NIST Run 1 Seq QA

Nate Olson

2017-02-16

Sequencing Data Quality Assessment

To generate summaries of QA metrics for the 384 datasets in the study (192 samples with forward and reverse reads) used the bioconductor `Rqc` package (REF) to calculate the quality metrics used in the following analysis.

Read Counts

Two barcoded experimental sample has less than 50,000 reads. 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. **TODO** Figure out why E01JH0011 titration 3 position D2 plate 2 is also low, look at picogreen post normaization data.

```
## # A tibble: 383 × 20
##                ID biosample_id titration pcr_16S_plate   pos barcode_lab
##             <chr>        <chr>     <dbl>          <chr> <chr>       <chr>
## 1  B1_M1_P1_L2_S2    E01JH0004        20              1    F6        NIST
## 2  B1_M1_P1_L2_S2    E01JH0004        20              1    F6        NIST
## 3  B1_M2_P1_L2_S2    E01JH0004         0              1    A1        NIST
## 4  B1_M2_P1_L2_S2    E01JH0004         0              1    A1        NIST
## 5  B1_M3_P1_L2_S2    E01JH0004         1              1    B1        NIST
## 6  B1_M3_P1_L2_S2    E01JH0004         1              1    B1        NIST
## 7  B1_M4_P1_L2_S2    E01JH0004         2              1    C1        NIST
## 8  B1_M4_P1_L2_S2    E01JH0004         2              1    C1        NIST
## 9  B1_M5_P1_L2_S2    E01JH0004         3              1    D1        NIST
## 10 B1_M5_P1_L2_S2    E01JH0004         3              1    D1        NIST
## # ... with 373 more rows, and 14 more variables: kit_version <chr>,
## #   For_Index_ID <chr>, Rev_Index_ID <chr>, seq_lab <chr>, pos_ns <chr>,
## #   ill_id <chr>, plate <chr>, Read <chr>, seq_ds_id <chr>,
## #   filename <chr>, pair <fctr>, group <chr>, reads <int>,
## #   total.reads <int>
```

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

exp_ntc	Read	plate	mean_lib_size	min_lib_size	median	max_lib_size
EXP	R1	plate1	56289.111	34727	54913.5	96550
EXP	R1	plate2	30255.180	16423	29666.0	59670
EXP	R2	plate1	56289.111	34727	54913.5	96550
EXP	R2	plate2	30309.322	16423	29667.0	59670
NTC	R1	plate1	57948.667	35169	55519.0	87628
NTC	R1	plate2	6647.667	3270	5830.5	14359
NTC	R2	plate1	57948.667	35169	55519.0	87628
NTC	R2	plate2	6647.667	3270	5830.5	14359

Potential issue with high seq number for no template controls.

