Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ShortRead Quality Assessment

Overview

This document provides a quality assessment of Genome Analyzer results. The assessment is meant to complement, rather than replace, quality assessment available from the Genome Analyzer and its documentation. The narrative interpretation is based on experience of the package maintainer. It is applicable to results from the 'Genome Analyzer' hardware single-end module, configured to scan 300 tiles per lane. The 'control' results refered to below are from analysis of PhiX-174 sequence provided by Illumina.

Run Summary

Subsequent sections of the report use the following to identify figures and other information.

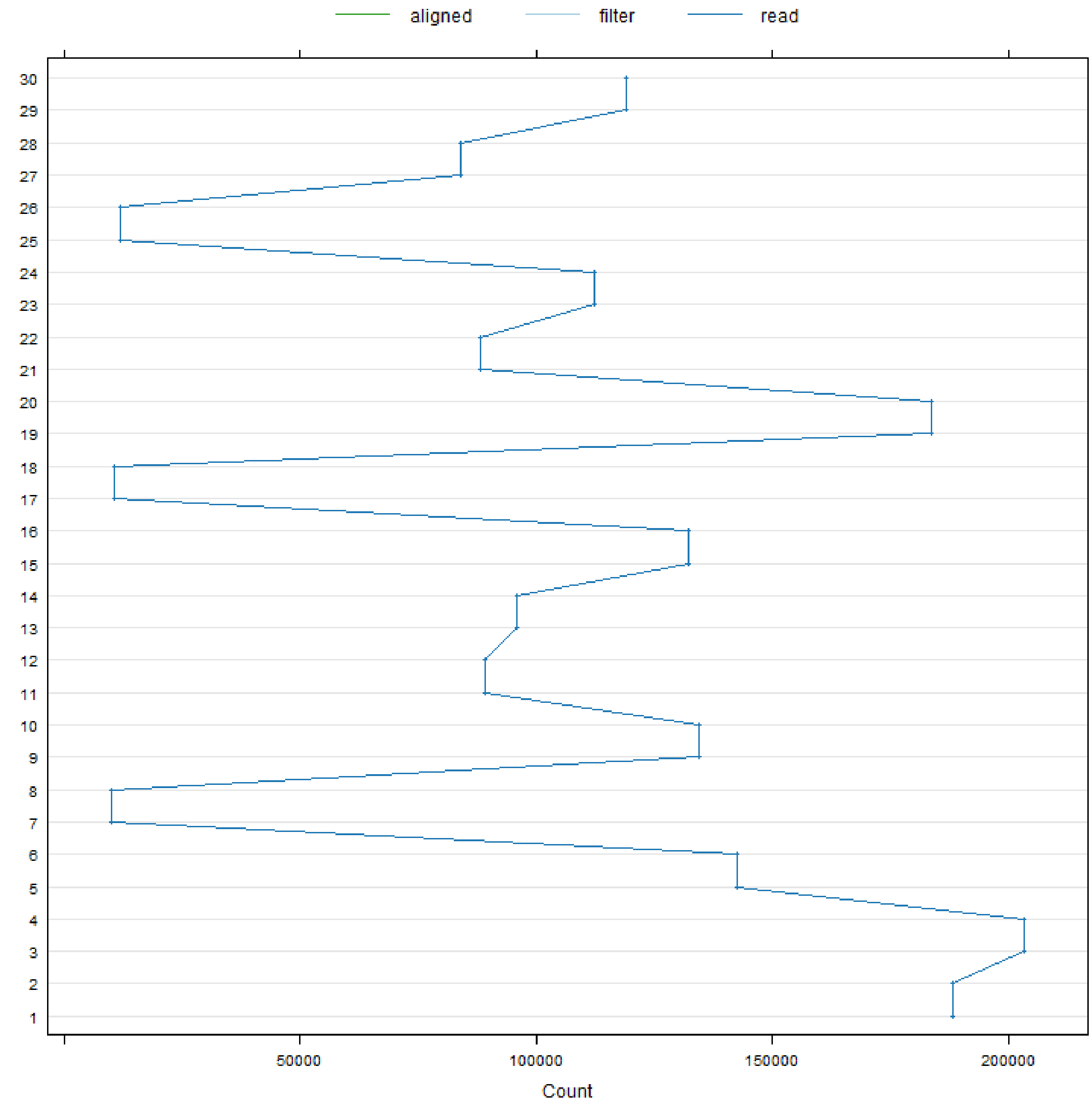
	Key
ANA2-KAPA_S14_L001_R1_001.fastq	1
ANA2-KAPA_S14_L001_R2_001.fastq	2
ANA2-Q5_S13_L001_R1_001.fastq	3
