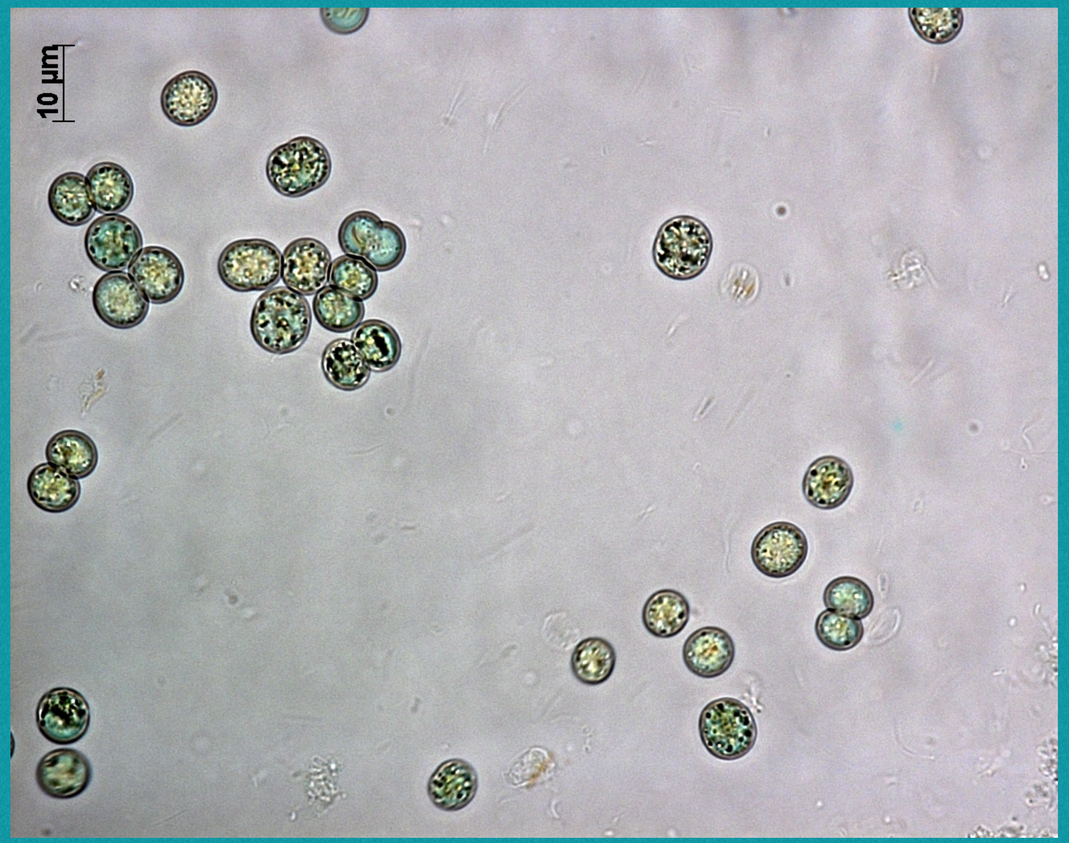
UNIVERSITY OF HELSINKI
FACULTY OF AGRICULTURE AND FORESTRY

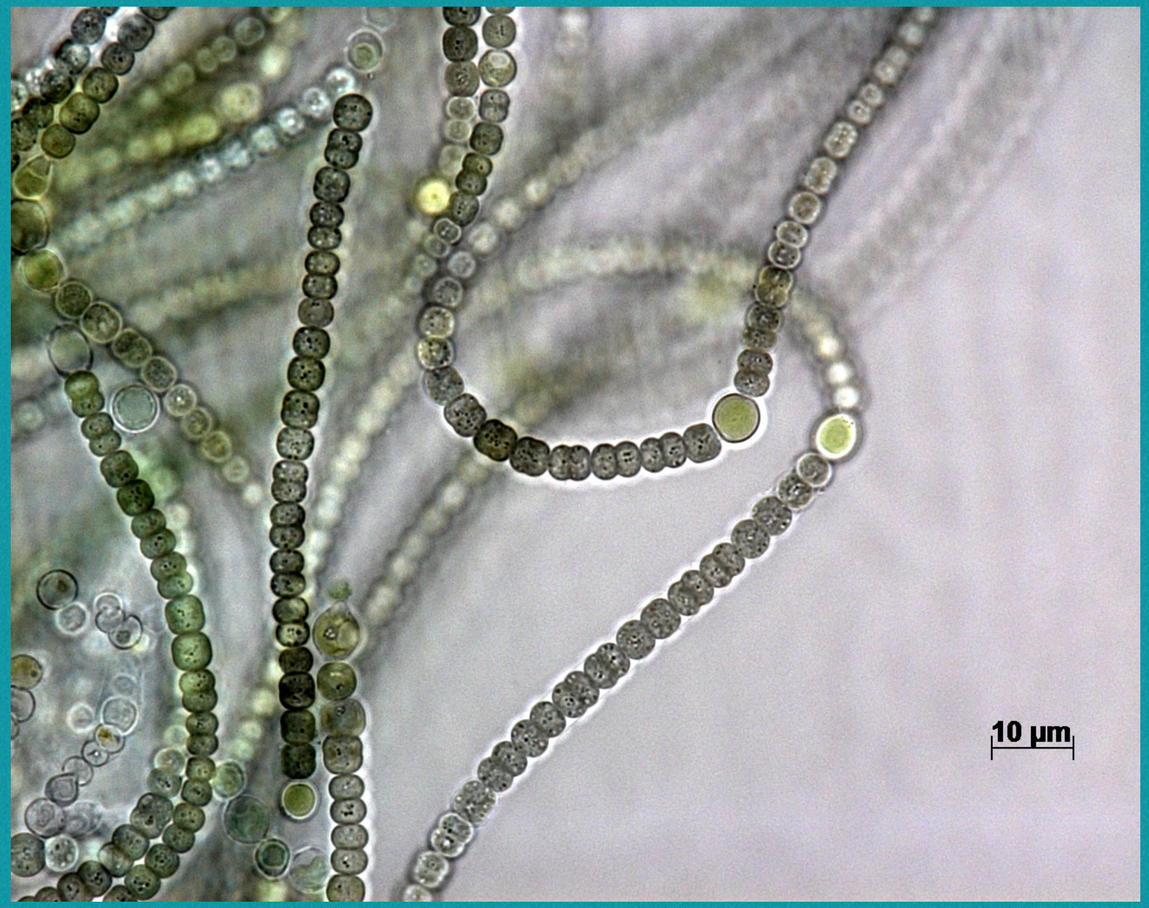
Assembly quality control

Tania Keiko Shishido Joutsen
tania.shishido@helsinki.fi

MBDP105 29.3.2022



Assembly quality control



Goals:

1. To learn how to “clean” your sample’s genomes
2. To know which tools can be used for quality control of assembled genomes
3. To be able to perform quality control on assembled genomes

DNA extraction

Sequencing

Where are we?