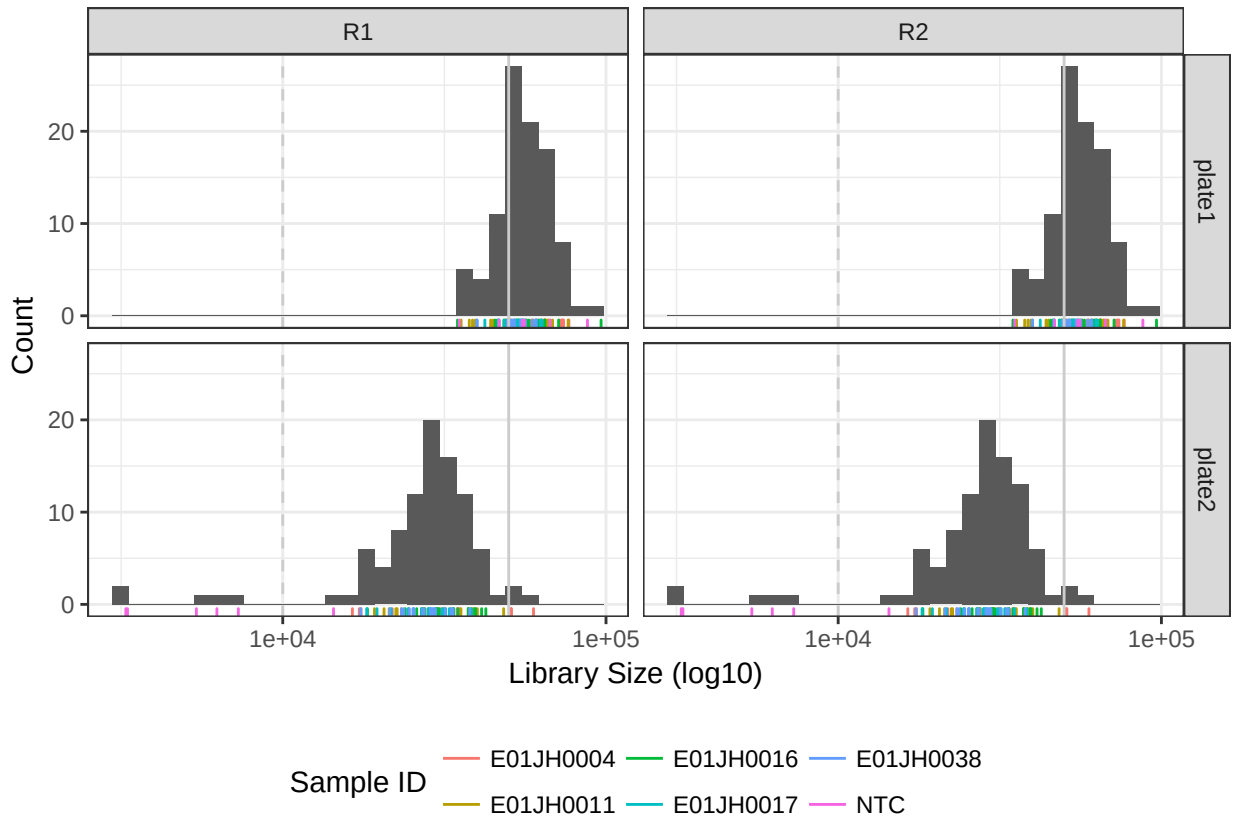

Figure 1: Number of reads per barcoded sample (Library Size), by read direction (X-facet) and replicate 16S PCR plate (Y-facet). Vertical line indicates 50,000 reads per barcoded sample.

Table 2: Barcoded experimental samples

sampleID	dilution	pos	plate	R1	R2
E01JH0004	0	A1	plate1	71860	71860
E01JH0004	0	A1	plate2	35266	35266
E01JH0004	0	A7	plate1	61072	61072
E01JH0004	0	A7	plate2	32894	32894
E01JH0004	1	B1	plate1	59573	59573
E01JH0004	1	B1	plate2	17249	17249
E01JH0004	1	B7	plate1	64436	64436
E01JH0004	1	B7	plate2	31618	31618
E01JH0004	2	C1	plate1	52594	52594
E01JH0004	2	C1	plate2	29650	29650
E01JH0004	2	C7	plate1	67432	67432
E01JH0004	2	C7	plate2	27070	27070
E01JH0004	3	D1	plate1	58732	58732
E01JH0004	3	D1	plate2	16423	16423
E01JH0004	3	D7	plate1	76762	76762
E01JH0004	3	D7	plate2	21820	21820
E01JH0004	4	E1	plate1	49240	49240
E01JH0004	4	E1	plate2	33689	33689
E01JH0004	4	E7	plate1	51090	51090
E01JH0004	4	E7	plate2	50811	50811
E01JH0004	5	F1	plate1	64942	64942
E01JH0004	5	F1	plate2	17470	17470
E01JH0004	5	F7	plate1	61602	61602
E01JH0004	5	F7	plate2	50998	50998
E01JH0004	10	G1	plate1	64192	64192
E01JH0004	10	G1	plate2	31377	31377
E01JH0004	10	G7	plate1	72868	72868
E01JH0004	10	G7	plate2	59670	59670
E01JH0004	15	H1	plate1	73539	73539
E01JH0004	15	H1	plate2	23281	23281
E01JH0004	15	H7	plate1	73458	73458
E01JH0004	15	H7	plate2	34639	34639
E01JH0004	20	F12	plate1	65035	65035
E01JH0004	20	F12	plate2	34823	34823
E01JH0004	20	F6	plate1	73880	73880
E01JH0004	20	F6	plate2	30688	30688
E01JH0011	0	A2	plate1	50775	50775
E01JH0011	0	A2	plate2	39008	39008
E01JH0011	0	A8	plate1	68341	68341
E01JH0011	0	A8	plate2	20576	20576
E01JH0011	1	B2	plate1	39418	39418
E01JH0011	1	B2	plate2	40004	40004
E01JH0011	1	B8	plate1	59445	59445
E01JH0011	1	B8	plate2	27353	27353
E01JH0011	2	C2	plate1	38643	38643
E01JH0011	2	C2	plate2	38601	38601
E01JH0011	2	C8	plate1	62956	62956
E01JH0011	2	C8	plate2	19222	19222
E01JH0011	3	D2	plate1	43962	43962
E01JH0011	3	D2	plate2	22592	22592