ANA2-Q5_S13_L001_R2_001.fastq	4
C-EXTRA_S5_L001_R1_001.fastq	5
C-EXTRA_S5_L001_R2_001.fastq	6
C-FIELD_S7_L001_R1_001.fastq	7
C-FIELD_S7_L001_R2_001.fastq	8
C-POOL_S6_L001_R1_001.fastq	9
C-POOL_S6_L001_R2_001.fastq	10
C1_S3_L001_R1_001.fastq	11
C1_S3_L001_R2_001.fastq	12
C3_S4_L001_R1_001.fastq	13
C3_S4_L001_R2_001.fastq	14
D-EXTRA_S10_L001_R1_001.fastq	15
D-EXTRA_S10_L001_R2_001.fastq	16
D-FIELD_S12_L001_R1_001.fastq	17
D-FIELD_S12_L001_R2_001.fastq	18
D-POOL_S11_L001_R1_001.fastq	19
D-POOL_S11_L001_R2_001.fastq	20
D1_S8_L001_R1_001.fastq	21
D1_S8_L001_R2_001.fastq	22
D3_S9_L001_R1_001.fastq	23
D3_S9_L001_R2_001.fastq	24
EXTRACTION_S2_L001_R1_001.fastq	25
EXTRACTION_S2_L001_R2_001.fastq	26
FILTRATION_S1_L001_R1_001.fastq	27
FILTRATION_S1_L001_R2_001.fastq	28
Undetermined_S0_L001_R1_001.fastq	29
Undetermined_S0_L001_R2_001.fastq	30