Illumina reads

QC

Trimmed Illumina reads

+

Nanopore reads

QC

Trimmed Nanopore reads

Spades

Hybrid assembly

Kaiju

Only cyanobacteria assembly

QC

Genome ready

Removal of contaminant contigs

Kaiju

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Kaiju

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "Kaiju" – news · newspapers · books · scholar · JSTOR *(July 2021)*
(Learn how and when to remove this template message)

Kaiju (Japanese: 怪獣, Hepburn: *Kaijū*, lit. 'Strange Beast') is a Japanese genre of films and television featuring giant monsters. The term *kaiju* can also refer to the giant monsters themselves, which are usually depicted attacking major cities and battling either the military or other monsters. The *kaiju* genre is a subgenre of *tokusatsu* (特撮, "special filming") entertainment.

The 1954 film *Godzilla* is commonly regarded as the first *kaiju* film. *Kaiju* characters are often somewhat metaphorical in nature; Godzilla, for example, serves as a metaphor for nuclear weapons, reflecting the fears of post-war Japan following the atomic bombings of Hiroshima and Nagasaki and the *Lucky Dragon 5* incident. Other notable examples of *kaiju* characters include Rodan, Mothra, King Ghidorah, and Gamera.

Contents [hide]
1 Origins
2 Terminology
2.1 *Kaijū eiga*

The *kaiju* Godzilla from the 1954 film *Godzilla*, one of the first Japanese films to feature a giant monster

kaiju.binf.ku.dk

Home Web Server Source

KAIJU

Fast and sensitive taxonomic classification for metagenomics

About

Kaiju is a program for sensitive taxonomic classification of high-throughput sequencing reads from metagenomic whole genome sequencing or metatranscriptomics experiments.

Each sequencing read is assigned to a taxon in the NCBI taxonomy by comparing it to a reference database containing microbial and viral protein sequences. By using protein-level classification, Kaiju achieves a higher sensitivity compared with methods based on nucleotide comparison.

Kaiju can use either the set of available complete genomes from NCBI RefSeq or the microbial subset of the NCBI BLAST non-redundant protein database *nr*, optionally also including fungi and microbial eukaryotes.

Reads are translated into amino acid sequences, which are then searched in the database using a modified backward search on a memory-efficient implementation of the Burrows-Wheeler transform, which finds maximum exact matches (MEMs), optionally allowing

https://kaiju.binf.ku.dk/

Menzel, P. et al. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7:11257

Removal of contaminant contigs

Kaiju

github.com/bioinformatics-centre/kaiju

README.md

Classification accuracy

The accuracy of the classification depends both on the choice of the reference database and the chosen options when running Kaiju. These choices also affect the speed and memory usage of Kaiju.

For highest sensitivity, it is recommended to use the *nr* database (+eukaryotes) as a reference database because it is the most comprehensive set of protein sequences. Alternatively, use proGenomes over Refseq for increased sensitivity.

Greedy run mode yields a higher sensitivity compared with MEM mode.

For fastest classification, use MEM mode and multiple parallel threads (`-z`); and for lowest memory usage use the proGenomes reference database. The number of parallel threads has only little impact on memory usage.

Further, the choice of the minimum required match length (`-m`) in MEM mode or match score (`-s`) in Greedy mode governs the trade-off between sensitivity and precision of the classification. Please refer to the paper for a discussion on this topic.