sampleID	dilution	pos	plate	R1	R2
E01JH0011	3	D8	plate1	64611	64611
E01JH0011	3	D8	plate2	22379	22379
E01JH0011	4	E2	plate1	35617	35617
E01JH0011	4	E2	plate2	37483	37483
E01JH0011	4	E8	plate1	56718	56718
E01JH0011	4	E8	plate2	30189	30189
E01JH0011	5	F2	plate1	49009	49009
E01JH0011	5	F2	plate2	48242	48242
E01JH0011	5	F8	plate1	66445	66445
E01JH0011	5	F8	plate2	27540	27540
E01JH0011	10	G2	plate1	44712	44712
E01JH0011	10	G2	plate2	35574	35574
E01JH0011	10	G8	plate1	76368	76368
E01JH0011	10	G8	plate2	38216	38216
E01JH0011	15	H2	plate1	37759	37759
E01JH0011	15	H2	plate2	27110	27110
E01JH0011	15	H8	plate1	53435	53435
E01JH0011	15	H8	plate2	21343	21343
E01JH0011	20	B12	plate1	66148	66148
E01JH0011	20	B12	plate2	33485	33485
E01JH0011	20	B6	plate1	65150	65150
E01JH0011	20	B6	plate2	29942	29942
E01JH0016	0	A3	plate1	46042	46042
E01JH0016	0	A3	plate2	30075	30075
E01JH0016	0	A9	plate1	62636	62636
E01JH0016	0	A9	plate2	39260	39260
E01JH0016	1	B3	plate1	48308	48308
E01JH0016	1	B3	plate2	33715	33715
E01JH0016	1	B9	plate1	61939	61939
E01JH0016	1	B9	plate2	34828	34828
E01JH0016	2	C3	plate1	52775	52775
E01JH0016	2	C3	plate2	26846	26846
E01JH0016	2	C9	plate1	57753	57753
E01JH0016	2	C9	plate2	30471	30471
E01JH0016	3	D3	plate1	56100	56100
E01JH0016	3	D3	plate2	18279	18279
E01JH0016	3	D9	plate1	96550	96550
E01JH0016	3	D9	plate2	23936	23936
E01JH0016	4	E3	plate1	45408	45408
E01JH0016	4	E3	plate2	25966	25966
E01JH0016	4	E9	plate1	55373	55373
E01JH0016	4	E9	plate2	41225	41225
E01JH0016	5	F3	plate1	49018	49018
E01JH0016	5	F3	plate2	37501	37501
E01JH0016	5	F9	plate1	34727	34727
E01JH0016	5	F9	plate2	42494	42494
E01JH0016	10	G3	plate1	50329	50329
E01JH0016	10	G3	plate2	30054	30054
E01JH0016	10	G9	plate1	64675	64675
E01JH0016	10	G9	plate2	39511	39511
E01JH0016	15	H3	plate1	54050	54050
E01JH0016	15	H3	plate2	24551	24551