Read counts. Filtered and aligned read counts are reported relative to the total number of reads (clusters; if only filtered or aligned reads are available, total read count is reported). Consult Genome Analyzer documentation for official guidelines. From experience, very good runs of the Genome Analyzer 'control' lane result in 25-30 million reads, with up to 95% passing pre-defined filters.

```
ShortRead:::.ppnCount(qa[["readCounts"]])
```

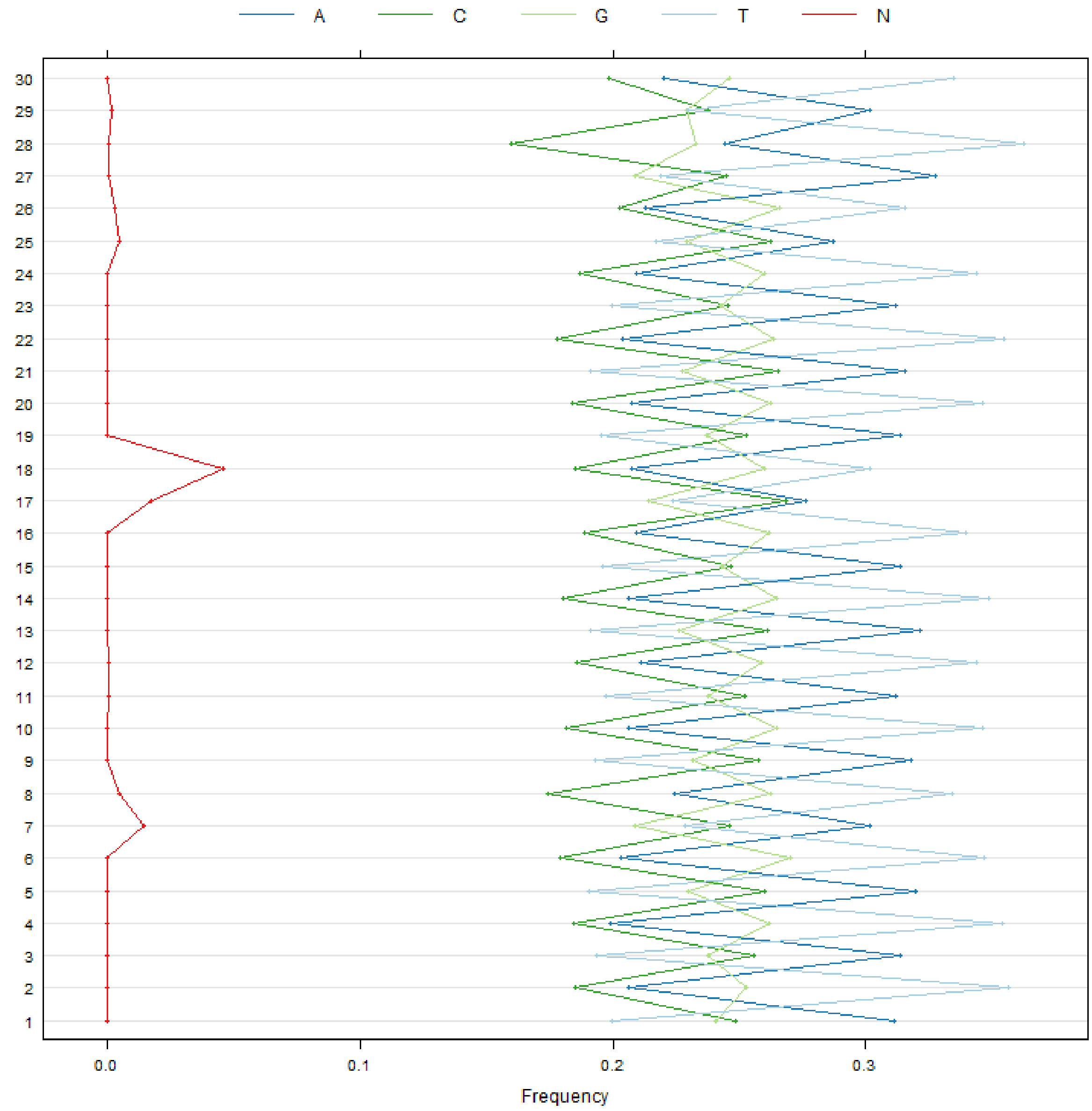
```
     read     filter   aligned
1    188154
2    188154
3    203311
4    203311
5    142434
6    142434
7    9933
8    9933
9    134453
10   134453
11   89233
12   89233
13   95836
14   95836
15   132332
16   132332
17   10658
18   10658
19   183581
20   183581
21   88114
22   88114
23   112235
24   112235
25   11837
26   11837
27   84077
28   84077
29   119143
30   119143
```

```
ShortRead:::.plotReadCount(qa)
```