Helper programs

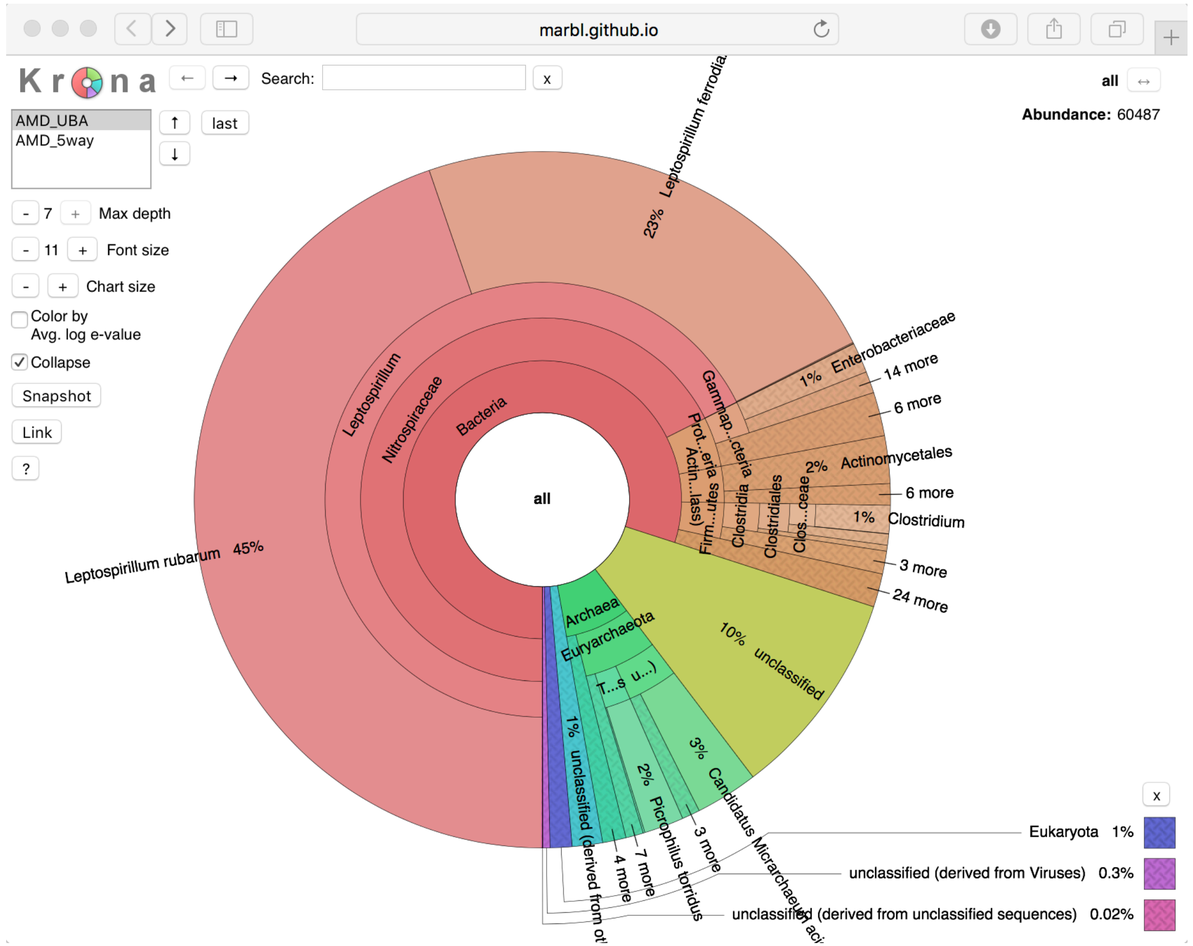
Creating input file for Krona

The program `kaiju2krona` can be used to convert Kaiju's tab-separated output file into a tab-separated text file, which can be imported into Krona. It requires the `nodes.dmp` and `names.dmp` files from the NCBI taxonomy for mapping the taxon identifiers from Kaiju's output to the corresponding taxon names.

```
kaiju2krona -t nodes.dmp -n names.dmp -i kaiju.out -o kaiju.out.krona
```

The file `kaiju.out.krona` can then be imported into Krona and converted into an HTML file using Krona's `ktImportText` program:

https://kaiju.binf.ku.dk/



https://github.com/marbl/Krona/wiki

DNA extraction

Sequencing

Where are we?

Illumina reads

QC

Trimmed Illumina reads

+

Nanopore reads

QC

Trimmed Nanopore reads

Spades

Hybrid assembly

Kaiju

Only cyanobacteria assembly

QC

Genome ready

Examples

TABLE 1 | Genome features of *Pantanalinema* GBBB05.

Features	Chromosome
Strain	*Pantanalinema* GBBB05
Number of contigs	94
L50 value	17
N50 value (bp)	142,797
Completeness	99.05%
Contamination	0.4%
Sequencing coverage	18x
GC content	48.43%
Estimated chromosome size (bp)	7,181,771
Protein-coding genes (CDS)[1]	5,976
rRNAs[1]	6
5S rRNAs	2
16S rRNAs	2
23S rRNA	2
tRNAs	106
ncRNAs	4
Pseudo genes	152
CRISPR[2*]	16 sequences
CAS[2*]	1 sequence
Phage[3*]	1

*Genome statistics were obtained through CheckM. [1]Prokaryotic Genome Annotation Pipeline (PGAP); [2]CRISPRCasFinder; [3]PHAST; *Complete results from analysis with CRISPRCasFinder and PHAST are shown in **Supplementary Tables 2** and **3**.*

Ferreira et al 2021 Frontiers in Ecology and Evolution
https://doi.org/10.3389/fevo.2021.639852

kit v.2. For PacBio sequencing, DNA was isolated using the Genomic-tip 100/G kit (Qiagen), libraries were prepared using the standard PacBio 20-kb protocol, and fragments were size selected (>10 kb) with BluePippin (Sage Science) and sequenced on a PacBio RS II system in one single-molecule real-time (SMRT) cell, using P6-C4 chemistry (6-h movie). Reads (50,981 reads; N_{50}, 20,257 bp) were filtered (>750 bp) and assembled using HGAP.3 (seed cutoff, 6 kb). The consensus sequence was polished by additional rounds of PacBio read mapping and was circularized using information from the bridge mapper tool, all within the SMRT Analysis software (v.2.3.0.140936), using default settings. MiSeq reads were mapped to the initial PacBio assembly using Geneious v.10 with default settings and used for additional quality control and manual correction of indel errors. Coverages were 45× and 175× for the MiSeq and PacBio reads, respectively. A single circular 2,630,292-bp assembly with a G+C content of 63.3% was obtained. The genome was initially annotated using RASTtk (9) and subsequently updated with the NCBI Prokaryotic Genome Annotation Pipeline (NCBI RefSeq database). The genome includes 2,693 protein-coding genes, 41 pseudogenes, 6 rRNAs, and 43 tRNAs.

Genes for viral resistance are often localized to genomic islands (hypervariable regions) in *Synechococcus* and *Prochlorococcus* spp. (4, 10). Using previously established criteria (10, 11), 13 genomic islands were identified in WH 8101 (Table 1). These regions were >8 kb and/or contained at least 10 genes that were not in synteny with the genome of the other clade VIII strain, *Synechococcus* sp. strain RS9917. Genomic islands that were identified in RS9917 (11) and present in WH 8101 were also included. This genomic sequence will be used to identify genetic determinants of cyanophage resistance.

Marston and Polson 2020 Microbiology Resource Announcements
https://doi.org/10.1128/MRA.01593-19

