# Pilot: Benefits of including quality for both reads individually as well as together

## Methods and Results

Sometimes there can be a noticeable problem with one of the reads if the data is inspected manually, where a number of cycles have a low proportion of clusters over Q30. This can particularly manifest towards the end of a read, and this can be indicative of issues with a sequencing run (e.g. reagents being used up or having degraded) or of the sequencing instrument itself (e.g. the M3 mirror mount requiring servicing). These cases should be identified as poor quality sequencing can lead to errors at the variant calling stage and an inaccurate report being delivered to patients. Often when this situation occurs, quality in the other read is sufficiently high to compensate for the lower quality cycles seen in the read, and the poor quality sequencing is missed.

A pilot study was undertaken to assess whether this impacted on the accuracy of automated QC and whether inclusion of a separate test on each read for paired end sequencing could improve automated QC performance. Method 1 did not include quality tests on individual reads, while Method 2 incorporated a test for each read individually (proportion of clusters over Q30), where sequencing was paired end (with index reads excluded). Runs where either of the reads fell under the threshold were considered to fail QC. This metric was chosen initially as it had a threshold specified by the instrument manufacturer, and therefore it was considered likely to be accurate and informative(1,2).

Table 1 Metrics included in initial automated QC: run quality

|  |  |  |
| --- | --- | --- |
| Method 1 metrics | Method 2 metrics | Times triggered |
| Percentage of clusters passing filter exceeds manually set threshold (passes) | Percentage of clusters passing filter exceeds manually set threshold (passes) | 11 (9%) |
| Coefficient of variation of cluster density exceeds manually set threshold (fails) | Coefficient of variation of cluster density exceeds manually set threshold (fails) | 10 (8%) |
| Cluster density passing filter median exceeds a manually set threshold from the median of total cluster density (fails) | Cluster density passing filter median exceeds a manually set threshold from the median of total cluster density (fails) | 19 (15%) |
| Proportion of clusters over Q30 for both reads exceeds an Illumina set threshold (passes) | Proportion of clusters over Q30 for both reads exceeds an Illumina set threshold (passes) | 9 (7%) |
|  | Proportion of clusters over Q30 for read 1 exceeds an Illumina set threshold (passes) | 5 (5%) |
|  | Proportion of clusters over Q30 for read 2 exceeds an Illumina set threshold (passes) | 16 (13%) |

Table 1: The initial quality metrics used to automate quality control. Method 1 does not incorporate any information about individual reads, while this is incorporated in Method 2. The number of times each metric was triggered and percentage of the runs for which each was triggered (rounded to the nearest whole number) is also recorded. Manually set thresholds were set on consultation with staff members and will need evaluation and refinement. Illumina set thresholds were specified by the manufacturer of the sequencer. Brackets indicate if exceeding the threshold means that the run passes or fails QC(1,3,2).

Table 2 Results from automated QC of data: run quality

|  |  |  |
| --- | --- | --- |
| Result | Number of runs | Percentage of runs  (approximate) |
| Method 1 Pass | 94 | 77% |
| Method 2 Pass | 87 | 71% |
| Method 1 Fail | 28 | 23% |
| Method 2 Fail | 35 | 29% |

Table 2: The number and percentage (rounded to the nearest integer) of runs which passed and failed the automated data quality control as described above. Method 2 was the same as Method 1 but involved also considering whether the proportion of clusters over Q30 for each read individually fell below the threshold set by Illumina (paired end sequencing only)(1,2). In this case, runs were considered to fail if either read 1 or read 2 failed to reach the required threshold. Raw data available in Appendix 1.

The results of the manual QC check and the automated QC check for both Method 1 and Method 2 were compared. The manual QC check was used as a truth set to evaluate the performance of the automated QC. The performance of Method 1 and Method 2 are shown below in Table 3 (raw data in GitHub- <https://github.com/smrfer/QCProject>/RawData/).

Table 3 Performance of automated QC methods: run quality

|  |  |  |
| --- | --- | --- |
|  | Method 1 | Method 2 |
| **Concordant results** | 98 (80%) | 101 (83%) |
| **Discordant results** | 24 (20%) | 21 (17%) |
|  |  |  |
| **False positives (incorrect fails)** | 12 | 14 |
| **False negatives (incorrect passes)** | 12 | 7 |
| **True positives (correct fails)** | 16 | 21 |
| **True negatives (correct passes)** | 82 | 80 |
| **Sensitivity** | 57.14%  (37.18%-75.54%) | 75.00%  (55.13%-89.31%) |
| **Specificity** | 87.23%  (78.76%-93.23%) | 87.23%  (78.76%-93.23%) |

Table 3: The number and percentage of results for which the manual and automated quality control methods were in agreement and disagreement for Methods 1 and 2. Percentages rounded to the nearest whole number. The sensitivity and specificity (reported to 2 decimal places) of each test was calculated from the number of true and false positives and negatives, with a positive result taken to be a failed run, as this is what the quality control check is aiming to achieve. The 95% confidence interval values are shown in brackets (Clopper-Pearson method used)(4).

It can be seen that Method 2 outperforms Method 1, with sensitivity increasing from 57.14% to 75.00%, and specificity remaining the same indicating that there is a benefit to quality control in evaluating each read individually for paired end sequencing as well as together.

Although quality control performance with Method 2 is better than Method 1, it does not perform sufficiently well to be implemented and therefore requires further refinement.

## Conclusions

There is benefit to including tests on each read individually for paired end sequencing as performance of automated QC is improved.

Additionally, runs which are problematic but are not identified by the current method of manual QC may also be identified by this method, as the data is being inspected from a different perspective. The consequences of such fails, how to categorise the run (outright fail, be wary, no problem with run but problem with sequencer) remain to be determined in collaboration with colleagues.

## References

1. Illumina Inc. MiSeq Specifications [Internet]. 2016 [cited 2016 Feb 12]. Available from: http://www.illumina.com/systems/miseq/performance\_specifications.html

2. Illumina Technical Assistance. Sequencing Analysis Viewer v1.10 Software Guide. (October 2015).

3. Illumina Technical Assistance. Cluster Density Specifications for Illumina Sequencing Platforms. 2014;

4. Field A, Miles J, Field Z. Discovering Statistics Using R. London, UK: SAGE Publications Ltd; 2012.