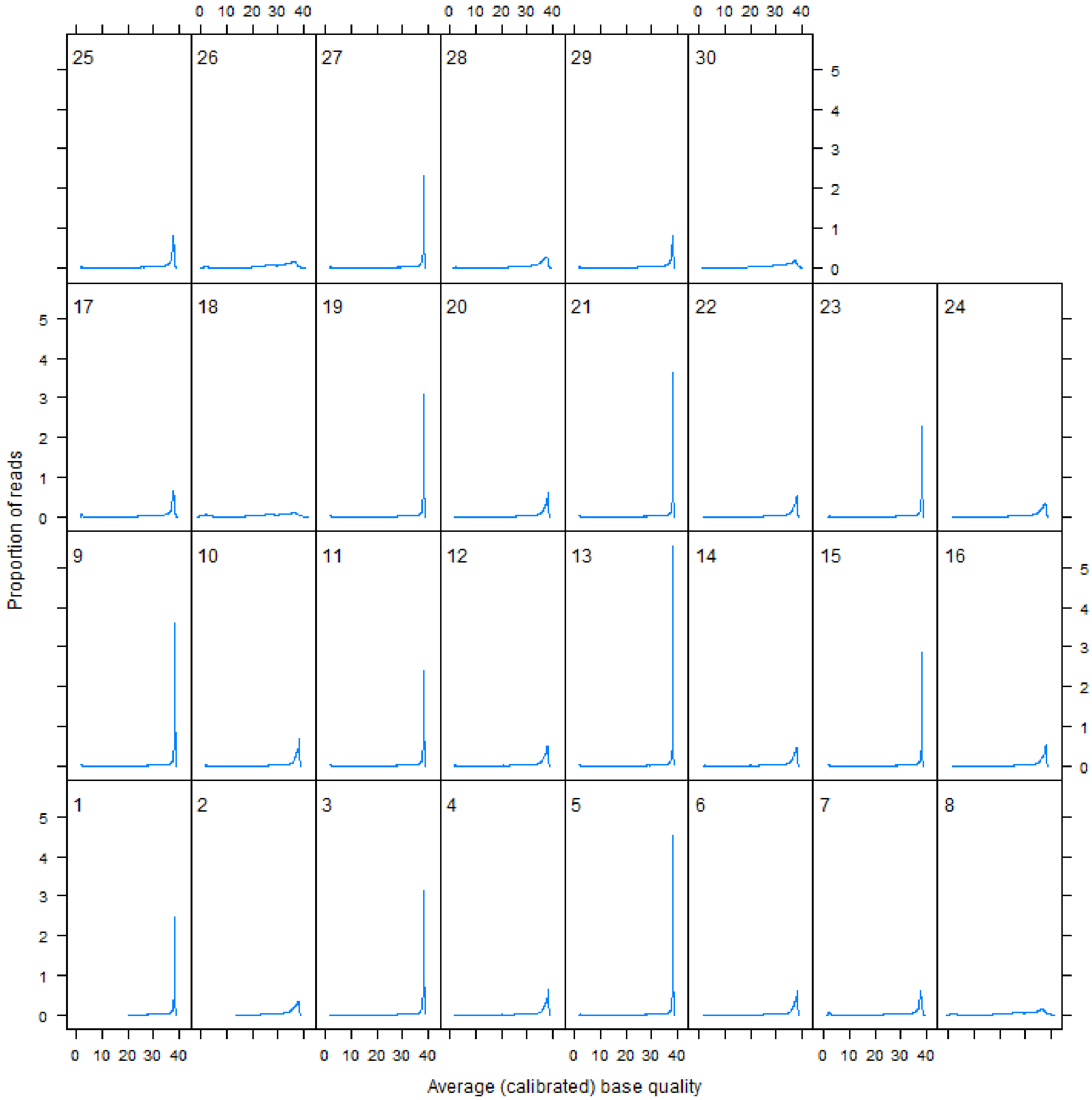
Base call frequency over all reads. Base frequencies should accurately reflect the frequencies of the regions sequenced.

```
ShortRead:::.plotNucleotideCount(qa)
```



Overall read quality. Lanes with consistently good quality reads have strong peaks at the right of the panel.

```
df <- qa[["readQualityScore"]]
ShortRead:::.plotReadQuality(df[df$type=="read",])
```