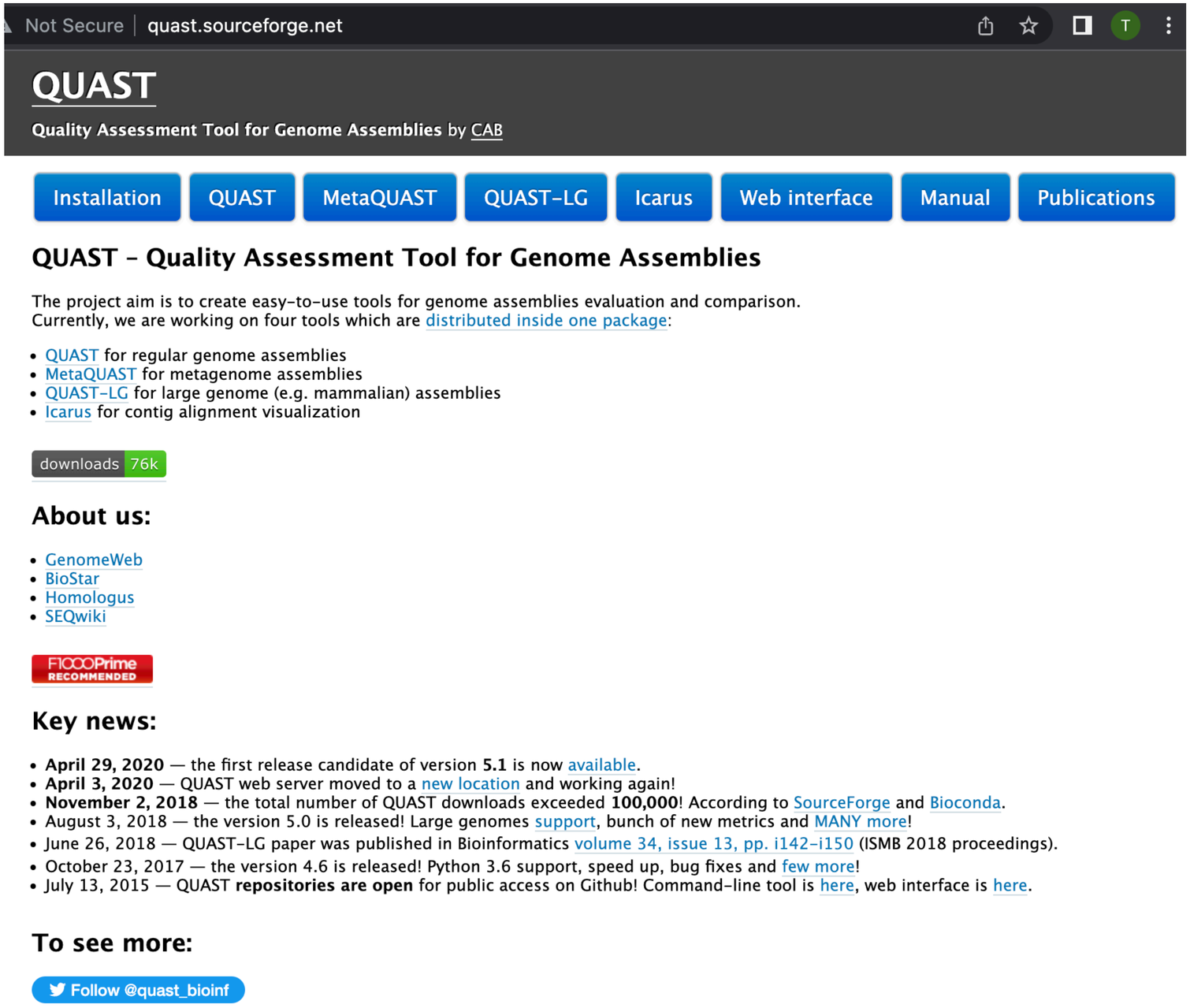
Attribute	Value	%
Genome size (bp)	7,502,480	100.0
DNA coding (bp)	6,017,946	80.2
DNA G + C (bp)	3,420,381	45.6
DNA contigs	296	—
Contigs N50	37,607	—
Longest contig	132,835	—
Total genes	7,648	100.0
Protein coding genes	6,349	83.0
RNA genes	68	0.9
Genes with function prediction	3,521	46.0
Genes assigned to COGs	5,349	69.9
Genes with signal peptides	636	8.3
Genes with transmembrane helices	1,280	16.7
CRISPR repeats	6	—

Table 1. Genome statistics of *Microcoleus asticus* sp. nov. Quality assessment and level of completeness of the genome assembly of *Microcoleus asticus* sp. nov. COGs - Clusters of Orthologous Groups of proteins, CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats.

Churro et al 2020 Scientific Reports
https://doi.org/10.1038/s41598-020-64774-y

Genomic statistics

QUAST



http://cab.cc.spbu.ru/quast/

Gurevich et al., 2013 Bioinformatics
doi: 10.1093/bioinformatics/btt086

Genomic statistics

Different assemblies

QUAST

Quality color coded

All statistics are based on contigs of size >= [illegible]bp, unless o[illegible]wise not[illegible].g., "# contig[illegible]= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best Show heat[illegible]

Statistics without reference	IDBA_UD	SOAPdenovo	SPAdes	Velvet
# contigs	111	181	93	120
# contigs (>= 0 bp)	126	930	152	192
# contigs (>= 1000 bp)	97	168	80	107
# contigs (>= 5000 bp)	71	127	56	82
# contigs (>= 10000 bp)	65	105	52	74
# contigs (>= 25000 bp)	49	62	45	57
# contigs (>= 50000 bp)	29	26	31	33
Largest contig	221 687	165 487	285 114	242 032
Total length	4 566 224	4 535 469	4 558 330	4 554 702
Total length (>= 0 bp)	4 571 021	4 614 535	4 570 605	4 569 214
Total length (>= 1000 bp)	4 557 071	4 526 969	4 549 301	4 546 082
Total length (>= 5000 bp)	4 503 399	4 429 815	4 496 699	4 492 466
Total length (>= 10000 bp)	4 456 362	4 256 422	4 467 005	4 432 370
Total length (>= 25000 bp)	4 213 787	3 571 206	4 364 167	4 147 479
Total length (>= 50000 bp)	3 485 235	2 277 663	3 878 287	3 287 286
N50	111 794	52 524	132 831	82 776
N75	56 778	29 555	67 340	42 907
L50	14	26	13	18
L75	28	56	24	36
GC (%)	50.75	50.74	50.74	50.74
Mismatches				
# N's	0	0	0	0
# N's per 100 kbp	0	0	0	0
Predicted genes				
# predicted genes (unique)	3601	3582	3588	3592
# predicted genes (>= 0 bp)	3593 + 8 part	3568 + 14 part	3579 + 9 part	3572 + 20 part
# predicted genes (>= 300 bp)	3374 + 8 part	3358 + 13 part	3364 + 8 part	3362 + 19 part
# predicted genes (>= 1500 bp)	657 + 1 part	651 + 3 part	660 + 1 part	655 + 2 part
# predicted genes (>= 3000 bp)	81 + 1 part	77 + 1 part	83 + 1 part	81 + 0 part

Genomic statistics

QUAST

Contigs

Size (bp)

