



How to Run MetaBAT



Workflow

- 1. Run MetaBAT and get bins
- 2. Use CheckM to evaluate those bins
-
-



dataset

- CAMI Low Complexity
- 19499 contigs
- Average 7972
- Standard deviation 45485
- Max 1684314
- Min 150



Install on Proteus

- Copy file 'metabat2' to Proteus
- or
- `$ bash install_on_proteus.sh`
- This will create a 'MetaBAT' folder in home directory.



Install on other environments











- <https://bitbucket.org/berkeleylab/metabat/src/master/>



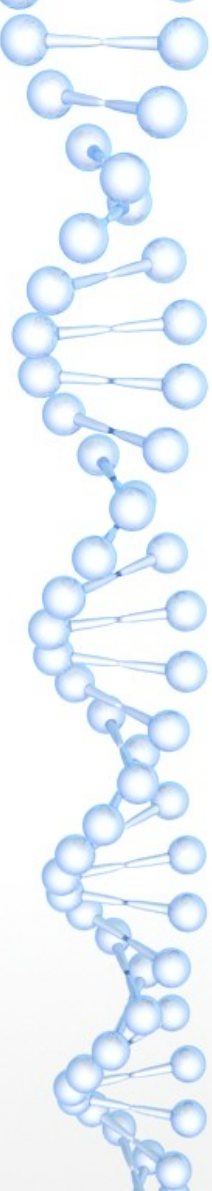
./metabat2

- -i <input file> (required) .fasta or .fasta.gz
- -o <*bin folder*>/<bin header> (required)
- -m (=2500) minimum size of a contig for binning
- --maxP arg (=95) Percentage of 'good' contigs considered for binning The greater, the more sensitive.
- --minS arg (=60) Minimum score of a edge for binning (should be between 1 and 99). The greater, the more specific.
- --maxEdges arg (=200) Maximum number of edges per node. The greater, the more sensitive.

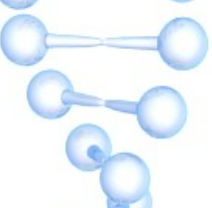
- 
- Bins in .fasta or .fa format

 0.fa	59.4 kB	Yesterday
 1.fa	4.6 MB	Yesterday
 2.fa	571.0 kB	Yesterday
 3.fa	3.5 MB	Yesterday
 4.fa	2.0 MB	Yesterday
 5.fa	18.5 MB	Yesterday
 6.fa	3.2 MB	Yesterday
 7.fa	6.7 MB	Yesterday
 8.fa	19.6 kB	Yesterday
 9.fa	142.4 kB	Yesterday
 10.fa	11.1 kB	Yesterday
		


Use CheckM to evaluate



```
## -S /bin/bash
## -cwd
## -M qz85@drexel.edu
## -P rosenclassPrj
## -l h_rt=48:00:00
## -l h_vmem=10G
## -l m_mem_free=10G
## -q all.q
. /etc/profile.d/modules.sh
module load proteus
module load gcc/4.8.1
module load sge/univa
module load shared
module load perl/5.20.0
module load oracle/jdk/1.7.0_current
module load samtools/1.2
module load bwa/master
module load groopm/0.3.5
checkm lineage_wf ./bins_2500/ ./bins_2500/SCG -x fa -t 8 -f ../bins_2500/CheckM.txt
```

