1 Assembly "Newbler"

1.1 General Information

This assembly uses the software "Newbler from Roche GS De Novo Assembler software v2.9" and uses a kmer size of 51, 61, 71 and 81 as well as a contig length cutoff of 1000 base pairs. There are 8 samples in this assembly, which are part of the project code `under_ice_rerun_bt` ("Rerun of under ice depth profile of lake BT").

1.2 Processing

- This report (and all the analysis) was generated using the GEFES project at:
  http://xapple.github.io/gefes/
- Version `0.1.5` of the pipeline was used.
- This document was generated at `2016-12-15 20:51:45 CET+0100`.
- The results and all the files generated for this sample can be found at:
  ssh://ww-hmem02.climb.cluster/home/lucas/GEFES/views/projects/under_ice_rerun_bt/
- The exact git hash of the latest commit was: 78325bb6760553550cb31b2adbd888f0395d099b

• 1.3 Also more simply referred to by its shortened tag `78325bb`.

1.4 Contigs

The first step is to take all reads from all samples and input them into the assembly algorithm. This results in the production of 44'165 contigs, of which the length distribution can be seen in figure 1:

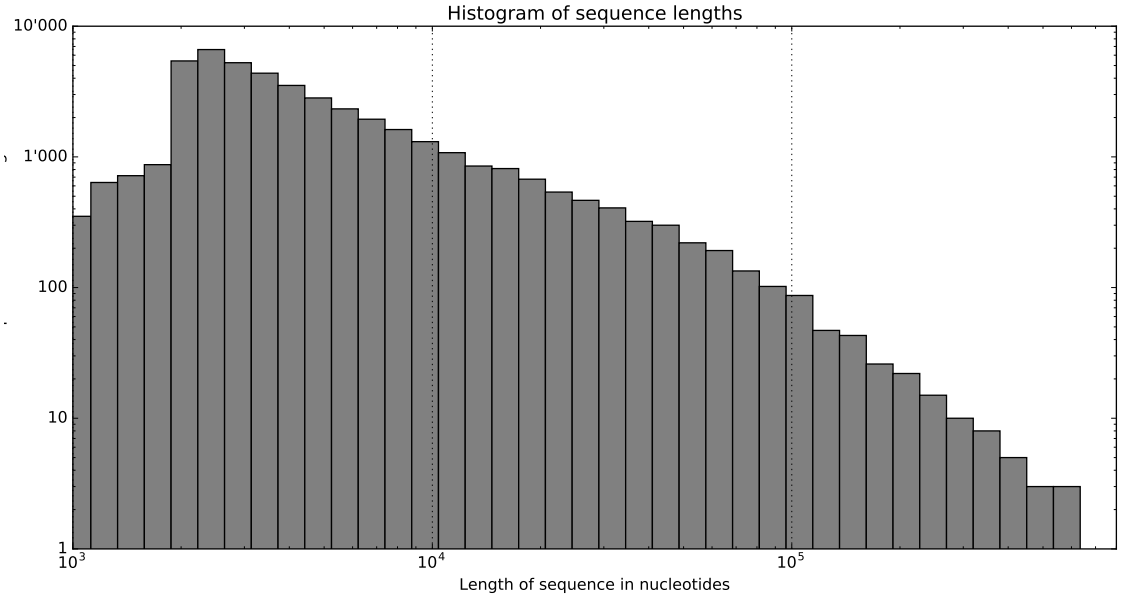

Figure 1. Assembly length distribution

The total amount of base pairs generated is 341'469'484.

1.5 Mapping

We can now take every read of every sample and attempt to map them back to the set of contigs we created, such that we compute, for every contig, a mean coverage in every sample. For this we use the "Bowtie2 2.2.5" software. A brief summary of the mapping step is included below:

	Name	Details	Reads	Did map
1	**bt1**	BT1	25'749'925	45.2%
2	**bt2**	BT2	32'220'952	43.2%
3	**bt3**	BT3	15'141'605	46.3%
4	**bt4**	BT4	19'901'153	45.9%
5	**bt5**	BT5	23'166'363	49.6%
6	**bt6**	BT6	18'766'706	48.2%
7	**bt7**	BT7	23'117'786	48.5%
8	**bt8**	BT8	21'646'381	32.3%

Table 1. Summary information for mapping of all samples.