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

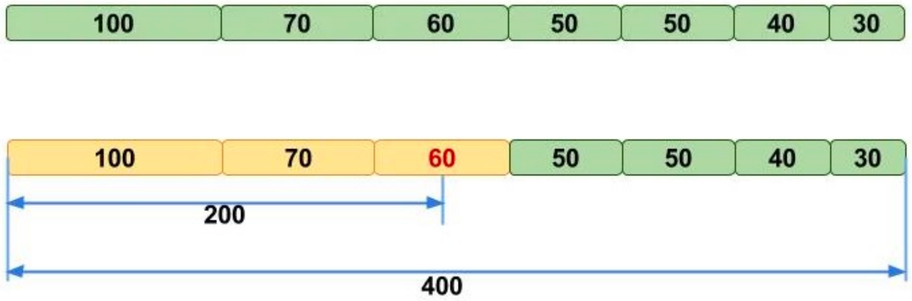
Show heatmap

Worst Median Best

Statistics without reference	IDBA_UD	SOAPdenovo	SPAdes	Velvet
# contigs	111	181	93	120
# contigs (>= 1000 bp)	97	168	80	107
# contigs (>= 5000 bp)	71	127	56	82
# contigs (>= 10000 bp)	65	105	52	74
# contigs (>= 25000 bp)	49	62	45	57
# contigs (>= 50000 bp)	29	26	31	33
Largest contig	221 687	165 487	285 114	242 032
Total length	4 566 224	4 535 469	4 558 330	4 554 702
Total length (>= 0 bp)	4 571 021	4 614 535	4 570 605	4 569 214
Total length (>= 1000 bp)	4 557 071	4 526 969	4 549 301	4 546 082
Total length (>= 5000 bp)	4 503 399	4 429 815	4 496 699	4 492 466
Total length (>= 10000 bp)	4 456 362	4 256 422	4 467 005	4 432 370
Total length (>= 25000 bp)	4 213 787	3 571 206	4 364 167	4 147 479
Total length (>= 50000 bp)	3 485 235	2 277 663	3 878 287	3 287 286
N50	111 794	52 524	132 831	82 776
N75	56 778	29 555	67 340	42 907
L50	14	26	13	18
L75	28	56	24	36
GC (%)	50.75	50.74	50.74	50.74
Mismatches				
# N's	0	0	0	0
# N's per 100 kbp	0	0	0	0
Predicted genes				
# predicted genes (unique)	3601	3582	3588	3592
# predicted genes (>= 0 bp)	3593 + 8 part	3568 + 14 part	3579 + 9 part	3572 + 20 part
# predicted genes (>= 300 bp)	3374 + 8 part	3358 + 13 part	3364 + 8 part	3362 + 19 part
# predicted genes (>= 1500 bp)	657 + 1 part	651 + 3 part	660 + 1 part	655 + 2 part
# predicted genes (>= 3000 bp)	81 + 1 part	77 + 1 part	83 + 1 part	81 + 0 part

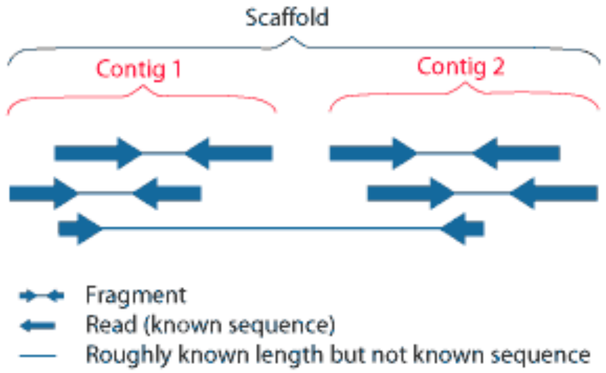
Genomic statistics

N50 (bp): distribution of the contigs length



https://www.molecularecologist.com/2017/03/29/whats-n50/

Another way of thinking about it, is that at least half of the nucleotides in the assembly belongs to contigs with the N50 length or longer. And if you have scaffolds, you can calculate the **scaffold N50** in addition to your **contig N50**.



https://mycocosm.jgi.doe.gov/help/scaffolds.jsf

Genomic statistics

QUAST

N50 (bp): distribution of the contigs length

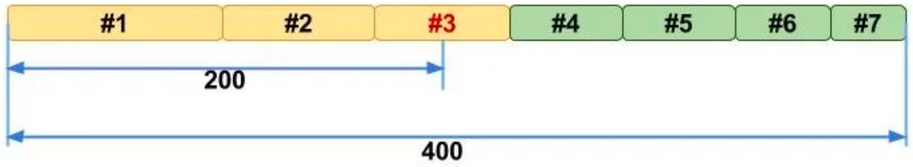
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	IDBA_UD	SOAPdenovo	SPAdes	Velvet
# contigs	111	181	93	120
# contigs (>= 0 bp)	126	930	152	192
# contigs (>= 1000 bp)	97	168	80	107
# contigs (>= 5000 bp)	71	127	56	82
# contigs (>= 10000 bp)	65	105	52	74
# contigs (>= 25000 bp)	49	62	45	57
# contigs (>= 50000 bp)	29	26	31	33
Largest contig	221 687	165 487	285 114	242 032
Total length	4 566 224	4 535 469	4 558 330	4 554 702
Total length (>= 0 bp)	4 571 021	4 614 535	4 570 605	4 569 214
Total length (>= 1000 bp)	4 557 071	4 526 969	4 549 301	4 546 082
Total length (>= 5000 bp)	4 503 399	4 429 815	4 496 699	4 492 466
Total length (>= 10000 bp)	4 456 362	4 256 422	4 467 005	4 432 370
Total length (>= 25000 bp)	4 213 787	3 571 206	4 364 167	4 147 479
Total length (>= 50000 bp)	3 485 235	2 277 663	3 878 287	3 287 286
N50	111 794	52 524	132 831	82 776
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
L50	14	26	13	18
L75	28	56	24	36
GC (%)	50.75	50.74	50.74	50.74
Mismatches				
# N's	0	0	0	0
# N's per 100 kbp	0	0	0	0
Predicted genes				
# predicted genes (unique)	3601	3582	3588	3592
# predicted genes (>= 0 bp)	3593 + 8 part	3568 + 14 part	3579 + 9 part	3572 + 20 part
# predicted genes (>= 300 bp)	3374 + 8 part	3358 + 13 part	3364 + 8 part	3362 + 19 part
# predicted genes (>= 1500 bp)	657 + 1 part	651 + 3 part	660 + 1 part	655 + 2 part
# predicted genes (>= 3000 bp)	81 + 1 part	77 + 1 part	83 + 1 part	81 + 0 part

