Sequencing Data Quality Assessment

Nate Olson

2017-03-14

Quality assessment of sequencing run summarizing the PhiX error rate, number of reads per sample, and quality score distributions over the length of the reads. Differences in the number of reads per sample are important for normalization and quality score distribution over read length is relevant to merging forward and reverse reads as well as the quality of the sequencing data in the middle of the amplicon.

Read Counts

Two barcoded experimental samples have less than 50,000 reads @ref(fig:readCount). The rest of the samples with less than 50,000 reads are negative PCR controls (NTC). Sample E01JH0016 titration 5 position F9 of plate 1 initial 16S PCR failed.
Excluding the one failed reaction the total range in the observed number of sequences per sample is approximately 40,000 to 150,000 reads.

TODO Figure out why E01JH0011 titration 3 position D2 plate 2 is also low, look at picogreen post normaization data. **TODO** Comparison to differences in number of reads per sample observed in other studies.

Table 1: Summary statistics for experimental and no template control samples by PCR plate and read.

exp_ntc	plate	mean_lib_size	min_lib_size	median	max_lib_size
EXP	plate1	89370.33	52302	86007.0	152267
EXP	plate2	96263.40	46328	97615.5	143110
NTC	plate1	3698.50	1305	1715.0	13146
NTC	plate2	13159.83	5216	9750.5	25349

PhiX Error Rate

The sequencing error rate data was obtained from the Basespace sequencing run report downloaded from Basespace (SAV file). Error rate is compared to a 16S public dataset on basespace (16S-Metagenomic-Library-Prep run id 3861867). The error rate for was higher at the ends of both R1 and R2 compared to the public dataset @ref(fig:phixError). This higher error rate may impact the number of reads pairs successfully merged and the accuracy of the sequence in the middle of the amplicon.

Base Quality Score

Cycle base quality score is more homogeneous from PCR plate 2 samples than plate 1 @ref(fig:lowcycleq). For the expected overlap region, based on primer positions and read lengths (16S PCR fig), the forward read has consistently higher base quality scores relative to the reverse read. The sample with a lower base quality score is titration 4 of biological replicate E01JH0017 (sample 1-F4).

Conclusion

The low variabiltiy in number of reads per sample suggests that normalization methods are less likely to impact the results compared other datasets with larger variability in the number of reads per sample. The

