Establishing quality control metrics for B cell receptor analysis using the next generation sequencing technologies

Moore Jr, Rodney, CDER; Zheng, Lily, CDER; Poslavsky, Stanislav, Milaboratories, Zhang, Yong, CDER; Mazor, Ronit, CBER; Chu, Shiang-Lin, CBER; Song, Chen, New England Biolabs; Erijman, Ariel, New England Biolabs; Langhorst, Bradley, New England Biolabs; Liu, Pingfang, New England Biolabs; Barry, Andrew, New England Biolabs; Zhang, Douglas, AbHelix; Farmer, Andrew, Takara Bio; Goyal, Shaveta, Takara Bio; Baines, Andrea, CDER; Westreich, Sam, PrecisionFDA; Busby, Ben, PrecisionFDA; Lan, Stanley, PrecisionFDA; Serang, Omar, PrecisionFDA; Johanson, Elaine, PrecisionFDA; Xiao, Wenming, CDER

Abstract

With the rapid development and adaptation of next-generation sequencing (NGS) technologies, we have the unprecedented opportunity to interrogate many fundamental and transformative questions surrounding the adaptive immune system and human disease. The convergence of high-throughput sequencing technologies, novel analysis methods, and advancements in immunotherapy and drug development has set the stage for standardizing the quantitative study of human immune cell receptor repertoire composition and diversity. Recently, to increase the sensitivity and specificity of screens, next-generation sequencing (NGS) has been evaluated to detect minimal residual disease (MRD) and to enable MRD-guided treatment decisions. However, studies by FDA/CDER MRD project team indicated that nearly 50% of MRD data in NDA/BLA submission between 2014-2021 were not included in U.S. Prescribing Information (USPI) due to concerns on assay performance and test validation. This study is to establish quality control (QC) metrics to ensure the performance of assay, sequencing technology, and analytical tools for B cell receptor (BCR) sequence analysis. With a comprehensive workflow on PrecisionFDA, we established QC metrics at three levels: 1) sequence reads; 2) Unique molecular index (UMI); 3) sequence alignment. These metrics are essential for meaningful downstream analysis and ensuring the performance of BCR sequencing assays and platform in detection accuracy and sensitivity. Detailed components of QC workflow and results from our BCR pilot study will be presented in this poster.

Introduction

Next-generation sequencing (NGS) is bioinformatic technology used to determine genetic (DNA/RNA) sequences usually for the purpose of identifying genetic variations, disease and disease treatments, other biological characteristics. Therefore, it is critical to have high-quality data to ensure proper down stream analyses. To ensure the quality of our data, we have determined 5 key metrics at the reads level.

Q30
Percentage of bases with a quality score ≥ 30

Reads redundancy
Percentage of unique reads

GC content
Percentage of G,C nucleotides present in sequence.

% passing filter
Percent of reads that pass filtration process

# of reads
Number of reads present in the sample

Figure 1. Read-level metric node display illustrating the relevant parameters recorded and evaluated at the 1st level of the qc metrics pipeline.The Q30 parameter measures the quality score of a sequencing read by measuring the probability of an incorrect base . Reads redundancy highlights the uniqueness amongst the reads to reduce duplication coverage. An acceptable GC content ranges between 40%-60% as sequences usually have a somewhat similar distribution of nitrogenous bases. Percent passing filter generally should be above 80%. There should be a sufficient number of reads to produce adequate results.

Methods

In the beginning of our pipeline, we created an application within precisionFDA, fastp, a FASTQ data pre-processing tool, to gain insight on our raw sequencing data at the reads level. This application receives FASTQ(s) (paired-end reads or single-read) as an input and outputs both an html and json file displaying relevant metrics visually and tabularly respectively.

As for analyzing the data at the UMI and alignment levels, we will use MiXCR and prestoR, two different bioinformatic tools for analyzing NGS data for immune profiling data. Both applications require FASTQ input(s) and outputs tabular and graphical data with optional inputs for a deeper analysis, similar to fastp. Listed below are all the parameters included in each level of the analysis:

Reads	UMI	Alignment
• Q30 • total # of reads • total # of reads • reads redundancy • GC content • % passing filter	• % of reads with perfect UMI • % of reads with unassigned UMI • error rate in UMI sequence • total # of UMI (molecular fragment) • distribution of reads per UMI (min, median, mean, max), primer/dimer	• reads mapped on BCR • %UMI mapped on BCR • % of reads mapped to other genes/locations • top 3-off target genes/locations • % of reads covering full length of receptor

Figure 2. List of parameters analyzed at each level of the pipeline.

We then will establish a workflow which performs the analysis our data at each level simultaneously. Once all the analyses have been performed, we extract all relevant parameters from the JSON output from each application into one tabular data table.

