Sample "rl1am"

General Information

This sample has the code "rl1am" and is named "rl1am". It is part of the project code "ice" ("ice") along with 39 other samples.

Meta-data details

The JSON file located online at:

https://github.com/limno/illumitag/tree/master/json/presamples/run010/run010-sample026.json

contains all the metadata associated with this sample. Here are the contents of this file:


```
"uppmax_id":    "b2014083",
"run_num":       10,
"run_id":        "140610_M00629_0001_000000000-A8H0L",
"sample_num":    26,
"sample_id":     "Sample_AE_POOL1_2014_26",
"forward_reads": "AE_POOL1_2014_26_CTAGTACG-TATCCTCT_L001_R1_001.fastq.gz",
"reverse_reads": "AE_POOL1_2014_26_CTAGTACG-TATCCTCT_L001_R2_001.fastq.gz",

"project":       "ice",
"project_name":  "ice",
"sample":        "rl1am",
"sample_name":   "rl1am",
"group":         "ice-rl",

"forward_mid":   "TATCCTCT",
"forward_num":   503,
"reverse_mid":   "CTAGTACG",
"reverse_num":   702,
"barcode_num":   26,

"primers": {
        "name":    "General bacteria primers",
        "sense":   "5' to 3'",
        "forward": {"name": "341F", "sequence": "NNNNCCTACGGGNGGCWGCAG"},
        "reverse": {"name": "805R", "sequence": "GACTACHVGGGTATCTAATCC"}
},

"library_strategy":     "AMPLICON",
"library_source":       "METAGENOMIC",
"library_selection":    "PCR",
"library_layout":       "Paired-end",
"platform":             "ILLUMINA",
"instrument_model":     "Illumina MiSeq",
"forward_read_length":  300,
"reverse_read_length":  300,
```

```
"date":         "0000-00-00",
"latitude":     ["XX°XX'XX'", "N"],
"longitude":    ["XX°XX'XX'", "E"],
"location":     "Country: XXX lake",
"organism":     "aquatic metagenome",

"bioproject":   "PRJNAXXXXXX",
"biosample":    "SAMNXXXXXXXX",

"dna_after_purification": [2.792, "ng/µl"]
```

Processing

This report (and all the analysis) was generated using the ILLUMITAG project at:

http://github.com/limno/illumitag

Version `1.0.0` of the pipeline was used. The exact git hash of the latest commit was:

```
e902cd63af4b634a255bb90c228c54ace07017d6
```

also refereed to by its tag `submission2-40-ge902cd6-dirty`. This document was generated at `2014-08-01 01:20:12 CEST+0200`.

A brief overview of what happens to the data can be viewed online here:

https://github.com/limno/illumitag/blob/master/documentation/pipeline_outline.pdf?raw=true

The results and all the files generated for this sample can be found on UPPMAX at:

```
/home/lucass/ILLUMITAG/views/presamples/run010-sample26/
```

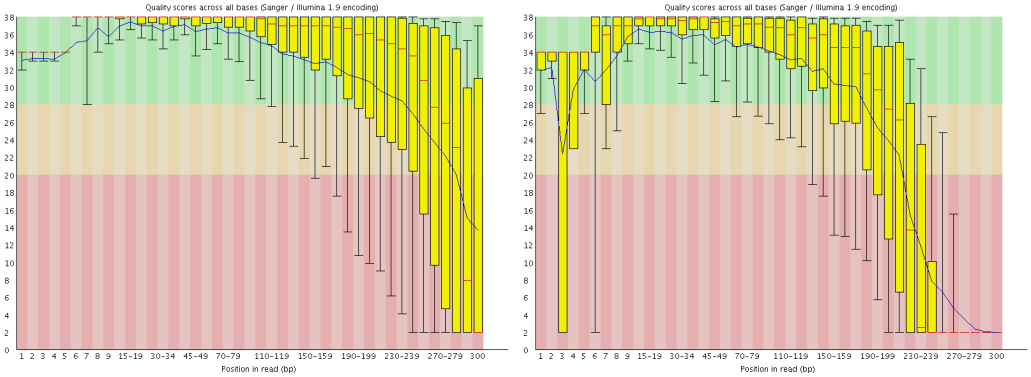
Raw data

- The forward read file weighed *30.5 MB* and contained *121'678* reads with an average PHRED quality of *31.45*
- The reverse read file weighed *30.6 MB* and contained *121'678* reads with an average PHRED quality of *25.14*

