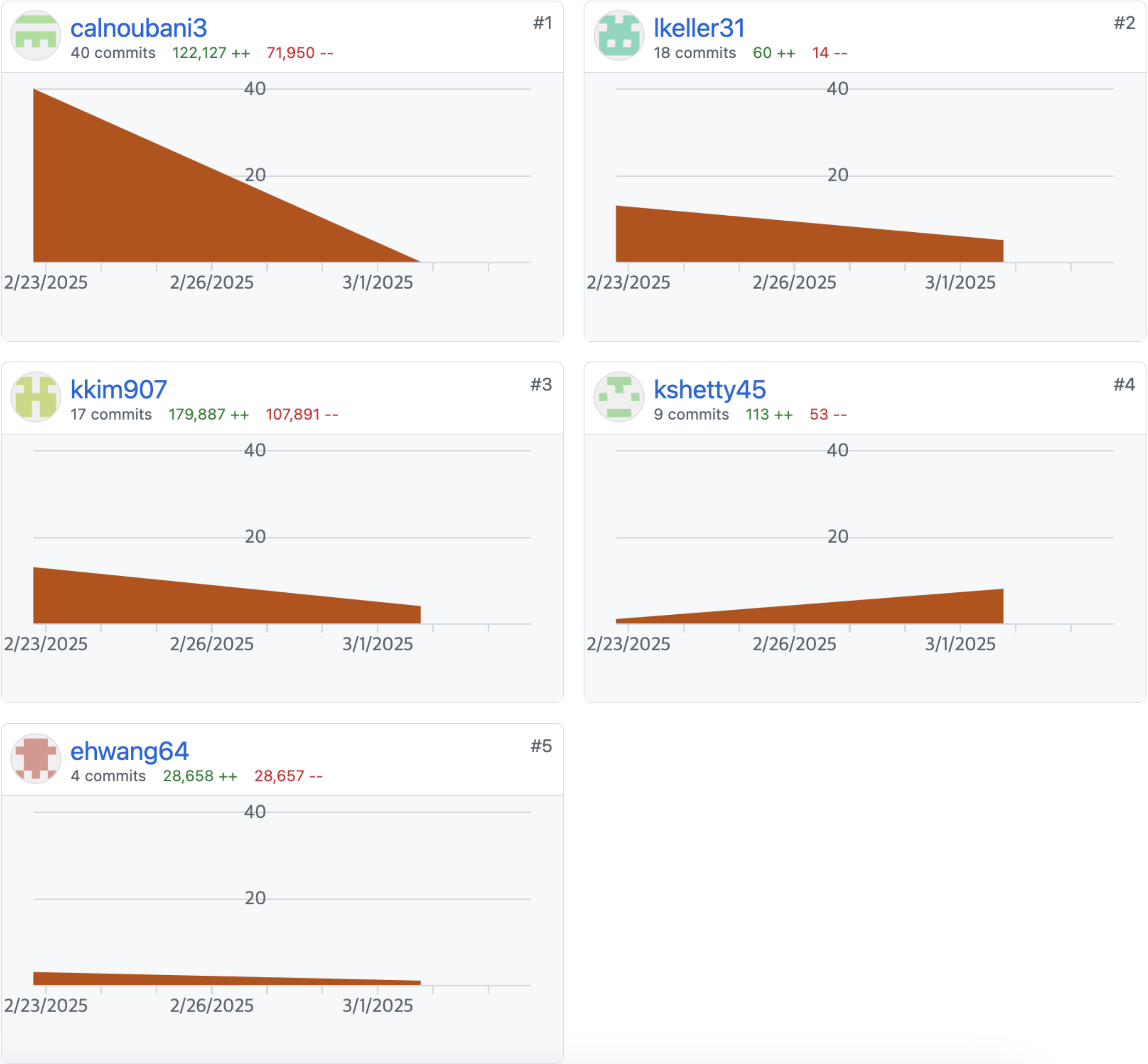
Thank you!