1.6 Binning

The next step is to determine which of the 44'165 contigs are likely to belong together and are part of the same population or genome. For this we use the "concoct v0.4.0" software. This produces exactly 193 different bins. Since each bin contains a varying number of contigs, we can plot a distribution as seen in figure 2:

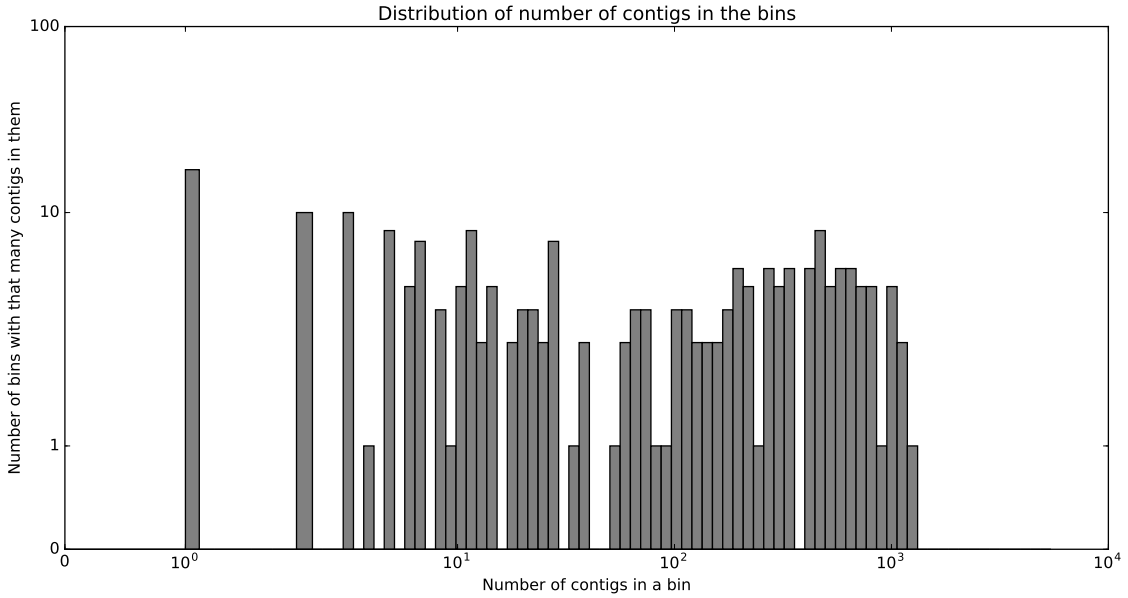

Figure 2. Bin number of contigs distribution

Of course, it is more interesting to look, not at the contig count, but at the total amount of nucleotides in a given bin in terms of base pair counts. This distribution can be seen in figure 3:

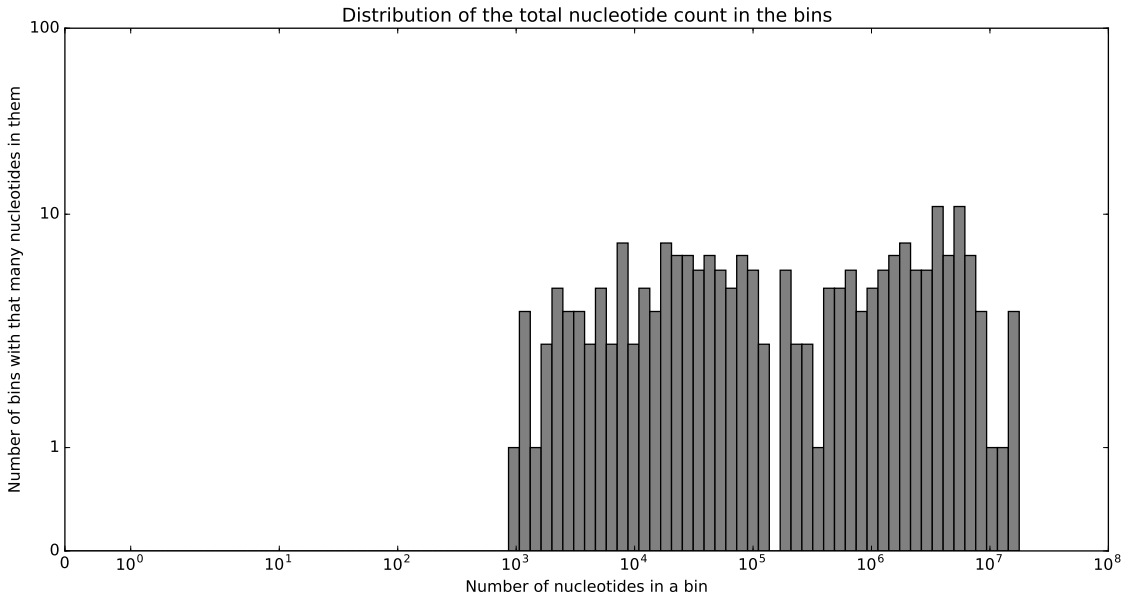

Figure 3. Bin total nucleotide size distribution

For further investigation please refer to the individual report of every bin.