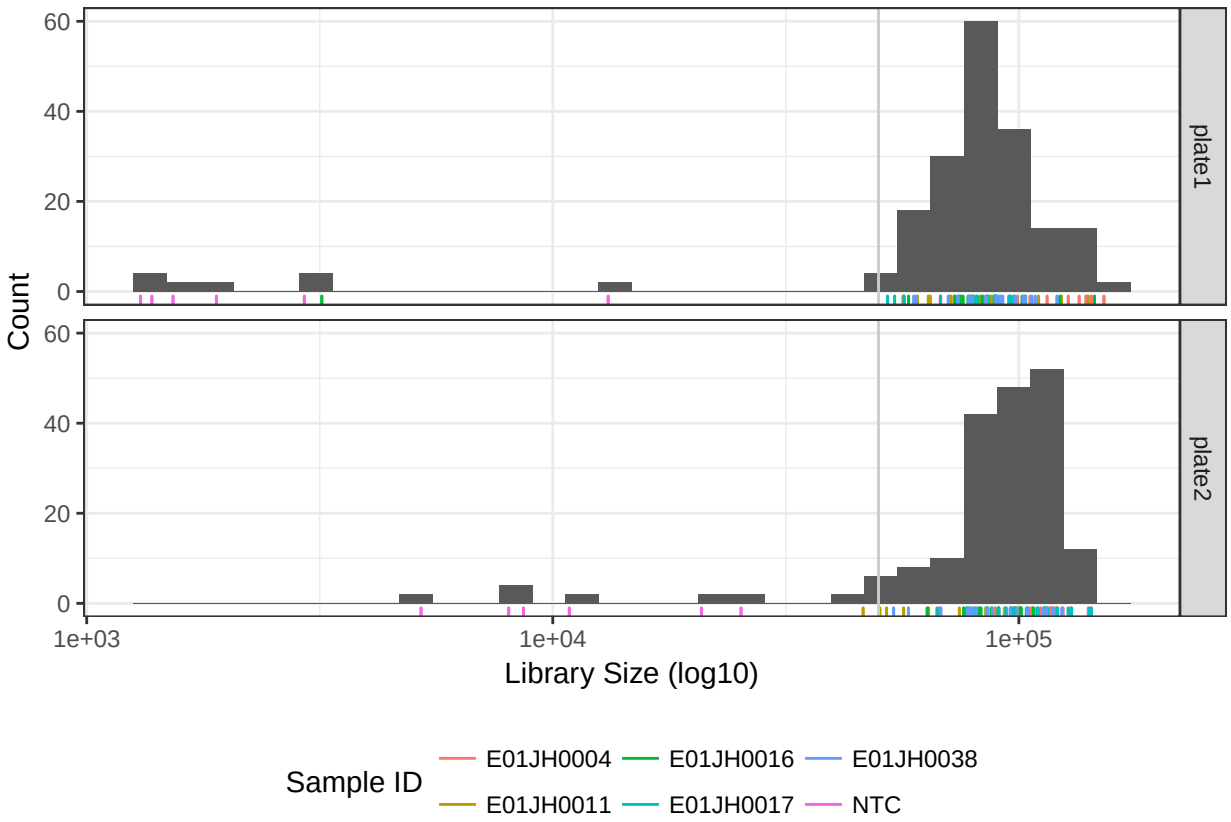

Figure 1: Number of reads per barcoded sample (Library Size), by read direction (X-facet) and replicate 16S PCR plate (Y-facet). Vertical line indicates 50,000 reads per barcoded sample.