Read Distribution

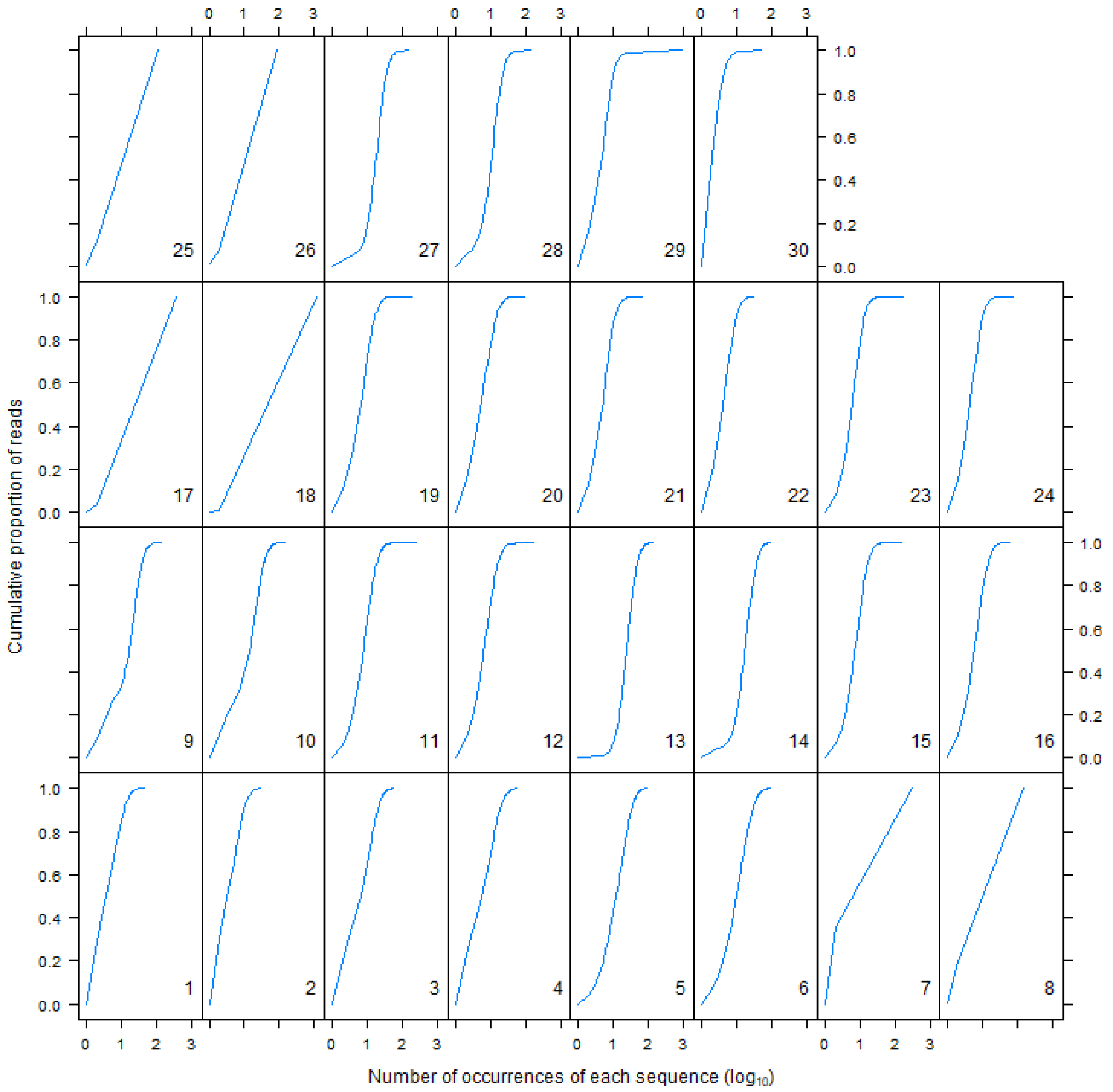
These curves show how coverage is distributed amongst reads. Ideally, the cumulative proportion of reads will transition sharply from low to high.

Portions to the left of the transition might correspond roughly to sequencing or sample processing errors, and correspond to reads that are represented relatively infrequently. 10-15%; of reads in a typical Genome Analyzer 'control' lane fall in this category.

Portions to the right of the transition represent reads that are over-represented compared to expectation. These might include inadvertently sequenced primer or adapter sequences, sequencing or base calling artifacts (e.g., poly-A reads), or features of the sample DNA (highly repeated regions) not adequately removed during sample preparation. About 5% of Genome Analyzer 'control' lane reads fall in this category.