1.7 Evaluation

Now we would like to measure the consistency or completeness of each bin. For this we use the "CheckM v0.9.7" software.

This includes running gene prediction on the contigs and identifying marker genes. A distribution of the different metrics can be seen below in the following figures:

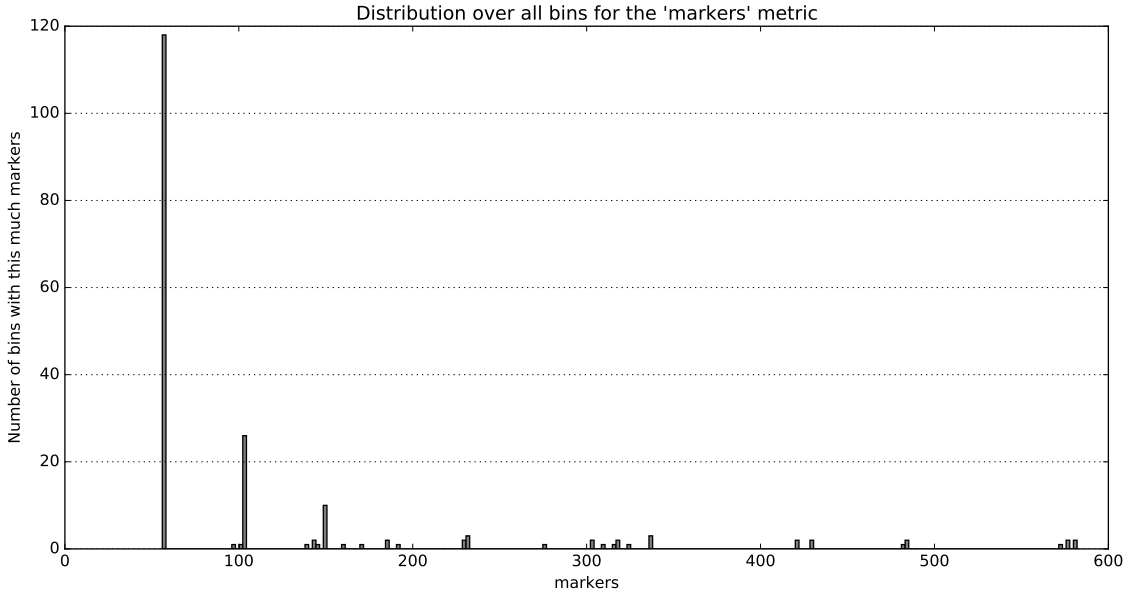

Figure 4. CheckMs 'markers' metric

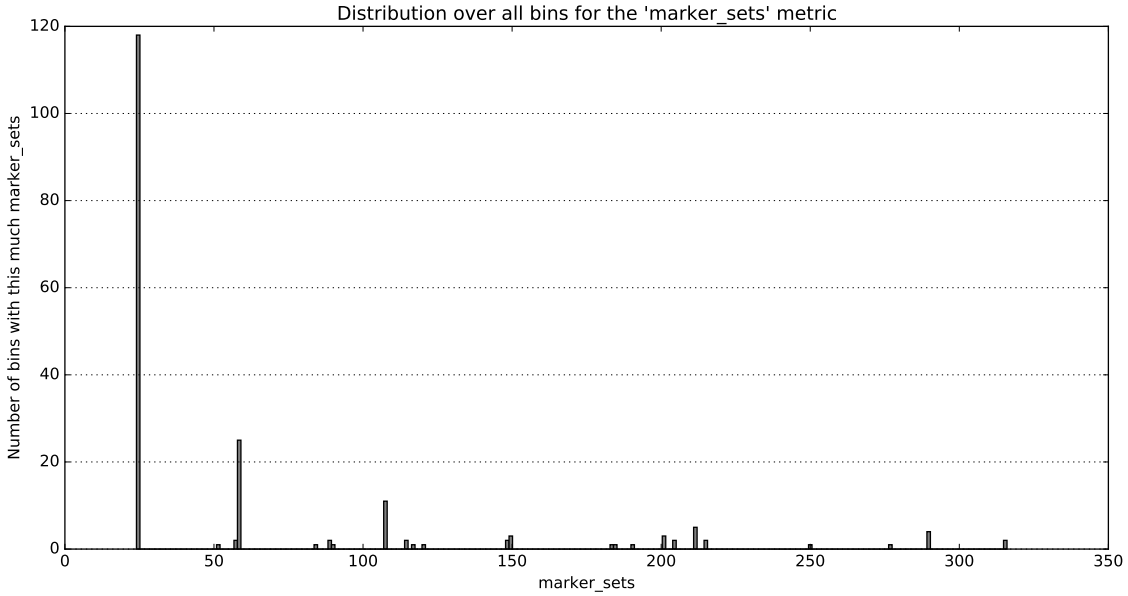

Figure 5. CheckMs 'marker_sets' metric

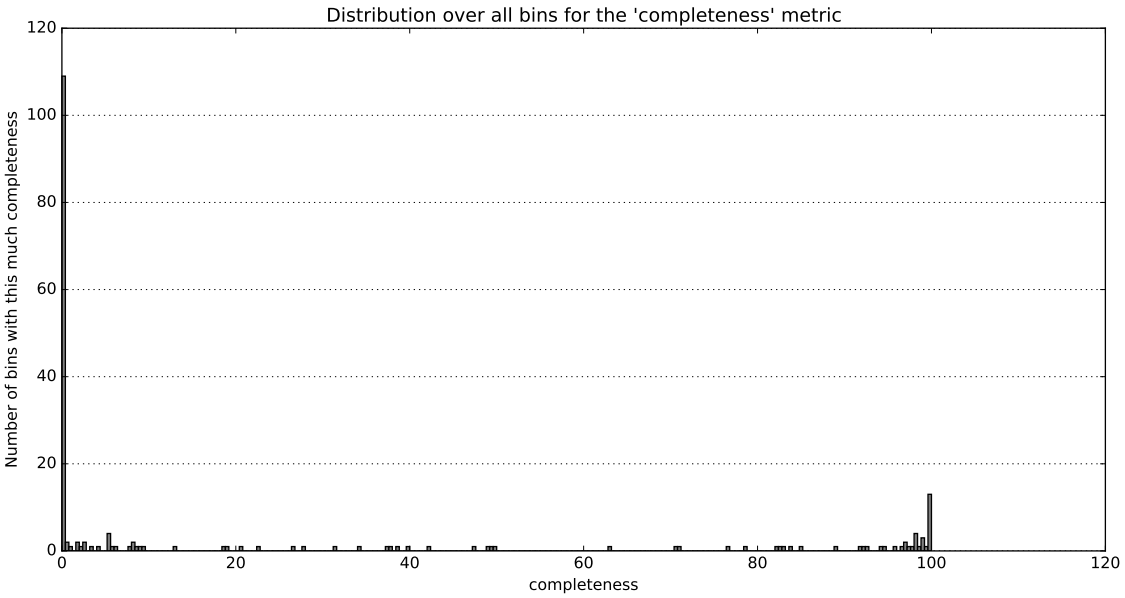

Figure 6. CheckMs 'completeness' metric