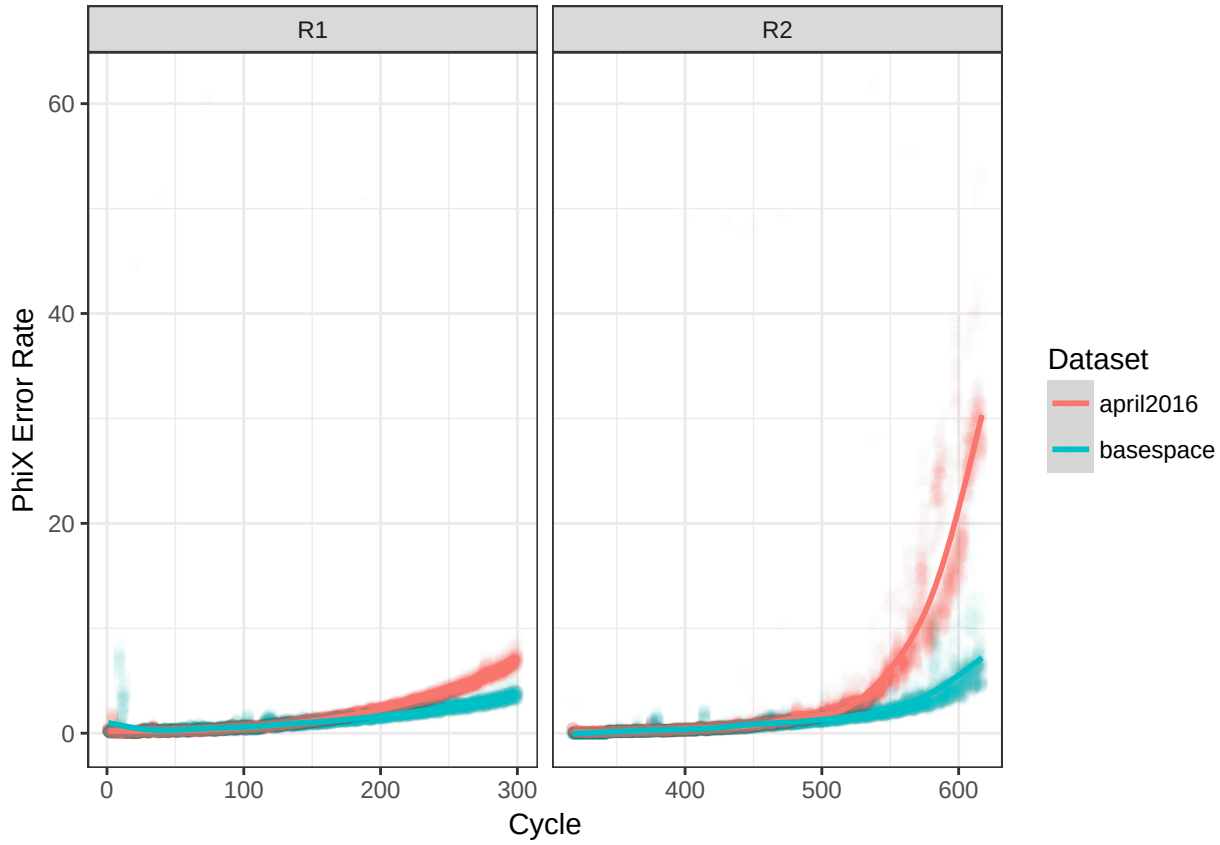

Figure 2: PhiX error rate for JHU barcoded samples compared to the public dataset. R1 and R2 are the forward and reverse reads and cycle is the position in the read.

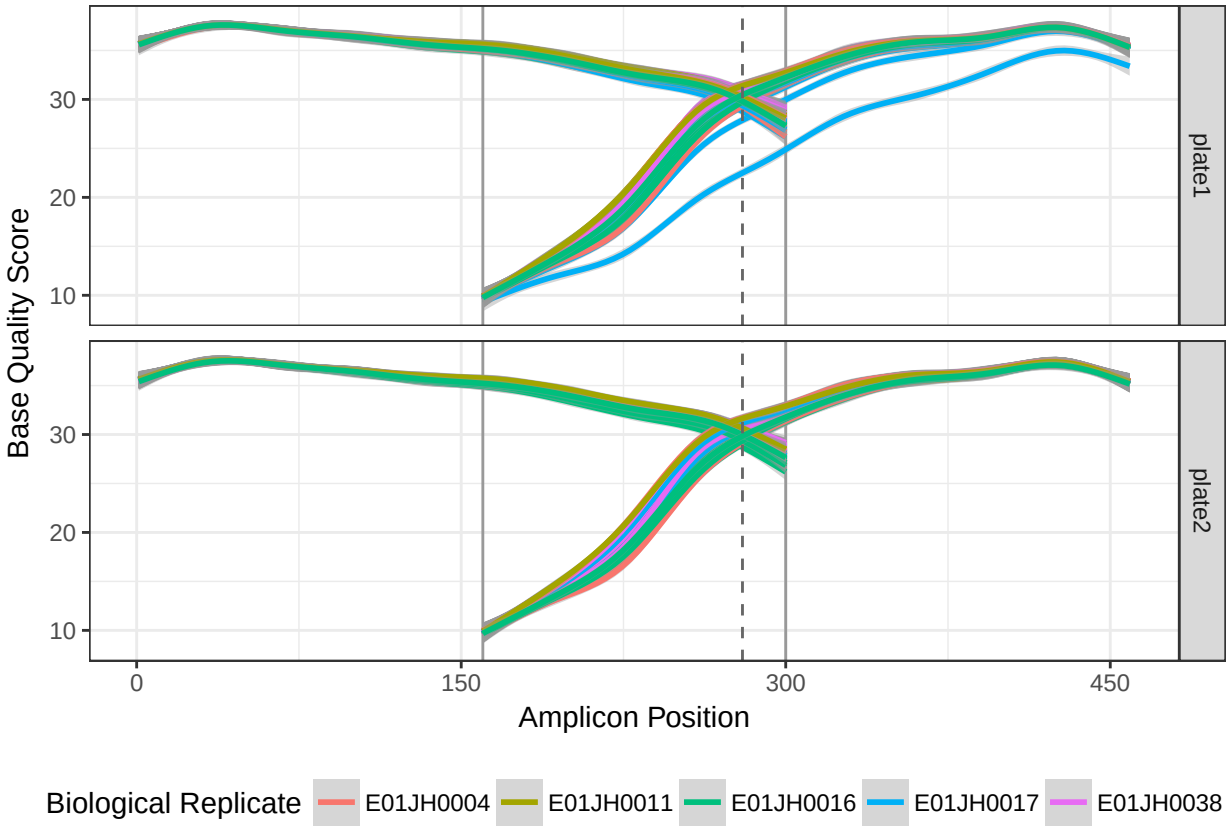

Figure 3: Smoothing spline of the base quality score by sequencing cycle. Vertical lines indicate approximate overlap region between forward and reverse reads. This is not a read level analysis but average quality score for individual barcoded datasets.