More information about the raw output of the sequencer for this sample can be found in the HTML report generated by the Illumina software here:

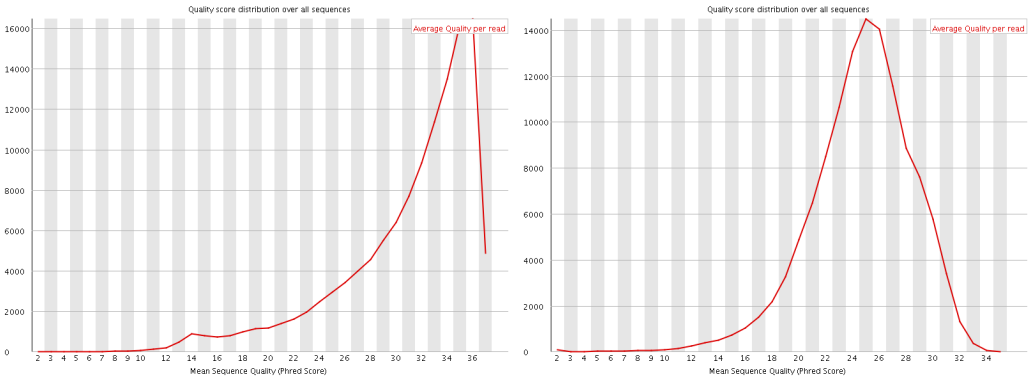
```
/home/lucass/proj/b2014083/INBOX/140610_M00629_0001_000000000-A8H0L/report.html
```

The average quality per base can be seen in figure 1 and the average quality per sequence in figure 2.



(a) Forward **(b)** Reverse

Figure 1. Per base quality



(a) Forward **(b)** Reverse

Figure 2. Per sequence quality

Joining

The first step in processing is joining the forward and reverse reads together. This is done with a program called "PandaSeq". Exactly 120'361 (98.9%) reads were joined, 778 (0.6%) where unable to be joined and 539 (0.4%) were deemed of too low quality by the algorithm. The size of the overlapping region varies for every read, hence a distribution of sequence lengths is produced after is step and can be seen in figure 3. In addition, you can see the per base quality of the joined reads in figure 4a and the per sequence quality in figure 4b.

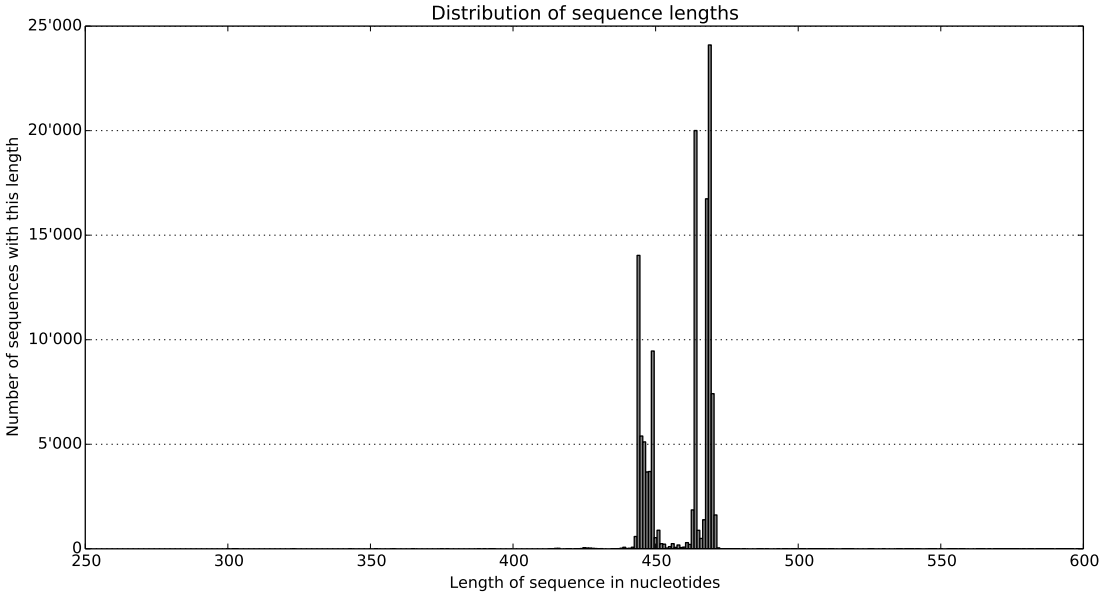
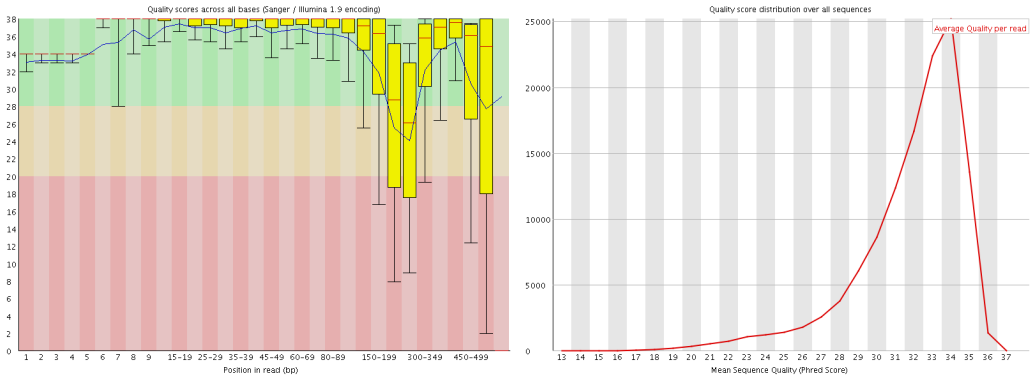