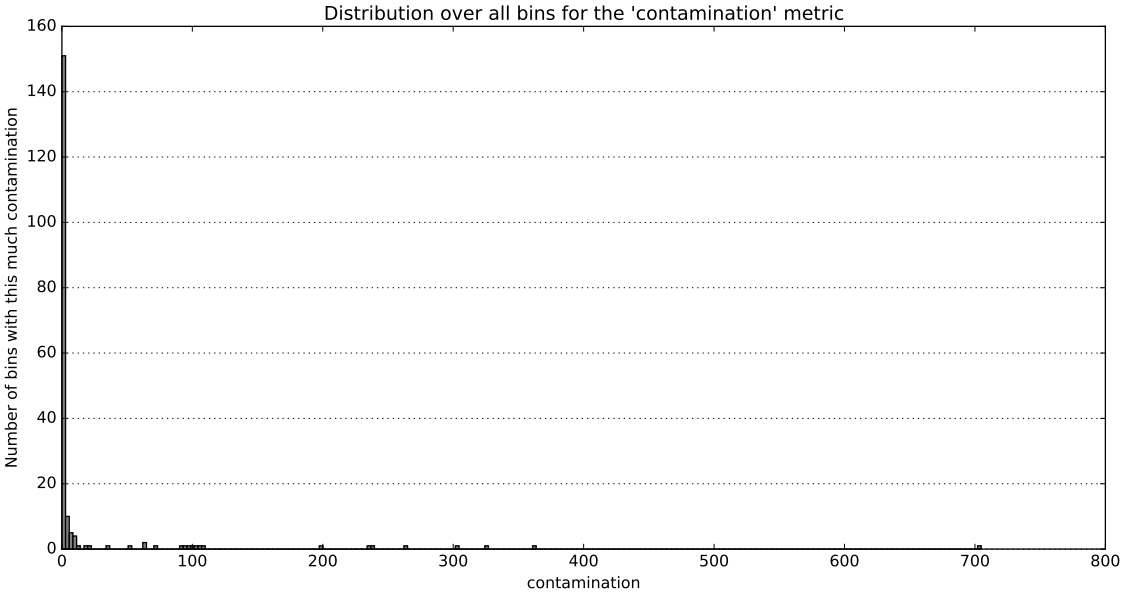

Figure 7. CheckMs 'contamination' metric

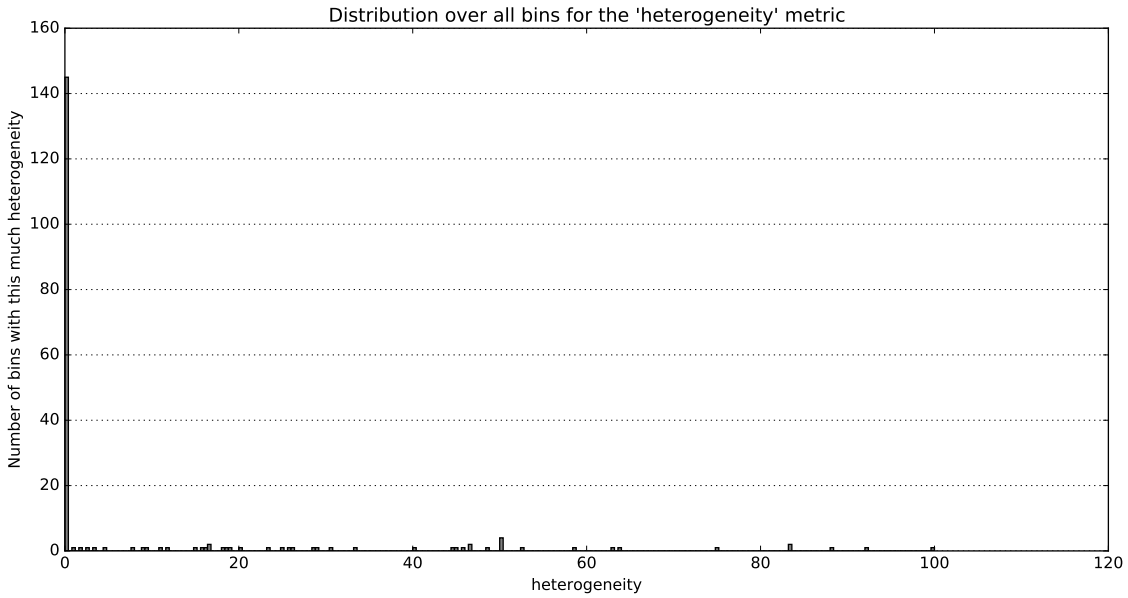

Figure 8. CheckMs 'heterogeneity' metric

We can view contamination, completeness and heterogeneity all in one graph (with heterogeneity as a color scale):