higher error rates at the ends of the reads has the potential to impact the quality of the reads in the overlap region and limit the number of successfully merged read pairs.

Session information

```
##  setting  value
##  version  R version 3.3.2 (2016-10-31)
##  system   x86_64, darwin15.6.0
##  ui       unknown
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-03-14
```

package	version	date	source
bbmle	1.0.18	2016-02-11	CRAN (R 3.3.2)
Biobase	2.34.0	2016-11-07	Bioconductor
BiocGenerics	0.20.0	2016-11-07	Bioconductor
BiocParallel	1.8.1	2016-11-07	Bioconductor
Biostrings	2.42.1	2016-12-19	Bioconductor
DESeq	1.26.0	2016-11-28	Bioconductor
DESeq2	1.15.28	2017-02-02	bioc (readonly/DESeq2@125913)
dplyr	0.5.0	2016-06-24	CRAN (R 3.3.2)
edgeR	3.16.5	2017-02-02	Bioconductor
forcats	0.2.0	2017-01-23	CRAN (R 3.3.2)
foreach	1.4.3	2015-10-13	CRAN (R 3.3.1)
GenomeInfoDb	1.10.2	2017-01-04	Bioconductor
GenomicAlignments	1.10.0	2016-11-07	Bioconductor
GenomicRanges	1.26.2	2017-01-04	Bioconductor
ggplot2	2.2.1	2016-12-30	CRAN (R 3.3.2)
glmnet	2.0-5	2016-03-17	CRAN (R 3.3.1)
IRanges	2.8.1	2016-11-18	Bioconductor
knitr	1.15.1	2016-11-22	CRAN (R 3.3.2)
lattice	0.20-34	2016-09-06	CRAN (R 3.3.2)
limma	3.30.9	2017-02-02	Bioconductor
locfit	1.5-9.1	2013-04-20	CRAN (R 3.3.1)
Matrix	1.2-8	2017-01-20	CRAN (R 3.3.2)
metagenomeSeq	1.16.0	2016-11-07	Bioconductor
modelr	0.1.0	2016-08-31	cran (@0.1.0)
permute	0.9-4	2016-09-09	CRAN (R 3.3.1)
phyloseq	1.19.1	2017-01-04	Bioconductor
ProjectTemplate	0.7	2016-08-11	CRAN (R 3.3.1)
purrr	0.2.2	2016-06-18	CRAN (R 3.3.1)
RColorBrewer	1.1-2	2014-12-07	CRAN (R 3.3.1)
readr	1.0.0	2016-08-03	CRAN (R 3.3.1)
readxl	0.1.1	2016-03-28	cran (@0.1.1)
Rqc	1.8.0	2016-11-07	Bioconductor
Rsamtools	1.26.1	2016-11-07	Bioconductor
S4Vectors	0.12.1	2016-12-19	Bioconductor
sads	0.3.1	2016-05-13	CRAN (R 3.3.2)
savR	1.12.0	2016-11-07	Bioconductor
ShortRead	1.32.0	2016-11-07	Bioconductor

package	version	date	source
stringr	1.1.0	2016-08-19	CRAN (R 3.3.1)
SummarizedExperiment	1.4.0	2016-11-07	Bioconductor
tibble	1.2	2016-08-26	CRAN (R 3.3.1)
tidyr	0.6.1	2017-01-10	CRAN (R 3.3.2)
tidyverse	1.1.1	2017-01-27	CRAN (R 3.3.2)
vegan	2.4-2	2017-01-17	CRAN (R 3.3.2)
XVector	0.14.0	2016-11-07	Bioconductor