Genomic statistics

L50: number of sequences



https://www.molecularecologist.com/2017/03/29/whats-n50/

N50 - distribution of the contigs length - bp

L50 - number of sequences

Genomic statistics

QUAST

L50: number of sequences

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	IDBA_UD	SOAPdenovo	SPAdes	Velvet
# contigs	111	181	93	120
# contigs (>= 0 bp)	126	930	152	192
# contigs (>= 1000 bp)	97	168	80	107
# contigs (>= 5000 bp)	71	127	56	82
# contigs (>= 10000 bp)	65	105	52	74
# contigs (>= 25000 bp)	49	62	45	57
# contigs (>= 50000 bp)	29	26	31	33
Largest contig	221 687	165 487	285 114	242 032
Total length	4 566 224	4 535 469	4 558 330	4 554 702
Total length (>= 0 bp)	4 571 021	4 614 535	4 570 605	4 569 214
Total length (>= 1000 bp)	4 557 071	4 526 969	4 549 301	4 546 082
Total length (>= 5000 bp)	4 503 399	4 429 815	4 496 699	4 492 466
Total length (>= 10000 bp)	4 456 362	4 256 422	4 467 005	4 432 370
Total length (>= 25000 bp)	4 213 787	3 571 206	4 364 167	4 147 479
Total length (>= 50000 bp)	3 485 235	2 277 663	3 878 287	3 287 286
N50	111 794	52 524	132 831	82 776
N75	56 778	29 555	67 340	42 907
L50	14	26	13	18
L75	28	56	24	36
GC (%)	50.75	50.74	50.74	50.74
Mismatches				
# N's	0	0	0	0
# N's per 100 kbp	0	0	0	0
Predicted genes				
# predicted genes (unique)	3601	3582	3588	3592
# predicted genes (>= 0 bp)	3593 + 8 part	3568 + 14 part	3579 + 9 part	3572 + 20 part
# predicted genes (>= 300 bp)	3374 + 8 part	3358 + 13 part	3364 + 8 part	3362 + 19 part
# predicted genes (>= 1500 bp)	657 + 1 part	651 + 3 part	660 + 1 part	655 + 2 part
# predicted genes (>= 3000 bp)	81 + 1 part	77 + 1 part	83 + 1 part	81 + 0 part

Genomic statistics

GC (%)

- *GC (%)*: The total number of G and C nucleotides in the assembly, divided by the total length of the assembly. This metric can be computed without a reference genome.

Table 1. Overview of genome features in the cyanobiont ('*Nostoc azollae*' 0708) of the water fern *Azolla filiculoides* Lam.

Feature	Chlp*	**NoAz**	Noss+	Nosp+
Symbiotic competence	Obligate	**Obligate**	None	Facultative
Genome size (bp)	154,478	**5,486,145**	7,211,789	9,059,191
Plasmids	0	**2**	6	5
Coding nucleotide proportion %	51	**52**	82	77
GC content %	36	**38**	41	41
Genes, total number	129	**5413**	6222	6791
Coding sequences	85	**3668**	6,130	6,690
Pseudogenes (%)	0	**1689 (31.2)**	0	0
rRNA	7	**12**	12	12
tRNA	37	**44**	70	88

For comparative purposes the genomes of a chloroplast (Arabidopsis) and genomes of two related cyanobacteria (Section IV), one being a facultative plant symbiont and the other a free-living species, are given. Chlp = Chloroplast of *Arabidopsis thaliana*, NoAz = '*Nostoc azollae*' 0708, Noss = *Nostoc* sp. PCC 7120, Nosp = *Nostoc punctiforme* PCC 73102.
*Data from NCBI database (http://www.ncbi.nlm.nih.gov/).
+Data from IMG database (http://img.jgi.doe.gov/).
doi:10.1371/journal.pone.0011486.t001

Ran et al 2010 PLoS ONE

Order	Genus	No. gen	Pld/gen	Size (Mb)	GC (%)	Genes	No. BGCs Chr	No. BGCs Pld	Total BGC
Gloeobacterales	*Gloeobacter*	2	0	4.69 ± 0.04	61.3 ± 1.1	4,497 ± 21	5 ± 3	0	7 ± 1
Synechococcales	*Cyanobium*	2	0	3.18 ± 0.23	68.7 ± 0.1	3,227 ± 251	8 ± 1	0	8 ± 1
	Leptolyngbya	6	2 ± 2	6.37 ± 0.85	47.9 ± 4.1	5,890 ± 965	10 ± 4	1 ± 1	11 ± 5
	Prochlorococcus	17	1 ± 0	1.83 ± 0.29	34.8 ± 6.3	2,018 ± 327	6 ± 5	1 ± 0	6 ± 5
	Pseudanabaena	2	4 ± 4	5.28 ± 0.54	44.2 ± 2.8	4,510 ± 789	3 ± 1	1 ± 0	4 ± 3
	Synechococcus	28	0 ± 1	2.79 ± 0.57	57.2 ± 5.4	2,847 ± 452	6 ± 4	0	5 ± 4
	Synechocystis	8	2 ± 3	3.72 ± 0.18	47.6 ± 0.2	3,465 ± 200	3 ± 0	0	3 ± 0
	Thermosynechococcus	5	0	2.55 ± 0.06	53.7 ± 0.3	2,543 ± 67	3 ± 0	0	3 ± 0
Oscillatoriales	*Arthrospira*	3	0	6.47 ± 0.35	44.4 ± 0.3	8,118 ± 3,081	3 ± 1	0	3 ± 1
	Moorea	2	2 ± 1	9.55 ± 0.23	43.6 ± 0.1	7,748 ± 31	42 ± 0	0	42 ± 0
	Oscillatoria	2	4 ± 2	8.04 ± 0.33	46.7 ± 1.3	6,479 ± 590	8 ± 1	1 ± 0	9 ± 2
	Planktothrix	2	5 ± 1	5.07 ± 0.02	39.6 ± 0.1	4,538 ± 13	6 ± 4	2 ± 0	11 ± 1
	Synechococcus	7	5 ± 1	3.14 ± 0.12	49.2 ± 0.1	3,183 ± 117	3 ± 0	1 ± 1	4 ± 1
Chroococcales	*Cyanobacterium*	2	1 ± 1	3.23 ± 0.10	38.2 ± 0.7	3,020 ± 251	4 ± 1	0	4 ± 1
	Gloeothece	2	6 ± 0	7.20 ± 0.91	39.2 ± 1	6,406 ± 795	8 ± 1	5 ± 4	13 ± 2
	Geminocystis	3	8 ± 4	4.24 ± 0.20	33.3 ± 1.0	3,831 ± 270	4 ± 1	0	4 ± 1
	Microcystis	7	0	5.19 ± 0.65	42.5 ± 0.3	5,243 ± 752	10 ± 1	0	10 ± 1
	Rippkaea	2	4 ± 1	4.80 ± 0.01	39.8 ± 0.0	4,540 ± 30	8 ± 0	0	8 ± 0
Pleurocapsales	*Stanieria*	2	3 ± 3	5.50 ± 0.06	36.4 ± 0.2	4,948 ± 30	11 ± 2	1 ± 0	12 ± 3
Nostocales	*Anabaena*	5	3 ± 3	6.44 ± 0.83	39.1 ± 1.3	5,713 ± 558	14 ± 3	1 ± 1	14 ± 3
	Calothrix	11	4 ± 3	8.70 ± 2.11	40.2 ± 1.5	7,273 ± 1,957	15 ± 6	1 ± 0	15 ± 6
	Cylindrospermum	2	4 ± 1	7.66 ± 0.06	42.1 ± ± 0.1	6,574 ± 65	21 ± 3	3 ± 2	24 ± 1
	Dolichospermum	2	2 ± 2	5.41 ± 0.33	38.2 ± 0.0	5,032 ± 349	11 ± 4	0	11 ± 4
	Fischerella	3	5 ± 4	6.54 ± 1.00	40.5 ± 0.7	5,547 ± 899	13 ± 1	5 ± 0	15 ± 3
	Nodularia	2	1 ± 1	5.43 ± 0.05	41.2 ± 0	4,866 ± 42	11 ± 1	0	11 ± 1
	Nostoc	20	5 ± 3	7.88 ± 1.22	41 ± 0.8	6,824 ± 1,096	16 ± 5	2 ± 2	18 ± 6
	Scytonema	2	6 ± 2	9.81 ± 0.06	43.6 ± 0.2	8,176 ± 76	25 ± 1	2 ± 1	27 ± 2
	Trichormus	2	4 ± 1	7.29 ± 0.25	41.4 ± 0	6,147 ± 291	14 ± 1	2 ± 1	15 ± 1