Broad transitions from low to high cumulative proportion of reads may reflect sequencing bias or (perhaps intentional) features of sample preparation resulting in non-uniform coverage. the transition is about 5 times as wide as expected from uniform sampling across the Genome Analyzer 'control' lane.

```
df <- qa[["sequenceDistribution"]]
ShortRead:::.plotReadOccurrences(df[df$type=="read",], cex=.5)
```



Common duplicate reads might provide clues to the source of over-represented sequences. Some of these reads are filtered by the alignment algorithms; other duplicate reads might point to sample preparation issues.

```
ShortRead:::.freqSequences(qa, "read")

sequence
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GGGGGCGTCGGTAAAACTCGTGCCAGCCACCGCGGTTATACGAGAGACCCAAGTTGACAGACAATCGGCGTAAAGAGTGGTTAAGTACTATACCCTACTAAAGCCAAACACCTTCAAAGCTGTTA
```

```
GGGGGGCATAGTGGGGTATCTAATCCCAGTTTGTGCCCTAGCTTTCGTGGGTTCAGTAAGTTTAAAGCCACTTTCGTGATTGGGCTTCTAACCTTCGGGTGCGTATAACAGCTTTGAAGGTGTTT
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GGGTGCGTCGGTAAAACTCGTGCCAGCCACCGCGGTTATACGAGAGACCCAAGTTGACAGACAATCGGCGTAAAGAGTGGTTAAGTACTATACCCTACTAAAGCCAAACACCTTCAAAGCTGTTA
GGGGGTGTCGGTAAAACTCGTGCCAGCCACCGCGGTTATACGAGAGACCCAAGTTGACAGACAATCGGCGTAAAGAGTGGTTAAGTACTATACCCTACTAAAGCCAAACACCTTCAAAGCTGTTA
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Common duplicate reads after filtering