sampleID	dilution	pos	plate	R1	R2
E01JH0016	15	H9	plate1	71348	71348
E01JH0016	15	H9	plate2	27407	27407
E01JH0016	20	C12	plate1	56801	56801
E01JH0016	20	C12	plate2	33814	33814
E01JH0016	20	C6	plate1	52913	52913
E01JH0016	20	C6	plate2	37530	37530
E01JH0017	0	A10	plate1	53711	53711
E01JH0017	0	A10	plate2	33642	33642
E01JH0017	0	A4	plate1	53059	53059
E01JH0017	0	A4	plate2	NA	35128
E01JH0017	1	B10	plate1	52034	52034
E01JH0017	1	B10	plate2	28869	28869
E01JH0017	1	B4	plate1	53726	53726
E01JH0017	1	B4	plate2	28264	28264
E01JH0017	2	C10	plate1	55182	55182
E01JH0017	2	C10	plate2	28150	28150
E01JH0017	2	C4	plate1	46486	46486
E01JH0017	2	C4	plate2	19567	19567
E01JH0017	3	D10	plate1	55461	55461
E01JH0017	3	D10	plate2	18183	18183
E01JH0017	3	D4	plate1	49620	49620
E01JH0017	3	D4	plate2	21658	21658
E01JH0017	4	E10	plate1	53547	53547
E01JH0017	4	E10	plate2	29660	29660
E01JH0017	4	E4	plate1	42191	42191
E01JH0017	4	E4	plate2	28504	28504
E01JH0017	5	F10	plate1	63104	63104
E01JH0017	5	F10	plate2	29666	29666
E01JH0017	5	F4	plate1	46463	46463
E01JH0017	5	F4	plate2	31291	31291
E01JH0017	10	G10	plate1	62711	62711
E01JH0017	10	G10	plate2	32777	32777
E01JH0017	10	G4	plate1	54836	54836
E01JH0017	10	G4	plate2	31087	31087
E01JH0017	15	H10	plate1	61048	61048
E01JH0017	15	H10	plate2	26876	26876
E01JH0017	15	H4	plate1	46671	46671
E01JH0017	15	H4	plate2	24423	24423
E01JH0017	20	E12	plate1	63416	63416
E01JH0017	20	E12	plate2	38374	38374
E01JH0017	20	E6	plate1	48459	48459
E01JH0017	20	E6	plate2	31935	31935
E01JH0038	0	A11	plate1	54531	54531
E01JH0038	0	A11	plate2	28651	28651
E01JH0038	0	A5	plate1	51726	51726
E01JH0038	0	A5	plate2	27079	27079
E01JH0038	1	B11	plate1	54991	54991
E01JH0038	1	B11	plate2	28975	28975
E01JH0038	1	B5	plate1	52119	52119
E01JH0038	1	B5	plate2	29668	29668
E01JH0038	2	C11	plate1	50790	50790
E01JH0038	2	C11	plate2	31813	31813

sampleID	dilution	pos	plate	R1	R2
E01JH0038	2	C5	plate1	46523	46523
E01JH0038	2	C5	plate2	25432	25432
E01JH0038	3	D11	plate1	56321	56321
E01JH0038	3	D11	plate2	21431	21431
E01JH0038	3	D5	plate1	39957	39957
E01JH0038	3	D5	plate2	17425	17425
E01JH0038	4	E11	plate1	60634	60634
E01JH0038	4	E11	plate2	33158	33158
E01JH0038	4	E5	plate1	39838	39838
E01JH0038	4	E5	plate2	25270	25270
E01JH0038	5	F11	plate1	54471	54471
E01JH0038	5	F11	plate2	23311	23311
E01JH0038	5	F5	plate1	50380	50380
E01JH0038	5	F5	plate2	23688	23688
E01JH0038	10	G11	plate1	60462	60462
E01JH0038	10	G11	plate2	29247	29247
E01JH0038	10	G5	plate1	49480	49480
E01JH0038	10	G5	plate2	27372	27372
E01JH0038	15	H11	plate1	51962	51962
E01JH0038	15	H11	plate2	24616	24616
E01JH0038	15	H5	plate1	60719	60719
E01JH0038	15	H5	plate2	25446	25446
E01JH0038	20	G12	plate1	59139	59139
E01JH0038	20	G12	plate2	37870	37870
E01JH0038	20	G6	plate1	51688	51688
E01JH0038	20	G6	plate2	33575	33575