Bin Id	Marker lineage	# genomes	# markers	# marker sets	0	1	2	3	4	5+	Completeness	Contamination	Strain heterogeneity
bin.8	o_Actinomycetales (UID1663)	488	309	185	0	308	1	0	0	0	100.00	0.18	0.00
bin.5	root (UID1)	5656	56	24	0	0	0	56	0	0	100.00	200.00	100.00
bin.4	root (UID1)	5656	56	24	0	0	5	37	14	0	100.00	201.62	29.50
bin.17	root (UID1)	5656	56	24	0	1	3	0	6	46	100.00	370.69	91.20
bin.16	root (UID1)	5656	56	24	0	0	56	0	0	0	100.00	100.00	98.21
bin.15	root (UID1)	5656	56	24	0	0	8	14	4	30	100.00	412.68	24.22
bin.6	o_Actinomycetales (UID1814)	148	572	276	5	566	1	0	0	0	99.35	0.36	0.00
bin.18	c_Betaproteobacteria (UID3888)	323	387	234	19	366	2	0	0	0	99.15	0.43	0.00
bin.2	o_Pseudomonadales (UID4488)	185	813	308	22	787	4	0	0	0	99.02	0.64	0.00
bin.20	f_Rhodobacteraceae (UID3340)	84	568	330	24	543	1	0	0	0	98.79	0.30	0.00
bin.14	k_Bacteria (UID1452)	924	151	101	2	130	19	0	0	0	98.68	10.89	0.00
bin.19	o_Burkholderiales (UID4000)	193	427	214	9	415	3	0	0	0	98.30	1.17	0.00
bin.7	c_Alphaproteobacteria (UID3337)	468	388	250	22	365	1	0	0	0	98.27	0.40	0.00
bin.23	c_Deltaproteobacteria (UID3216)	83	247	155	22	225	0	0	0	0	96.77	0.00	0.00
bin.12	c_Gammaproteobacteria (UID4202)	67	481	276	50	431	0	0	0	0	91.98	0.00	0.00
bin.3	o_Clostridiales (UID1212)	172	263	149	66	180	8	0	2	7	79.21	8.66	48.42
bin.22	p_Firmicutes (UID1022)	100	295	158	60	235	0	0	0	0	76.58	0.00	0.00
bin.9	k_Bacteria (UID2328)	3167	126	75	55	38	22	10	1	0	64.52	53.33	96.55
bin.13	root (UID1)	5656	56	24	42	1	0	0	2	11	58.33	208.33	99.18
bin.10	s_algicola (UID2847)	33	496	263	204	286	5	1	0	0	56.15	0.91	12.50
bin.1	root (UID1)	5656	56	24	14	0	1	0	2	39	41.67	159.03	97.77
bin.21	k_Bacteria (UID203)	5449	104	58	89	15	0	0	0	0	22.41	0.00	0.00
bin.11	p_Firmicutes (UID241)	930	213	118	135	77	1	0	0	0	20.99	0.42	0.00



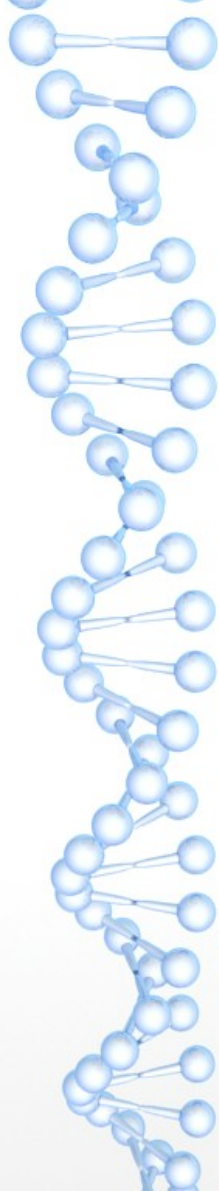


Recall

Precision	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
0.6	16	16	15	14	14	14	14	12	12	10
0.7	16	16	15	14	14	14	14	12	12	10
0.8	16	16	15	14	14	14	14	12	12	10
0.9	15	15	14	13	13	13	13	11	11	9
0.95	15	15	14	13	13	13	13	11	11	9
0.99	12	12	11	10	10	10	10	9	9	8

m = 2500(Default) removeStrain = F

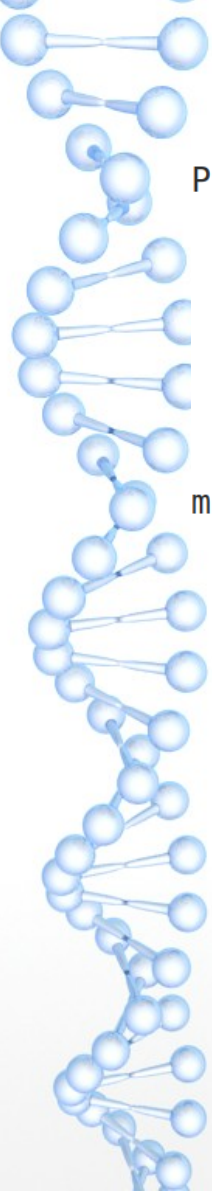




$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$



	Recall									
Precision	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
0.6	16	16	15	15	14	14	14	12	11	9
0.7	15	15	14	14	13	13	13	11	10	9
0.8	15	15	14	14	13	13	13	11	10	9
0.9	14	14	13	13	12	12	12	10	9	8
0.95	14	14	13	13	12	12	12	10	9	8
0.99	11	11	10	10	9	9	9	8	8	7

m = 2000 removeStrain = F

	Recall									
Precision	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
0.6	15	15	13	13	13	12	12	10	10	9
0.7	15	15	13	13	13	12	12	10	10	9
0.8	15	15	13	13	13	12	12	10	10	9
0.9	14	14	12	12	12	11	11	9	9	8
0.95	13	13	11	11	11	10	10	9	9	8
0.99	12	12	10	10	10	9	9	8	8	7

m = 1500 removeStrain = F



References

- <https://bitbucket.org/berkeleylab/metabat/src/master/>
- <https://github.com/Ecogenomics/CheckM/wiki>