--Team B Group 3

Appendix

- Commands and outputs

Student Roles - Github



calnoubani3 #1
40 commits 122,127 ++ 71,950 --

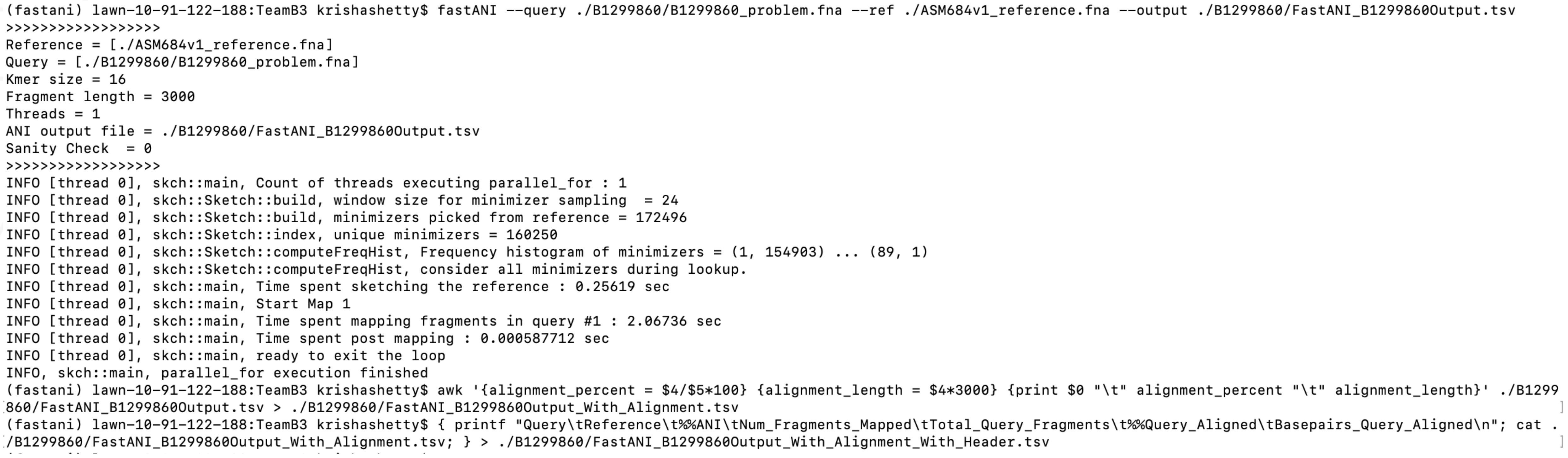
lkeller31 #2
18 commits 60 ++ 14 --

kkim907 #3
17 commits 179,887 ++ 107,891 --

kshetty45 #4
9 commits 113 ++ 53 --

ehwang64 #5
4 commits 28,658 ++ 28,657 --

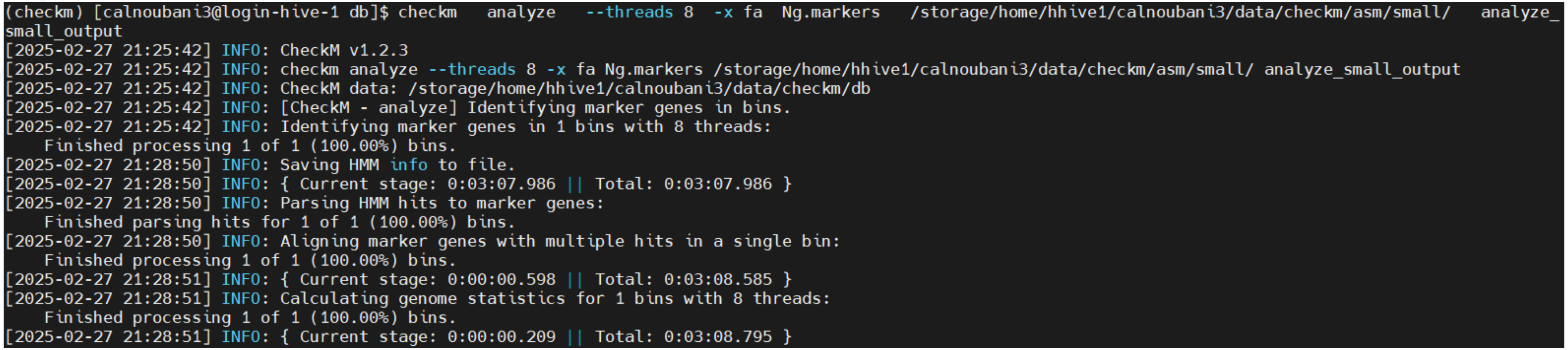
FastANI Running- B1299860



```
(fastani) lawn-10-91-122-188:TeamB3 krishashetty$ fastANI --query ./B1299860/B1299860_problem.fna --ref ./ASM684v1_reference.fna --output ./B1299860/FastANI_B1299860Output.tsv
>>>>>>>>>>>>>>>>>>
Reference = [./ASM684v1_reference.fna]
Query = [./B1299860/B1299860_problem.fna]
Kmer size = 16
Fragment length = 3000
Threads = 1
ANI output file = ./B1299860/FastANI_B1299860Output.tsv
Sanity Check  = 0
>>>>>>>>>>>>>>>>>>
INFO [thread 0], skch::main, Count of threads executing parallel_for : 1
INFO [thread 0], skch::Sketch::build, window size for minimizer sampling  = 24
INFO [thread 0], skch::Sketch::build, minimizers picked from reference = 172496
INFO [thread 0], skch::Sketch::index, unique minimizers = 160250