Averages and standard deviations of genome size, GC content, number of genes, BGCs in the chromosomes and plasmids, and the total number of BGCs in the genom were calculated.
Gen, genome; Pld, plasmid; BGC, biosynthetic gene cluster; Chr, chromosome.

Popin et al 2021, Front. Microbiol.

Genomic statistics

GC (%)

QUAST

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

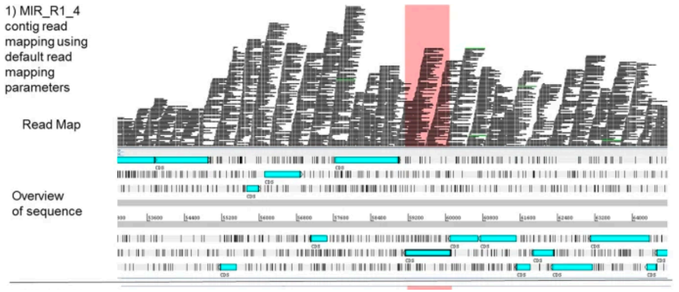
Worst Median Best

☑ Show heatmap

Statistics without reference	IDBA_UD	SOAPdenovo	SPAdes	Velvet
# contigs	111	181	93	120
# contigs (>= 0 bp)	126	930	152	192
# contigs (>= 1000 bp)	97	168	80	107
# contigs (>= 5000 bp)	71	127	56	82
# contigs (>= 10000 bp)	65	105	52	74
# contigs (>= 25000 bp)	49	62	45	57
# contigs (>= 50000 bp)	29	26	31	33
Largest contig	221 687	165 487	285 114	242 032
Total length	4 566 224	4 535 469	4 558 330	4 554 702
Total length (>= 0 bp)	4 571 021	4 614 535	4 570 605	4 569 214
Total length (>= 1000 bp)	4 557 071	4 526 969	4 549 301	4 546 082
Total length (>= 5000 bp)	4 503 399	4 429 815	4 496 699	4 492 466
Total length (>= 10000 bp)	4 456 362	4 256 422	4 467 005	4 432 370
Total length (>= 25000 bp)	4 213 787	3 571 206	4 364 167	4 147 479
Total length (>= 50000 bp)	3 485 235	2 277 663	3 878 287	3 287 286
N50	111 794	52 524	132 831	82 776
N75	56 778	29 555	67 340	42 907
L50	14	26	13	18
L75	28	56	24	36
GC (%)	50.75	50.74	50.74	50.74
Mismatches				
# N's	0	0	0	0
# N's per 100 kbp	0	0	0	0
Predicted genes				
# predicted genes (unique)	3601	3582	3588	3592
# predicted genes (>= 0 bp)	3593 + 8 part	3568 + 14 part	3579 + 9 part	3572 + 20 part
# predicted genes (>= 300 bp)	3374 + 8 part	3358 + 13 part	3364 + 8 part	3362 + 19 part
# predicted genes (>= 1500 bp)	657 + 1 part	651 + 3 part	660 + 1 part	655 + 2 part
# predicted genes (>= 3000 bp)	81 + 1 part	77 + 1 part	83 + 1 part	81 + 0 part

Genomic statistics

Genome coverage



Adapted from Lehri et al 2017 Scientific Reports

Bowtie2 → **Samtools** → **Bedtools**

Map the reads

Bowtie 2 Papers

- Langmead B, Wilks C., Antonescu V., Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. bty648.
- Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.
-

To sort and make an index for the mapped reads

Twelve years of SAMtools and BCFtools Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li *GigaScience*, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008

Calculate the coverage

- Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

Genomic statistics

CheckM

Method

CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

Donovan H. Parks,[1] Michael Imelfort,[1] Connor T. Skennerton,[1] Philip Hugenholtz,[1,2] and Gene W. Tyson[1,3]

[1]*Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia;* [2]*Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia;* [3]*Advanced Water Management Centre, The University of Queensland, St. Lucia, QLD 4072, Queensland, Australia*