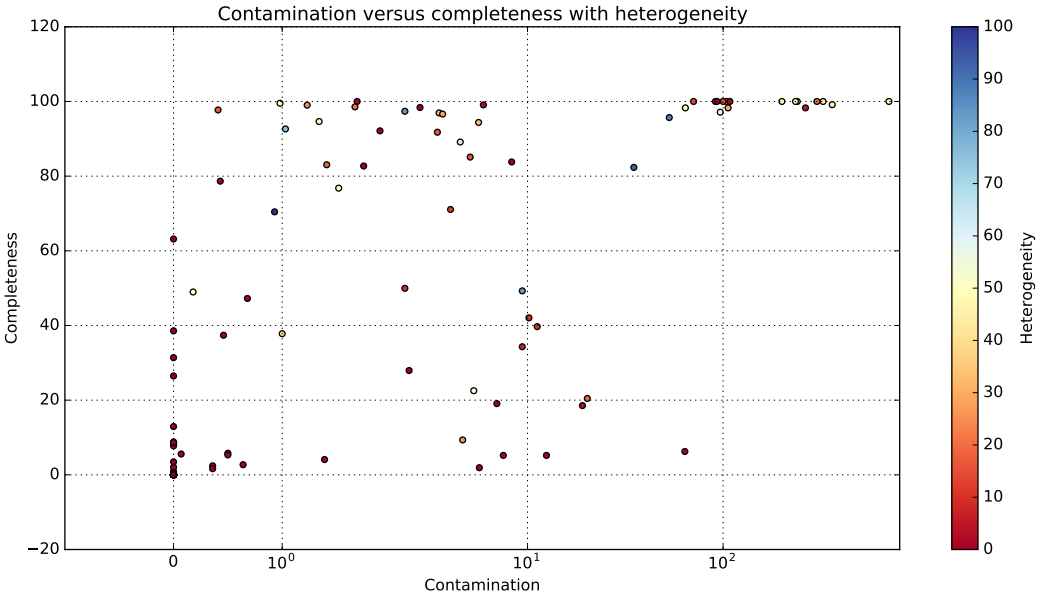

Figure 9. Contamination versus completeness with heterogeneity

Finally, we can produce a table listing only the bins which are more than 60% complete and less than 10% contamination. Here is the list of these best bins:

#	Compl.	Conta.	Heter.	Prots.	Avg. cov.	Assignment
50	100	1.69	0	1735	0.56	Rickettsiales (o)
6	99.51	0.98	50	4440	0.48	Bacteroidetes (p)
116	99.08	5.95	2.56	6042	1.43	Acidobacteria (p)
192	99.02	1.23	25	3943	0.41	Bacteroidetes (p)
99	98.56	1.67	20	4119	0.83	Methylococcaceae (f)
190	98.38	2.82	0	5363	0.62	Verrucomicrobiaceae (f)
66	97.72	0.41	16.67	4004	0.8	Nitrosomonadaceae (f)
112	97.37	2.36	83.33	6073	2.03	Rhodospirillales (o)
55	96.93	3.53	30.77	3509	1.74	Methylococcaceae (f)
141	96.63	3.68	28.57	4438	1.29	Planctomycetacia (c)
118	94.63	1.34	50	1475	0.23	Verrucomicrobia (p)
167	94.38	5.62	33.33	5062	0.58	Planctomycetacia (c)
134	92.64	1.03	75	3622	0.29	Bradyrhizobiaceae (f)
156	92.13	1.9	0	3361	0.28	Bacteroidetes (p)
86	91.76	3.46	16.67	2433	0.25	Xanthomonadaceae (f)
51	89.14	4.53	63.16	1701	0.72	Actinomycetales (o)
36	85.1	5.09	15.79	1408	0.12	Actinomycetales (o)
78	83.8	8.3	0	3588	0.24	Bacteroidetes (p)
139	83.05	1.41	18.18	2605	0.19	Methylococcaceae (f)
100	82.71	1.75	0	2894	0.22	Verrucomicrobiaceae (f)
88	78.66	0.43	0	5106	0.33	Acidobacteria (p)
126	76.78	1.52	50	3680	0.25	Rhodospirillales (o)
166	71.07	4.04	11.76	2303	0.13	Comamonadaceae (f)
150	70.44	0.93	100	3184	0.17	Chloroflexi (p)
111	63.16	0	0	3832	0.26	Planctomyces (g)

Table 2. Summary table for the best bins in this assembly.

In total 11.365% of the original reads from all samples map back to these 9'338 contigs.

1.8 Visualization

To better view the binning that was performed, we can visualize a few of the properties of the bins. The first graph is a regression of all bins comparing their predicted number of genes against

the number of predicted PFAMs.