INFO [thread 0], skch::Sketch::computeFreqHist, Frequency histogram of minimizers = (1, 154903) ... (89, 1)
INFO [thread 0], skch::Sketch::computeFreqHist, consider all minimizers during lookup.
INFO [thread 0], skch::main, Time spent sketching the reference : 0.25619 sec
INFO [thread 0], skch::main, Start Map 1
INFO [thread 0], skch::main, Time spent mapping fragments in query #1 : 2.06736 sec
INFO [thread 0], skch::main, Time spent post mapping : 0.000587712 sec
INFO [thread 0], skch::main, ready to exit the loop
INFO, skch::main, parallel_for execution finished
(fastani) lawn-10-91-122-188:TeamB3 krishashetty$ awk '{alignment_percent = $4/$5*100} {alignment_length = $4*3000} {print $0 "\t" alignment_percent "\t" alignment_length}' ./B1299
860/FastANI_B1299860Output.tsv > ./B1299860/FastANI_B1299860Output_With_Alignment.tsv
(fastani) lawn-10-91-122-188:TeamB3 krishashetty$ { printf "Query\tReference\t%%ANI\tNum_Fragments_Mapped\tTotal_Query_Fragments\t%%Query_Aligned\tBasepairs_Query_Aligned\n"; cat .
/B1299860/FastANI_B1299860Output_With_Alignment.tsv; } > ./B1299860/FastANI_B1299860Output_With_Alignment_With_Header.tsv
```