Large-scale recovery of genomes from isolates, single cells, and metagenomic data has been made possible by advances in computational methods and substantial reductions in sequencing costs. Although this increasing breadth of draft genomes is providing key information regarding the evolutionary and functional diversity of microbial life, it has become impractical to finish all available reference genomes. Making robust biological inferences from draft genomes requires accurate estimates of their completeness and contamination. Current methods for assessing genome quality are ad hoc and generally make use of a limited number of "marker" genes conserved across all bacterial or archaeal genomes. Here we introduce CheckM, an automated method for assessing the quality of a genome using a broader set of marker genes specific to the position of a genome within a reference genome tree and information about the collocation of these genes. We demonstrate the effectiveness of CheckM using synthetic data and a wide range of isolate-, single-cell-, and metagenome-derived genomes. CheckM is shown to provide accurate estimates of genome completeness and contamination and to outperform existing approaches. Using CheckM, we identify a diverse range of errors currently impacting publicly available isolate genomes and demonstrate that genomes obtained from single cells and metagenomic data vary substantially in quality. In order to facilitate the use of draft genomes, we propose an objective measure of genome quality that can be used to select genomes suitable for specific gene- and genome-centric analyses of microbial communities.

ecogenomics.github.io/CheckM/

CheckM

Overview

CheckM provides a set of tools for assessing the quality of genomes recovered from isolates, single cells, or metagenomes. It provides robust estimates of genome completeness and contamination by using collocated sets of genes that are ubiquitous and single-copy within a phylogenetic lineage. Assessment of genome quality can also be examined using plots depicting key genomic characteristics (e.g., GC, coding density) which highlight sequences outside the expected distributions of a typical genome. CheckM also provides tools for identifying genome bins that are likely candidates for merging based on marker set compatibility, similarity in genomic characteristics, and proximity within a reference genome tree.

News

- CheckM v1.1.3 was released on July 9, 2020 and requires Python 3.

Use CheckM

Before using CheckM you need a set of putative genomes. These may come from isolates, single cells, or metagenomic data. Our companion tool GroopM can be used to recover genomes from metagenomic data.

For information on using CheckM visit the wiki.

Cite CheckM

If you find this software useful, we'd love for you to cite us:

- Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2014. Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25: 1043-1055.

Talk to us

We'd love to hear from you. All comments and suggestions can be sent to Donovan Parks:

- donovan_dot_parks_at_gmail_dot_com

Licensing

CheckM is licensed using the GNU General Public License version 3 as published by the Free Software Foundation.

The CheckM logo is a product of Mike Imelfort's mind.

This site was created using a template created by the wonderful people at bootswatch.

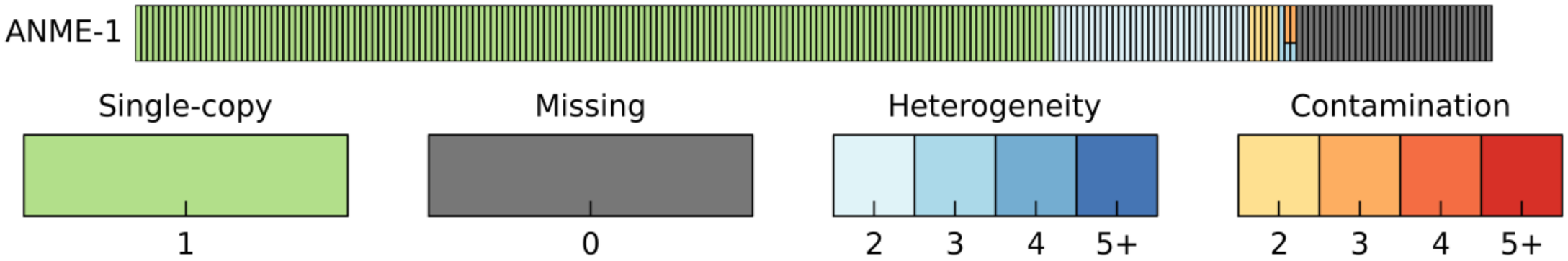
https://github.com/Ecogenomics/CheckM/wiki/Introduction#about

Genomic statistics

CheckM

Completeness (%) and Contamination (%)

Marker genes that are specific to a genome's inferred lineage within a reference genome tree



Supplemental Figure S13. Identification of the 213 marker genes within the Meyerdierks et al. (2010) ANME-1 genome. Each bar represents a marker gene. Bars in green represent markers identified exactly once, while bars in grey represent missing markers. **Markers identified multiple times in the genome are represented by shades of blue or red** depending on the AAI between pairs of multi-copy genes and the total number of copies present (2-5+). Pairs of **multi-copy genes with an AAI ≥90% are indicated with shades of blue,** while genes with **less amino acid similarity are shown in red**. A gene present 3 or more times may have pairs with an AAI ≥90% and pairs with an AAI < 90%. Plot produced with CheckM.

(Parks et al 2015)

AAI - amino acid identity

Genomic statistics

CheckM

Completeness (%) and Contamination (%)

CheckM_summary_table

Bin Name	Marker Lineage	# Genomes	# Markers	# Marker Sets	0	1	2	3	4	5	Completeness	Contamination
Test	p__Cyanobacteria	79	583	457	1	579	3	0	0	0	99.78	0.51

TABLE 1 | Genome features of *Pantanalinema* GBBB05.

Features	Chromosome
Strain	*Pantanalinema* GBBB05
Number of contigs	94
L50 value	17
N50 value (bp)	142,797
Completeness	99.05%
Contamination	0.4%
Sequencing coverage	18x
GC content	48.43%
Estimated chromosome size (bp)	7,181,771
Protein-coding genes (CDS)[1]	5,976
rRNAs[1]	6
5S rRNAs	2
16S rRNAs	2
23S rRNA	2
tRNAs	106
ncRNAs	4
Pseudo genes	152
CRISPR[2*]	16 sequences
CAS[2*]	1 sequence
Phage[3*]	1

Genome statistics were obtained through CheckM. [1]*Prokaryotic Genome Annotation Pipeline (PGAP);* [2]*CRISPRCasFinder;* [3]*PHAST; *Complete results from analysis with CRISPRCasFinder and PHAST are shown in* ***Supplementary Tables 2*** *and* ***3****.*

Ferreira et al 2021 Frontiers in Ecology and Evolution
https://doi.org/10.3389/fevo.2021.639852

Further

PROKKA

https://github.com/tseemann/prokka

PHASTER

https://phaster.ca/

CRISPRone

https://omics.informatics.indiana.edu/CRISPRone/denovo.php

DNA extraction

Sequencing

Where are we?

Illumina reads

QC

Trimmed Illumina reads

Nanopore reads

QC

Trimmed Nanopore reads

+

Spades

Hybrid assembly

Kaiju

Only cyanobacteria assembly

QC

Genome ready

Hands on:
Eliminate contaminant contigs
Assembly QC
Calculate the genome coverage
Genome completeness and contamination