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 the first sequencing run and a 16S public dataset on basespace (16S-Metagenomic-Library-Prep run id 3861867). The error rate for the second run was lower for both R1 and R2 compared to the first run but still higher than the error rate for the public dataset. **NOTE** Not sure whether or not to include in publication as supplemental material.

Base Quality Score

Read BQ

Differences in forward and reverse read average base quality score distributions consistent between replicate plates. A distinct population of barcoded datasets, NTC vs experimental samples, with a higher proportion of lower base quality scores for forward read datasets. For reverse reads the popultion of datasets with lower base quality scores is more hetergeneous.

Cycle BQ

Cycle base quality score is more homogeneous from PCR plate 2 samples than plate 1. 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. **NOTE** Figure out which sample is the low reverse read for plate 1 sample id 17.

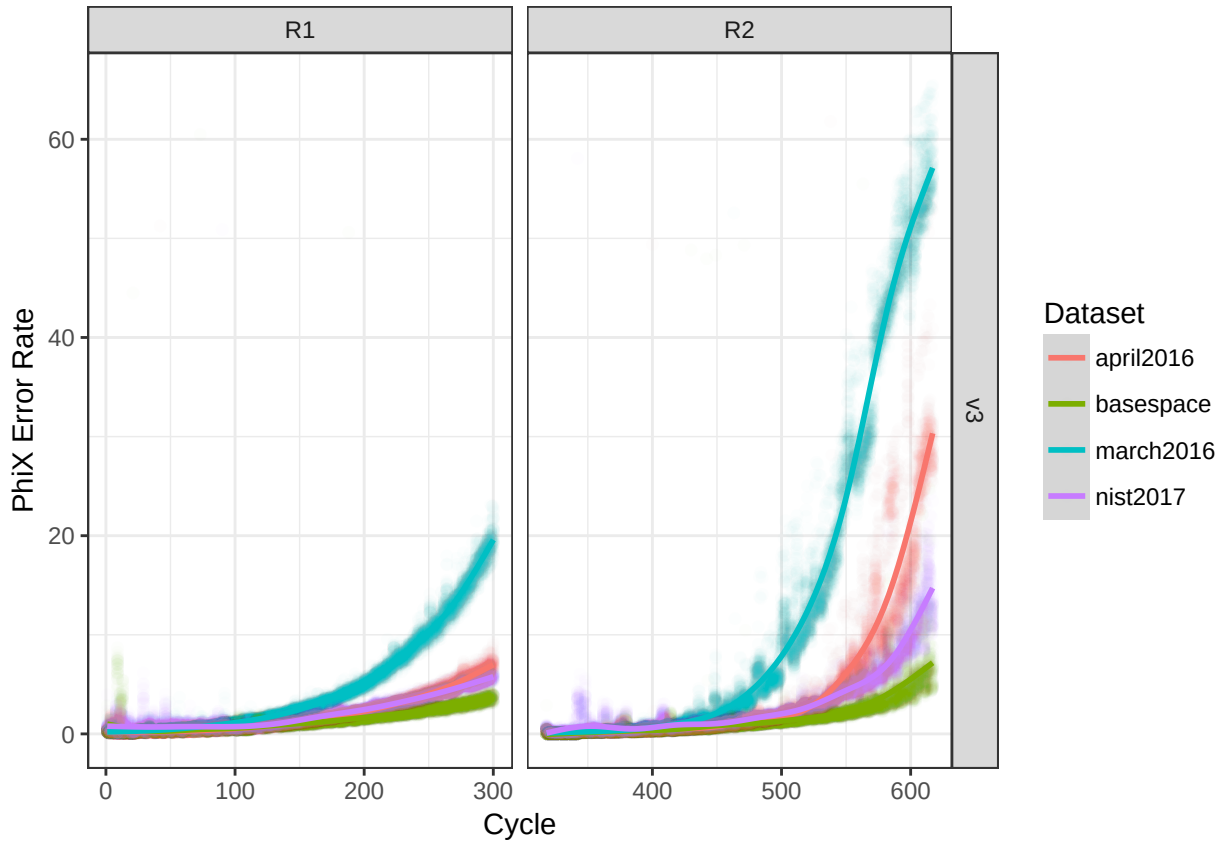

Figure 2: PhiX error rate for initial and reseuqencing of JHU barcoded samples compared to the public dataset.

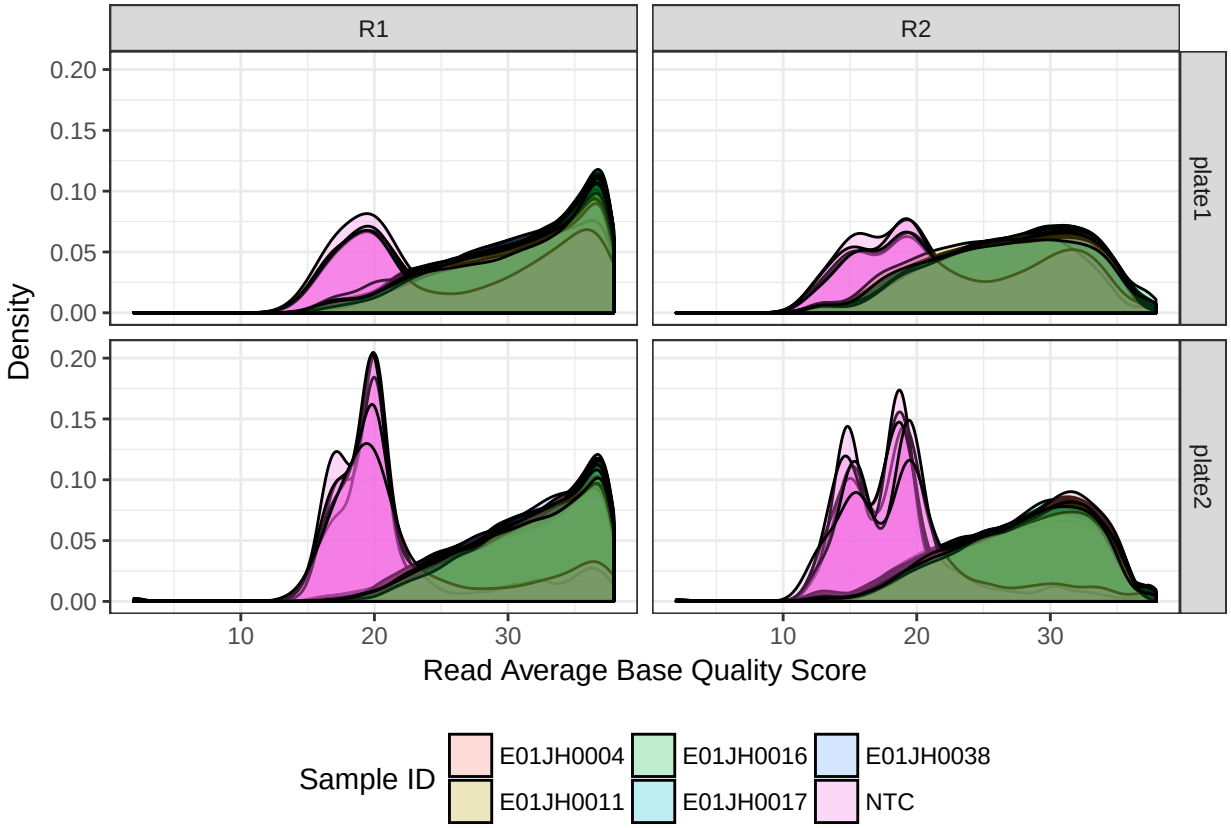

Figure 3: Distribution of base quality scores per barcoded samples.