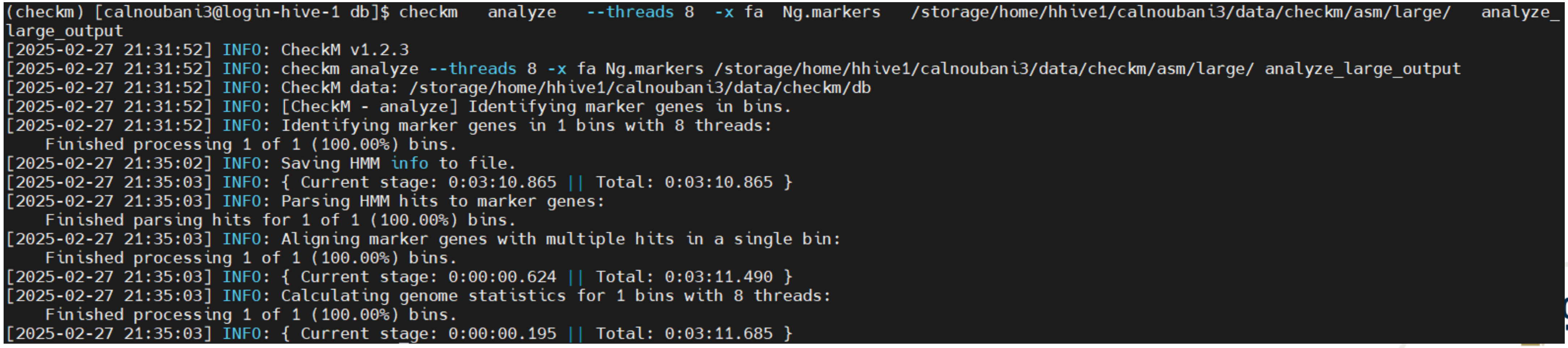
CheckM Running

- Small



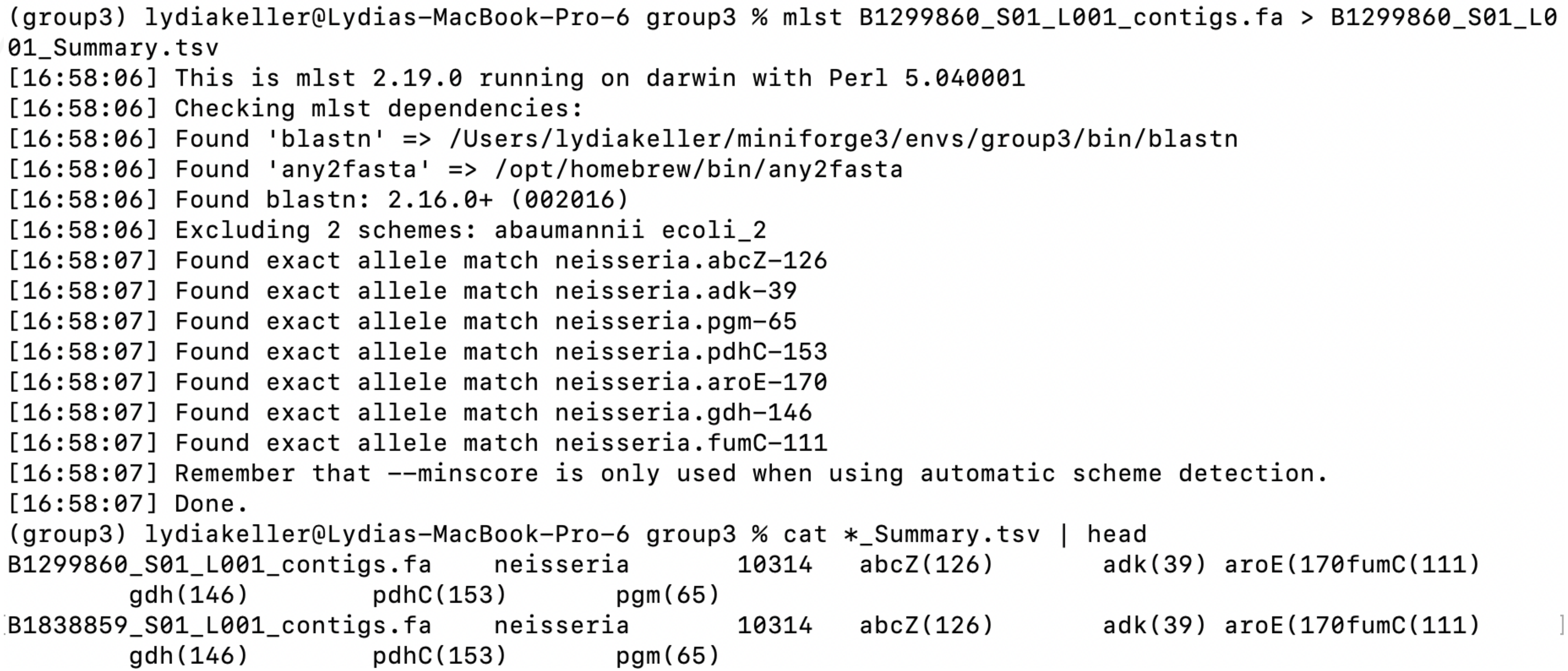
```
(checkm) [calnoubani3@login-hive-1 db]$ checkm   analyze   --threads 8  -x fa  Ng.markers   /storage/home/hhive1/calnoubani3/data/checkm/asm/small/   analyze_
small_output
[2025-02-27 21:25:42] INFO: CheckM v1.2.3
[2025-02-27 21:25:42] INFO: checkm analyze --threads 8 -x fa Ng.markers /storage/home/hhive1/calnoubani3/data/checkm/asm/small/ analyze_small_output
[2025-02-27 21:25:42] INFO: CheckM data: /storage/home/hhive1/calnoubani3/data/checkm/db
[2025-02-27 21:25:42] INFO: [CheckM - analyze] Identifying marker genes in bins.
[2025-02-27 21:25:42] INFO: Identifying marker genes in 1 bins with 8 threads:
    Finished processing 1 of 1 (100.00%) bins.
[2025-02-27 21:28:50] INFO: Saving HMM info to file.
[2025-02-27 21:28:50] INFO: { Current stage: 0:03:07.986 || Total: 0:03:07.986 }
[2025-02-27 21:28:50] INFO: Parsing HMM hits to marker genes:
    Finished parsing hits for 1 of 1 (100.00%) bins.
[2025-02-27 21:28:50] INFO: Aligning marker genes with multiple hits in a single bin:
    Finished processing 1 of 1 (100.00%) bins.
[2025-02-27 21:28:51] INFO: { Current stage: 0:00:00.598 || Total: 0:03:08.585 }
[2025-02-27 21:28:51] INFO: Calculating genome statistics for 1 bins with 8 threads:
    Finished processing 1 of 1 (100.00%) bins.
[2025-02-27 21:28:51] INFO: { Current stage: 0:00:00.209 || Total: 0:03:08.795 }
```