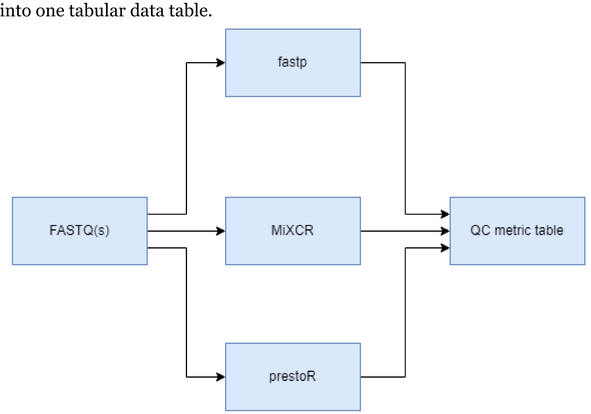

Figure 2. Flowchart illustrating the QC metric workflow available on precisionFDA.

Results and Discussion

We performed a reads level analysis by using fastp on all of the pilot study data which included reads from New England Biolabs (NEB), AbHelix, iRepertoire, Takara Bio, and Upenn. After filtering and trimming the reads, each of our samples are processed to produce our numerical metrics which can be found below in table 1.

Source	Q30 (%)	Reads Redundancy (%)	GC content (%)	% passing filter	# of reads
NEB	73.67	0.007	54	76.61	1.59 M
AbHelix Hifi	99.53	0.9	52.9	100	552.07 K
iRepertoire	54.6	0.016	53.88	87.43	1.56 M
TakaraBio	74.27	0.874	55.08	96.13	3.08 M
Upenn	82.54	11.819	55.27	90.85	359.68 K
AbHelix	93.41	0.414	53.3	99.04	8.48 M

Table 1. Average QC metrics for NEB, AbHelix Hifi, iRepertoire, Takara Bio, Upenn, and AbHelix samples.

Our table highlights a few potential warnings in terms of quality control for some samples. Most notably:

- NEB samples on average have a sub 80% Q30 and passing filter rate. These warnings imply that there may be some base errors or ambiguities across the sequence.
- The avg reads redundancy of Upenn samples are high which may lead to a lack of coverage in the genome.
- With the exception of these outliers, most of the results are consistent with our expectation in terms of quality.

Further investigating the quality of our samples, we analyzed the base quality at each position of the reads. One result is pictured below in figures 3 and 4.

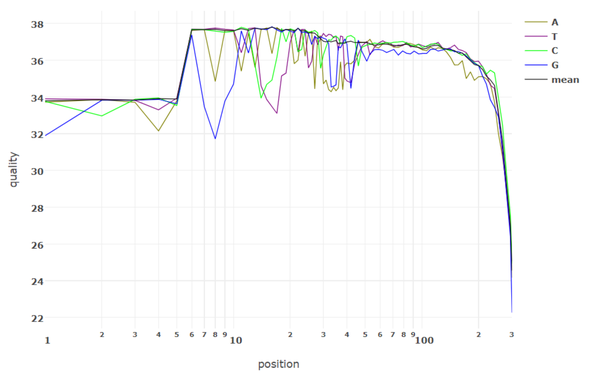

Figure 3. Upenn sample quality after filtering. Notice that the quality of the sequence remains >30% until the sequence nears its end where we notice a significant drop. This behavior is expected due to the sequencing method being used (weaker signal intensity toward end of read).

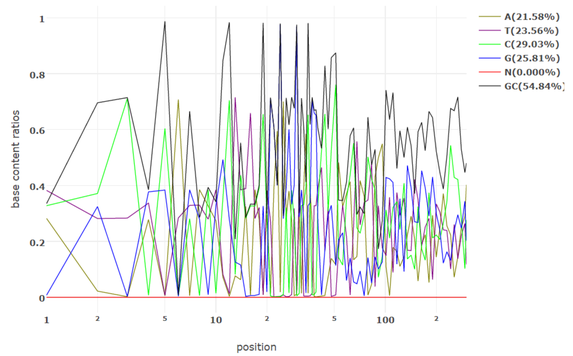

Figure 4. Upenn sample base content ratios at each position of the read. All bases have a similiar frequency ranging between 20%-30% with a GC content of 54.56%.

- Examining the results from Table 1, most of the data is consistent with our expectations. However, the few discussed metrics for the NEB samples are problematic, especially considering that <80% of the reads passed filtration. Coupling the low filtration rate with an equally low Q30 prompts further investigation into the NEB samples.
- Besides NEB, performance across assay type appears to be consistent suggesting that the workflow is not assay-dependent to produce reliable results.
- Of the 5 metrics, % passing filter and Q30 could be considered more critical in determining the overall quality. Probable causes for these failing tests could most likely due to contamination or some other unforeseen error.

In the future

- UMI and alignment levels of the workflow will further assess the quality of our sequencing data.
- It is hypothesized that due to NEB's poor performance at the reads level, that it will also perform sub-optimally at these levels too suggesting that the samples are unsatisfactory to use for downstream analysis.
- During the alignment level, we expect to see more favorable quality bias toward DNA-based reads as RNA-seq reads are considerably more difficult to map to the genome.

Conclusion

- We have decided on 5 key metrics that are most important to determining the quality of our reads. Although each of the discussed metrics provide insight to the quality of the data, together they create a comprehensive profile.
- Moving forward, the qc workflow will be more robust as the UMI and alignment levels are implemented.