Figure 3. Distribution of sequence lengths after joining



(a) Per base **(b)** Per sequence

Figure 4. Joined sequence quality

Filtering

Next, we filter the sequences based on several criteria. These many sequences are lost:

- Checking for the presence of both primers at their exact positions with at most 1 mismatches allowed discards 24'563 sequences (95'798 left).
- Checking for the absence of undetermined "N" bases anywhere in the reads (expect in the primers) discards 9'168 sequences (86'630 left).

- Running a 10 base pair window over the sequences and checking that quality doesn't drop below 5 discards 10'111 sequences (76'519 left).
- Checking that no sequence is shorter than 400 base pairs discards 7 sequences (76'512 left).

This leaves us with 62.9% of the original sequences. The next step in the pipeline is generating outputs that are compatible with "Qiime" and "Mothur" for those who are interested in that. We will also trim the primers on both ends of sequences now.

Taxonomy

Further analysis is usually performed not at the single sample level but on a bunch of samples at the same time (samples are grouped into cluster), so you should go see the corresponding cluster report now. This will include OTU generation, taxonomic assignment and more filtering. Nonetheless we can present some of the results from that analysis here. Here are the 20 most abundant species observed in this sample:

Rank	Clade	Reads	OTUs
1	vadinHA64	4'235	3
2	Candidate division OD1	4'018	2747
3	hgcI clade	2'078	15
4	Candidate division OP3	2'022	573
5	Polynucleobacter	1'929	2
6	Flavobacterium	1'661	19
7	Chthoniobacter	1'549	14
8	Comamonadaceae	1'394	16
9	Candidatus Solibacter	1'273	11
10	TM214	1'042	21
11	OPB35 soil group	1'008	99
12	Methylococcales	938	8
13	Oxalobacteraceae	878	13
14	Candidate division TM7	719	232
15	Methylophilaceae	701	11
16	Acetobacteraceae	644	16
17	Opitutus	596	19
18	Limnohabitans	578	5
19	Chitinophagaceae	563	35
20	CL500-3	548	6

Table 1. The 20 most abundant species in this sample.

Diversity

One of the basic questions one can ask is "What is the diversity in this sample ?", or more specifically "What are the richness and evenness ?". Several estimators are available. Moreover, we can calculate these estimates at several rarefaction values to check if the sampling depth was sufficient. We input the OTU count vector for this specific sample (containing 44'435 sequences and 819 OTUs) and down-sample it iteratively (without replacement). The Chao1, ACE, Shannon and Simpsons estimates can be seen in figures 5, 6, 7 and 8.