- Large



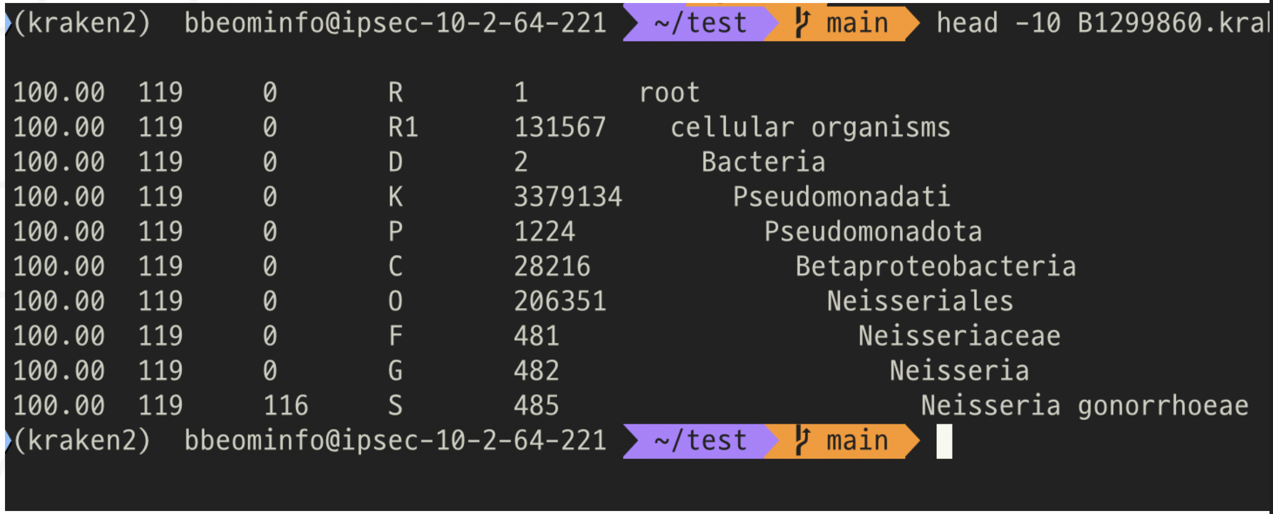
```
(checkm) [calnoubani3@login-hive-1 db]$ checkm   analyze   --threads 8  -x fa  Ng.markers   /storage/home/hhive1/calnoubani3/data/checkm/asm/large/   analyze_
large_output
[2025-02-27 21:31:52] INFO: CheckM v1.2.3
[2025-02-27 21:31:52] INFO: checkm analyze --threads 8 -x fa Ng.markers /storage/home/hhive1/calnoubani3/data/checkm/asm/large/ analyze_large_output
[2025-02-27 21:31:52] INFO: CheckM data: /storage/home/hhive1/calnoubani3/data/checkm/db
[2025-02-27 21:31:52] INFO: [CheckM - analyze] Identifying marker genes in bins.
[2025-02-27 21:31:52] INFO: Identifying marker genes in 1 bins with 8 threads:
    Finished processing 1 of 1 (100.00%) bins.
[2025-02-27 21:35:02] INFO: Saving HMM info to file.
[2025-02-27 21:35:03] INFO: { Current stage: 0:03:10.865 || Total: 0:03:10.865 }
[2025-02-27 21:35:03] INFO: Parsing HMM hits to marker genes:
    Finished parsing hits for 1 of 1 (100.00%) bins.
[2025-02-27 21:35:03] INFO: Aligning marker genes with multiple hits in a single bin:
    Finished processing 1 of 1 (100.00%) bins.
[2025-02-27 21:35:03] INFO: { Current stage: 0:00:00.624 || Total: 0:03:11.490 }
[2025-02-27 21:35:03] INFO: Calculating genome statistics for 1 bins with 8 threads:
    Finished processing 1 of 1 (100.00%) bins.
[2025-02-27 21:35:03] INFO: { Current stage: 0:00:00.195 || Total: 0:03:11.685 }
```