```
ShortRead:::.freqSequences(qa, "filtered")
```

```
NA
```

Common aligned duplicate reads are

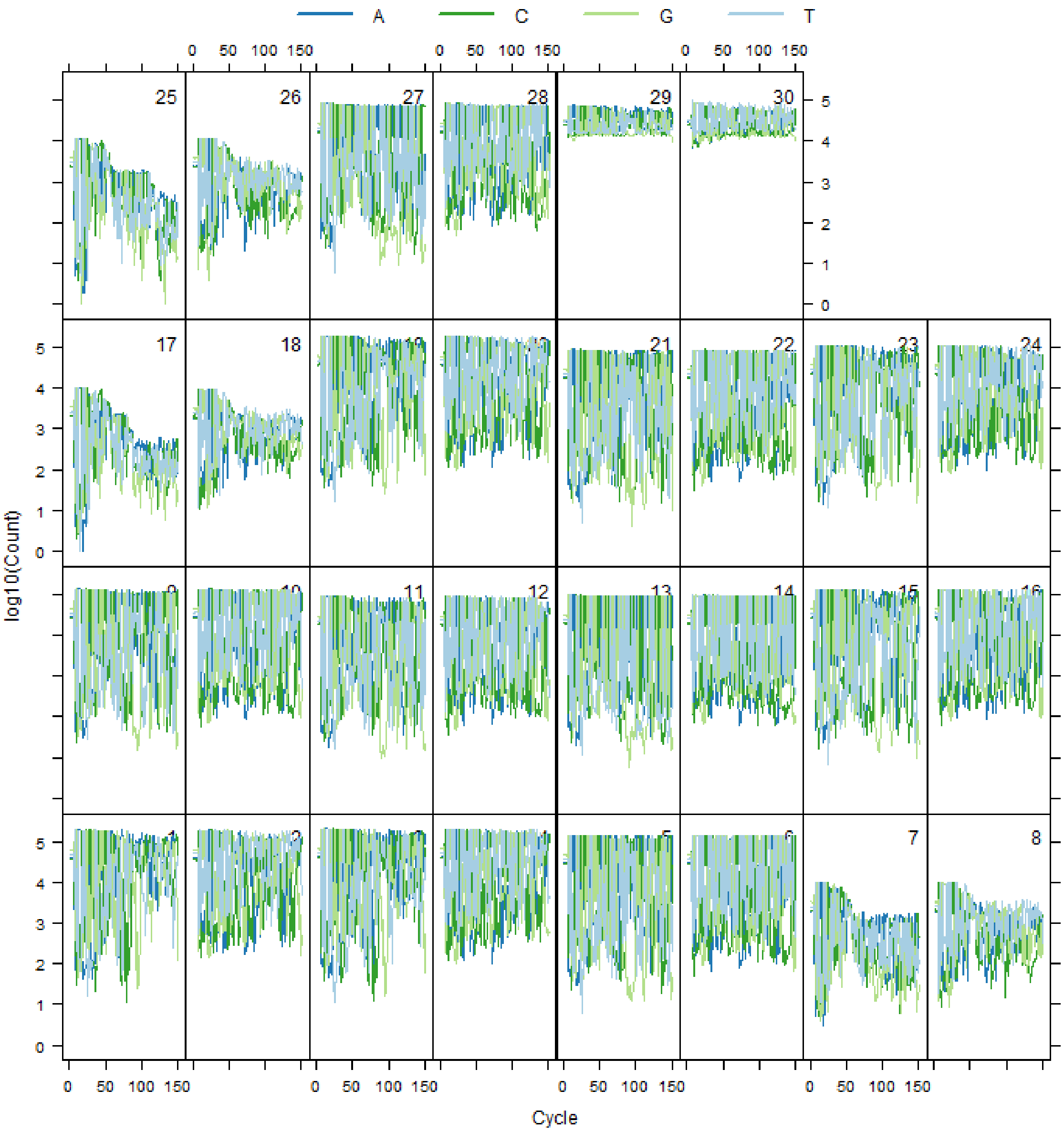
```
ShortRead:::.freqSequences(qa, "aligned")
```