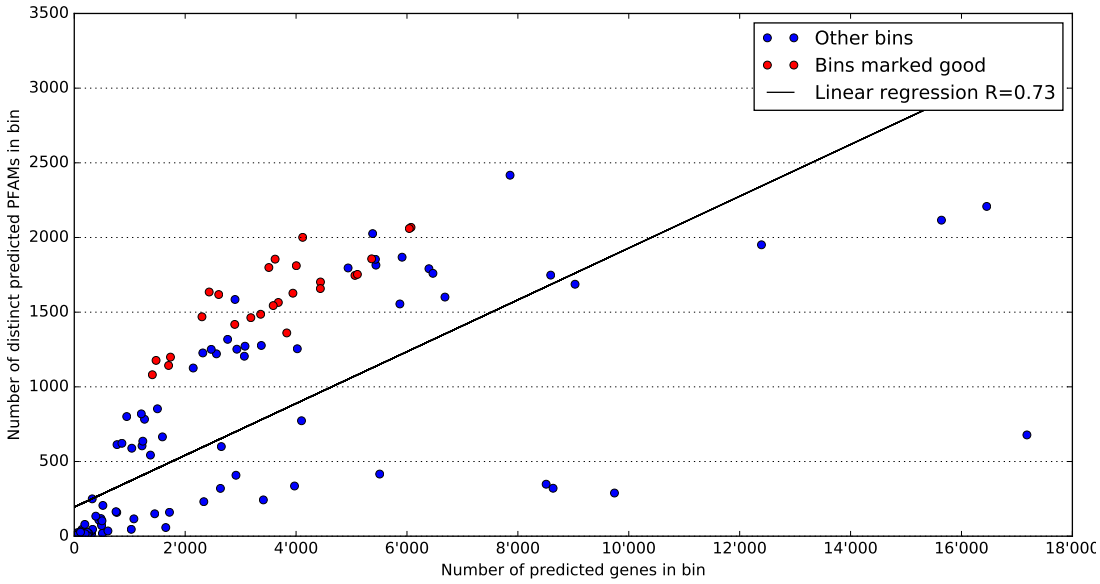

Figure 10. Comparison of predicted number of genes against the number of predicted PFAMs

The second graph is a regression of all bins comparing their cumulative length in base pairs against the number of predicted genes.

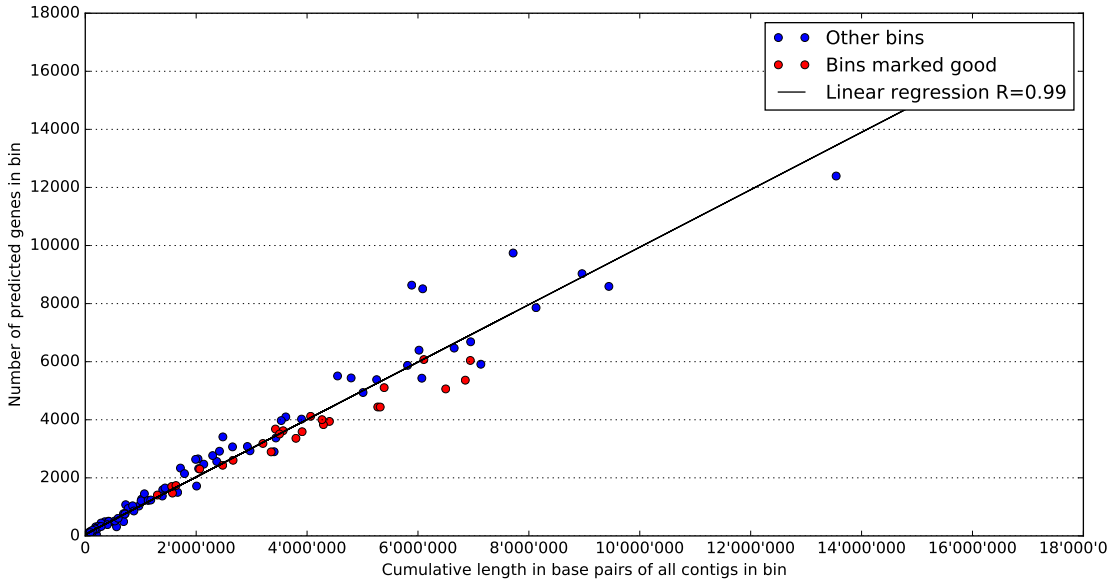

Figure 11. Comparison of cumulative length in base pairs against the number of predicted genes

One can plot every contig in an ordination plot, where contigs with similar coverage and nucleotide composition are placed closer together.