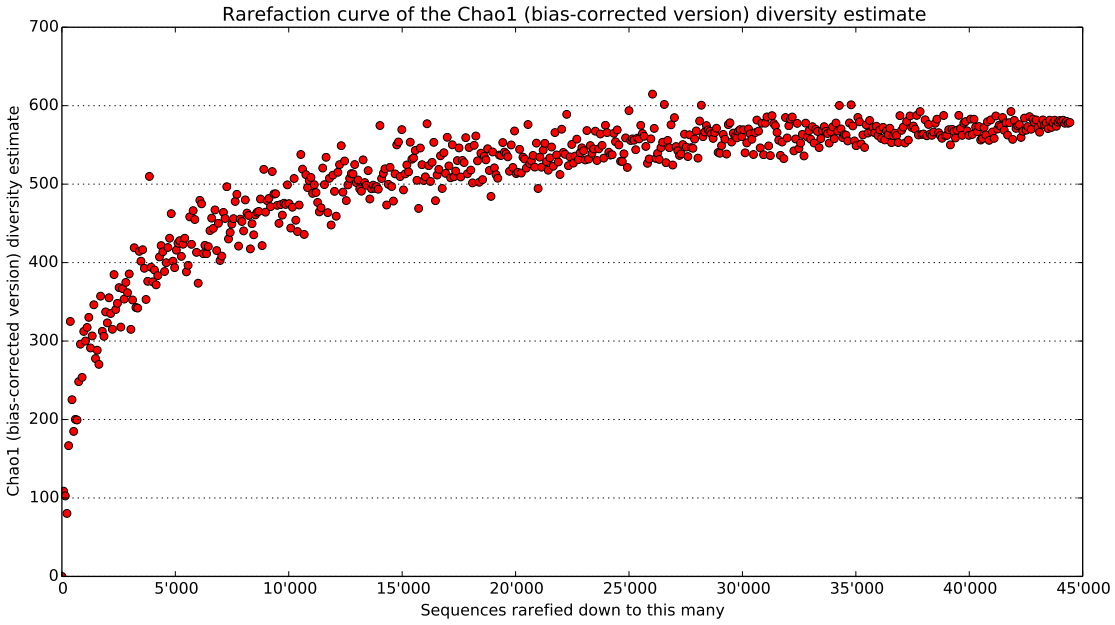


Figure 5. Chao1 rarefaction curve

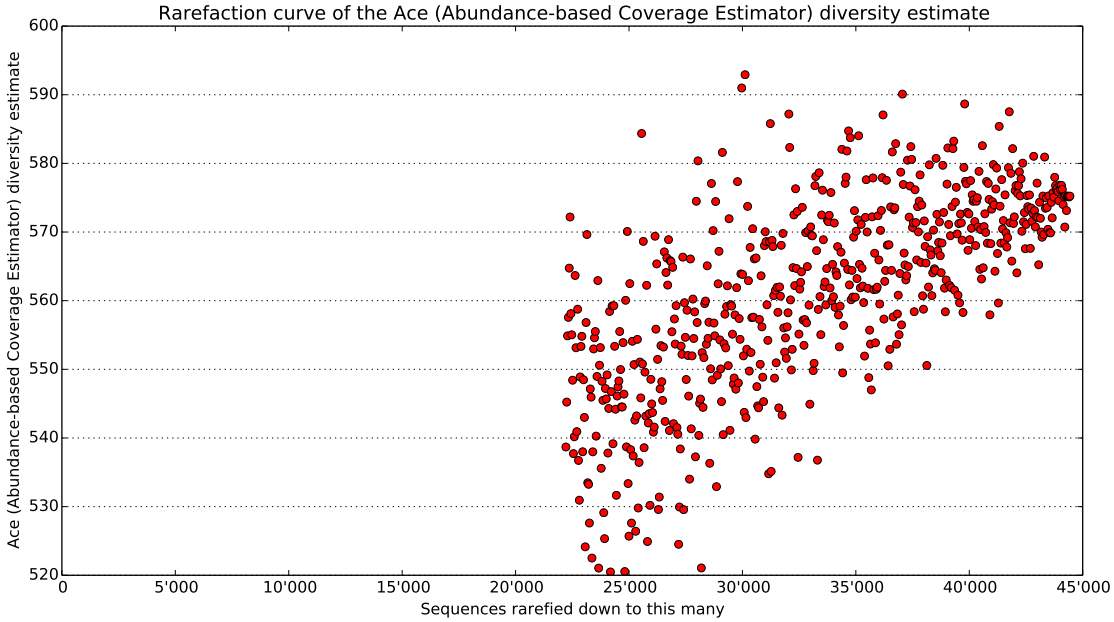


Figure 6. Ace rarefaction curve

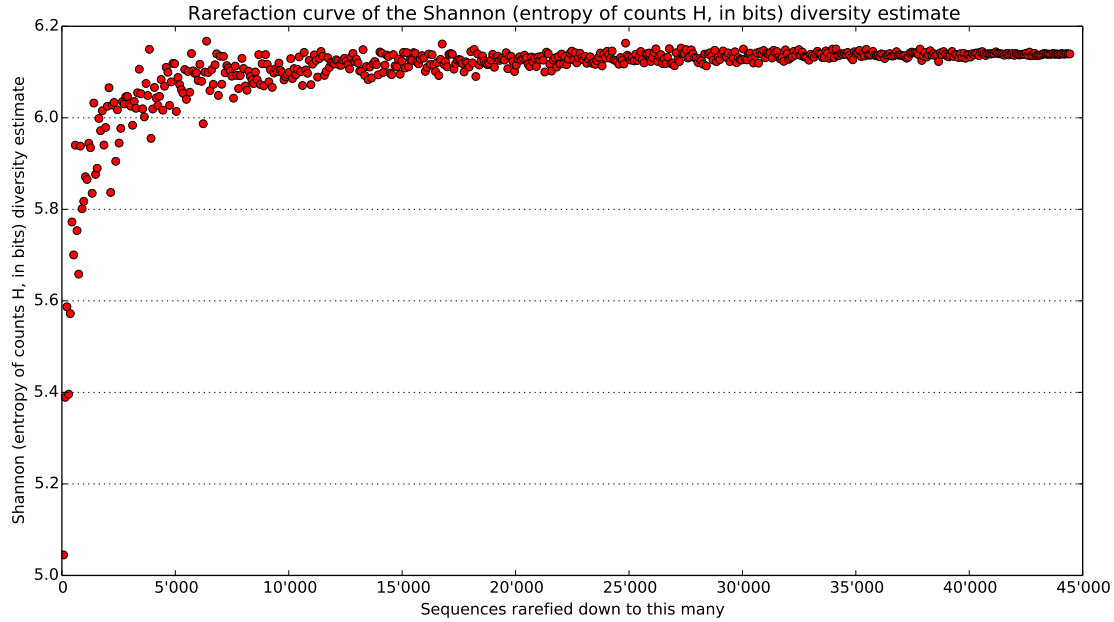


Figure 7. Shannon rarefaction curve