MLST

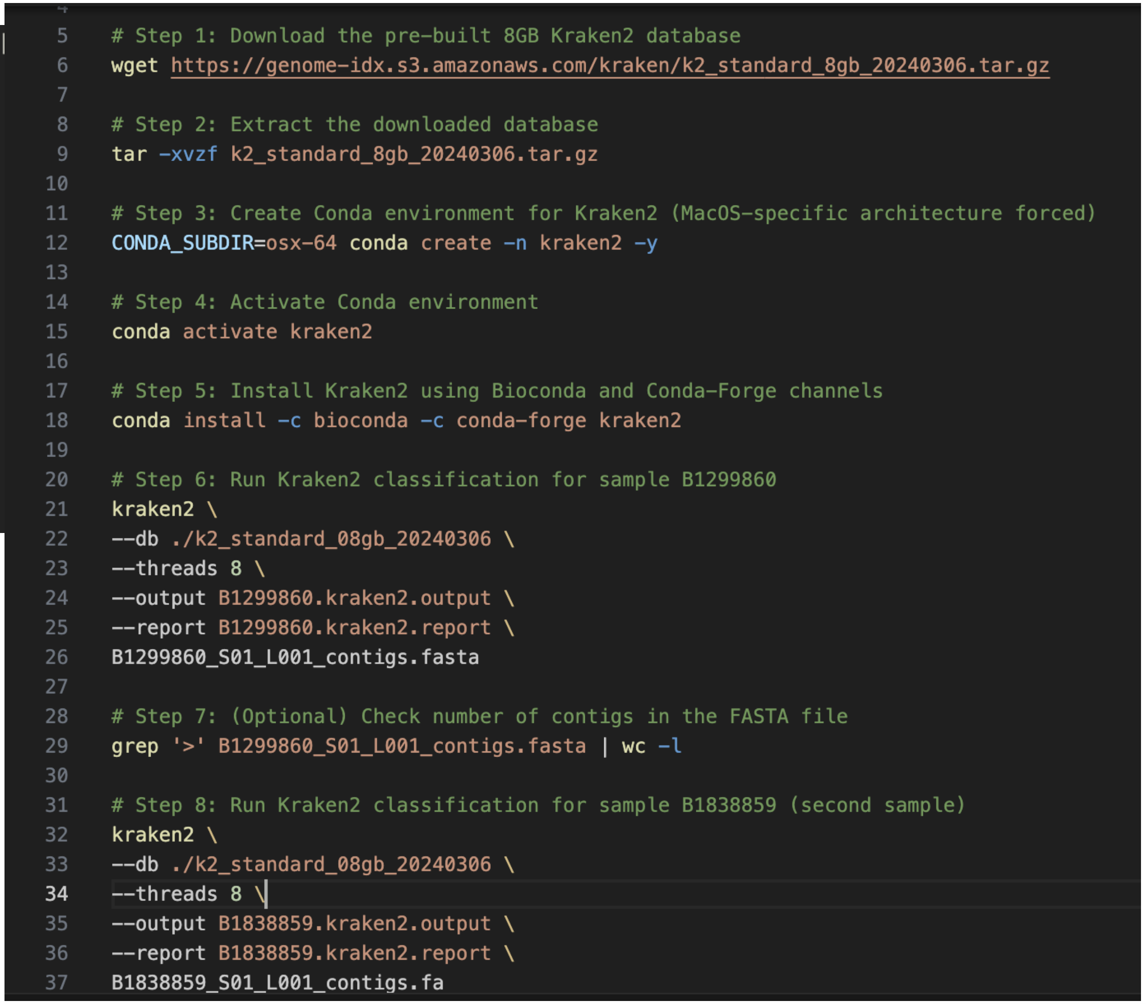


```
(group3) lydiakeller@Lydias-MacBook-Pro-6 group3 % mlst B1299860_S01_L001_contigs.fa > B1299860_S01_L0
01_Summary.tsv
[16:58:06] This is mlst 2.19.0 running on darwin with Perl 5.040001
[16:58:06] Checking mlst dependencies:
[16:58:06] Found 'blastn' => /Users/lydiakeller/miniforge3/envs/group3/bin/blastn
[16:58:06] Found 'any2fasta' => /opt/homebrew/bin/any2fasta
[16:58:06] Found blastn: 2.16.0+ (002016)
[16:58:06] Excluding 2 schemes: abaumannii ecoli_2
[16:58:07] Found exact allele match neisseria.abcZ-126
[16:58:07] Found exact allele match neisseria.adk-39
[16:58:07] Found exact allele match neisseria.pgm-65
[16:58:07] Found exact allele match neisseria.pdhC-153
[16:58:07] Found exact allele match neisseria.aroE-170
[16:58:07] Found exact allele match neisseria.gdh-146
[16:58:07] Found exact allele match neisseria.fumC-111
[16:58:07] Remember that --minscore is only used when using automatic scheme detection.
[16:58:07] Done.
(group3) lydiakeller@Lydias-MacBook-Pro-6 group3 % cat *_Summary.tsv | head
B1299860_S01_L001_contigs.fa    neisseria      10314   abcZ(126)       adk(39) aroE(170fumC(111)
        gdh(146)        pdhC(153)       pgm(65)
B1838859_S01_L001_contigs.fa    neisseria      10314   abcZ(126)       adk(39) aroE(170fumC(111)
        gdh(146)        pdhC(153)       pgm(65)
```