```
NA
```

Cycle-Specific Base Calls and Read Quality

Per-cycle base call should usually be approximately uniform across cycles. Genome Analyzer `control' lane results often show a deline in A and increase in T as cycles progress. This is likely an artifact of the underlying technology.

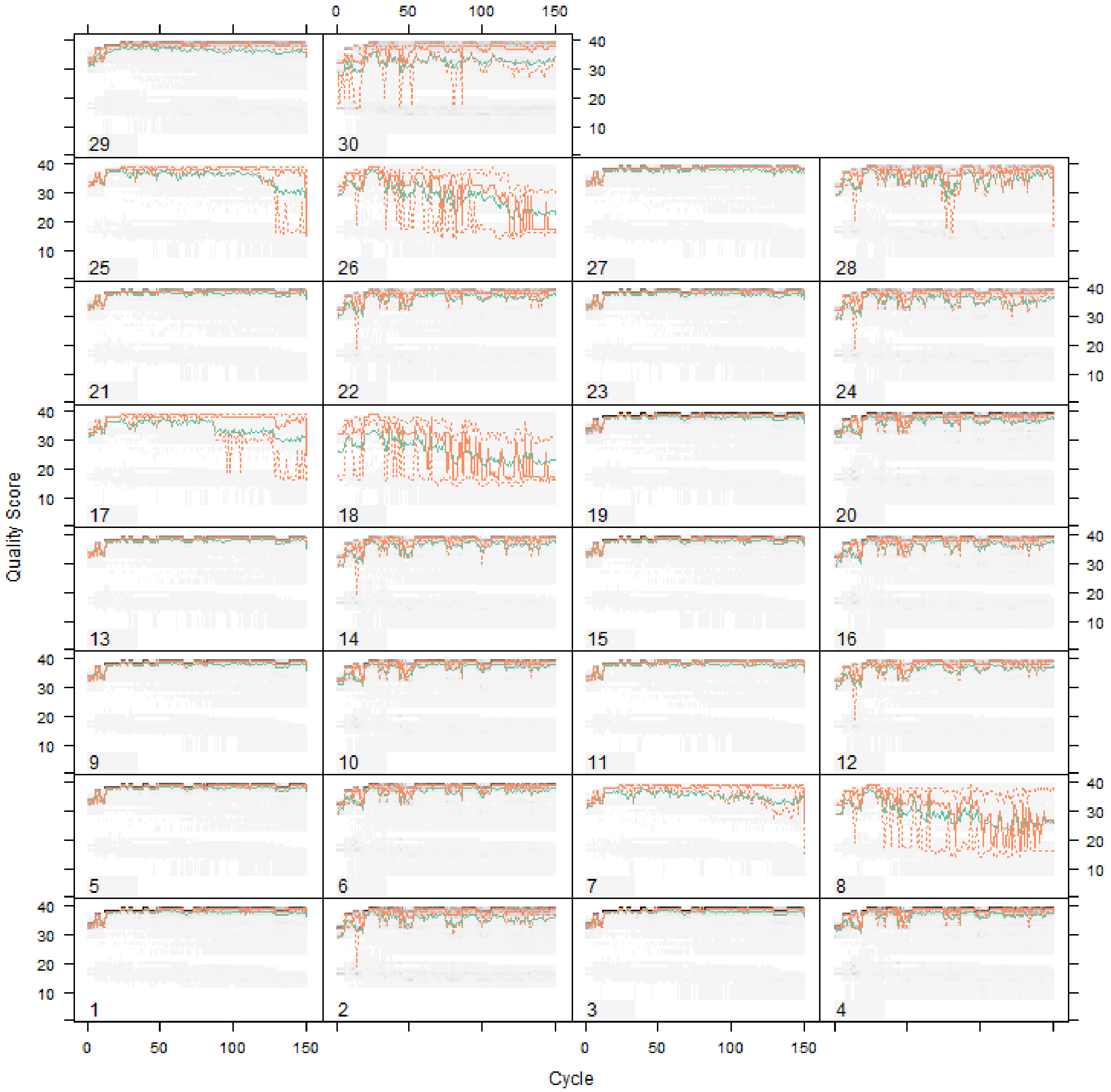
```
perCycle <- qa[["perCycle"]]
ShortRead:::.plotCycleBaseCall(perCycle$baseCall)
```



Per-cycle quality score. Reported quality scores are `calibrated', i.e., incorporating phred-like adjustments following sequence alignment. These typically decline with cycle, in an accelerating manner. Abrupt transitions in quality between cycles toward the end of the read might result when only some of the cycles are used for alignment: the cycles included in the alignment are calibrated more effectively than the reads excluded from the alignment.

The reddish lines are quartiles (solid: median, dotted: 25, 75), the green line is the mean. Shading is proporitional to number of reads.

```
  perCycle <- qa[["perCycle"]]
  ShortRead:::.plotCycleQuality(perCycle$quality)
```



Adapter Contamination

Adapter contamination is defined here as non-genetic sequences attached at either or both ends of the reads. The 'contamination' measure is the number of reads with a right or left match to the adapter sequence over the total number of reads. Mismatch rates are 10% on the left and 20% on the right with a minimum overlap of 10 nt.

```
ShortRead:::.ppnCount(qa[["adapterContamination"]])
```

```
     contamination
1    NA
2    NA
3    NA
4    NA
5    NA
6    NA
7    NA
8    NA
9    NA
10   NA
11   NA
12   NA
13   NA
```

```
  14   NA
  15   NA
  16   NA
  17   NA
  18   NA
  19   NA
  20   NA
  21   NA
  22   NA
  23   NA
  24   NA
  25   NA
  26   NA
  27   NA
  28   NA
  29   NA
  30   NA
```

Wed Jun 22 13:51:35 2022; ShortRead v. 1.54.0
Report template: Martin Morgan