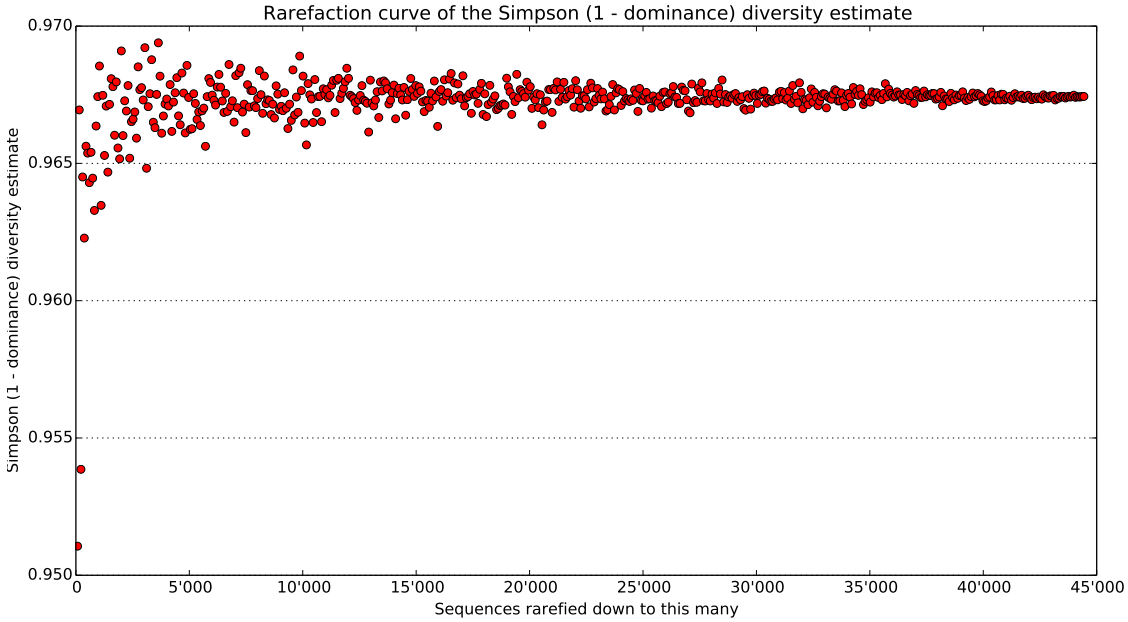


Figure 8. Simpson rarefaction curve

A list of other estimators that can easily be added can be seen here:

http://scikit-bio.org/docs/0.1.4/math.diversity.alpha.html