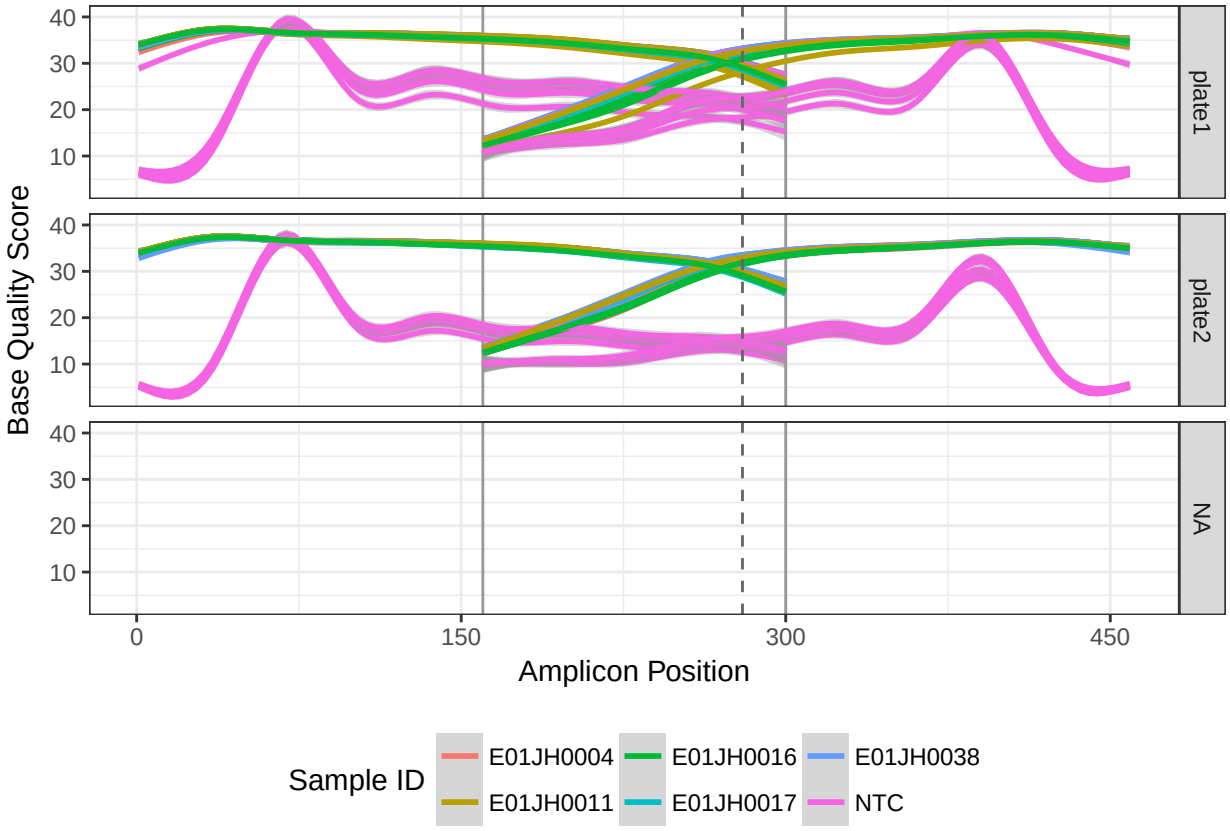

Figure 4: Smoothing spline of the base quality score by sequencing cycle. Vertical lines indicate approximate overlap region between forward and reverse reads. This is not a read level analysis but average quality score for individual barcoded datasets.