Kraken2



```
(kraken2)  bbeominfo@ipsec-10-2-64-221  ~/test  main  head -10 B1299860.kra

100.00  119     0       R       1       root
100.00  119     0       R1      131567    cellular organisms
100.00  119     0       D       2           Bacteria
100.00  119     0       K       3379134       Pseudomonadati
100.00  119     0       P       1224            Pseudomonadota
100.00  119     0       C       28216             Betaproteobacteria
100.00  119     0       O       206351              Neisseriales
100.00  119     0       F       481                   Neisseriaceae
100.00  119     0       G       482                     Neisseria
100.00  119     116     S       485                       Neisseria gonorrhoeae
(kraken2)  bbeominfo@ipsec-10-2-64-221  ~/test  main
```



```
# Step 1: Download the pre-built 8GB Kraken2 database
wget https://genome-idx.s3.amazonaws.com/kraken/k2_standard_8gb_20240306.tar.gz

# Step 2: Extract the downloaded database
tar -xvzf k2_standard_8gb_20240306.tar.gz

# Step 3: Create Conda environment for Kraken2 (MacOS-specific architecture forced)
CONDA_SUBDIR=osx-64 conda create -n kraken2 -y

# Step 4: Activate Conda environment
conda activate kraken2

# Step 5: Install Kraken2 using Bioconda and Conda-Forge channels
conda install -c bioconda -c conda-forge kraken2

# Step 6: Run Kraken2 classification for sample B1299860
kraken2 \
--db ./k2_standard_08gb_20240306 \
--threads 8 \
--output B1299860.kraken2.output \
--report B1299860.kraken2.report \
B1299860_S01_L001_contigs.fasta

# Step 7: (Optional) Check number of contigs in the FASTA file
grep '>' B1299860_S01_L001_contigs.fasta | wc -l

# Step 8: Run Kraken2 classification for sample B1838859 (second sample)
kraken2 \
--db ./k2_standard_08gb_20240306 \
--threads 8 \
--output B1838859.kraken2.output \
--report B1838859.kraken2.report \
B1838859_S01_L001_contigs.fa
```