



Published in final edited form as:

J Comput Biol. 2026 ; 33(5-6): 705–715. doi:10.1177/15578666261436820.

Assessment of PHRED Score Characteristics in Illumina MiSeq Amplicon Sequencing

SETH SIMS^{1,2}, YURY KHUDYAKOV¹, ALEXANDER ZELIKOVSKY^{2,3}

¹Molecular Epidemiology and Bioinformatics Laboratory, Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia, USA.

²Department of Computer Science, College of Arts and Sciences, Georgia State University, Atlanta, Georgia, USA.

³The Laboratory of Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia.

Abstract

PHRED scores are confidence values associated with each basecall generated by sequencers. The score is defined as a monotonic function of the probability that the basecall is incorrect. The calibration of PHRED scores has previously been examined by evaluating errors made in reading known sequences. We investigated the calibration of the Illumina MiSeq instrument PHRED model using data from a large dataset. We also derive calibration methods for the PHRED scores in datasets similar to those produced by the Global Hepatitis Outbreak and Surveillance Technology (GHOST). The GHOST protocol uses a short amplicon, resulting in many positions having two base calls, one coming from each of the paired reads. A maximum likelihood model of redundant base calls that match each other was used to estimate corrected probabilities of the PHRED scores. The PHRED scores showed only small absolute deviations from their target values. These differences are statistically significant deviations ($p < 0.0001$) from being calibrated. The accuracy of the scores varied significantly between the MiSeq instrument runs. Recalibration produced quality scores that improved Brier scores for the dataset by an average relative improvement of 2.83%. Methods developed to create calibration curves for PHRED scores will be useful in improving error-correction pipelines based on redundant deep sequencing of amplicon data. However, quality scores are relatively uninformative of substitution errors. The quality scores assigned are determined more by the global error rate of the sequencing run in the current machine cycle than by the characteristics of the specific base call.

Address correspondence to: Dr. Seth Sims, Molecular Epidemiology and Bioinformatics Laboratory, Division of Viral Hepatitis, Centers for Disease Control and Prevention, Clifton Road, 33 Gilmer Street SE, Atlanta, GA 30302-3965, USA, xzy3@cdc.gov.

AUTHORS' CONTRIBUTIONS

S.S.: Conceptualization, methodology, software, formal analysis, data curation, writing—original draft, and visualization. Y.K.: Writing—review and editing, project administration, and supervision. A.Z.: Writing—review and editing, project administration, and supervision.

DISCLAIMER

The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the U.S. CDC.

AUTHOR DISCLOSURE STATEMENT

The authors declare that they have no competing interests.

An earlier version of this article appeared on ResearchSquare (<https://doi.org/10.21203/rs.3.rs-6107843/v1>).

Keywords

amplicon sequencing; Illumina MiSeq; paired-end sequencing; PHRED score

1. INTRODUCTION

Viral hepatitis represents a significant public health challenge. In 2017, the National Academies of Science, Engineering, and Medicine released a strategy for the elimination of hepatitis B and C viruses (HBV, HCV) from the United States. Associated recommendations included working with states to improve viral hepatitis screening and information collection (Buckley and Strom, 2017).

The Global Hepatitis Outbreak and Surveillance Technology (GHOST) system is a program of the Centers for Disease Control and Prevention (CDC) to address the enhanced hepatitis screening recommendation. GHOST is an epidemiological decision support and molecular data collection system. The system clusters related cases, allowing the source tracing of hepatitis outbreaks. It also highlights unusual situations useful for epidemiological investigations, such as superinfections of HCV, which are transient phenomena that are only regularly observable in high-risk environments such as injection drug use. GHOST uses a deep-amplicon-based sequencing protocol that targets a region of the HCV genome called hypervariable region 1 (HVR1; Longmire et al., 2017). The sequencing protocol is intended to sample the structure of the intra-host HCV population found in a sample. To do so, we have developed an error-correction and filtering pipeline that uses a 2×300 bp sequencing kit on an amplicon with a design length of 264. Therefore, many bases of each molecule are sequenced twice, once in the first read and once in the second read.

As part of each cycle, the sequencing machine estimates a confidence value known as the PHRED score, which is output along with the base call. Originally developed for Sanger Sequencing, the PHRED score is defined in equation 1 where p is the estimated probability that a base was called incorrectly (Ewing and Green, 1998) (Ewing et al., 1998).

$$Q = \text{Round}(-10 \log p) \quad (1)$$

For such a binary confidence value to be calibrated, the events that are assigned a probability of 30% should have a long-run error frequency of 30% (Dawid, 1982). The most direct method of evaluating the calibration of the PHRED scores involves measuring substitution errors made in a known sequence. This has been used by Schirmer et al. (2015) to measure the calibration of sequencing amplicons on the MiSeq instrument. These experiments are useful in characterizing the MiSeq instrument, but they rely on a known ground truth of the sequences. Such experiments are not commonly performed because they do not produce other information. Examination of the overlap of paired end sequences has been used to evaluate the overall usefulness of quality scores by Eren et al. (2013). Their article developed an error filtering method based on overlapping reads matching across the entire

overlap length. However, they did not investigate the quality scores themselves, instead examining how quality score-based read filtering methods changed the fraction of reads that merged using their criteria. Zhang et al. (2017) developed a method for estimating the PHRED values from raw data generated during HiSeq operation. Their method used runs of phiX174 sequences as known sequences. Illumina MiSeq uses PhiX174 to increase the complexity of the sequences on the chip to improve cluster detection in the raw image data captured from the flow cell. Errors were identified using a short-read alignment to the phiX174 genome. They compared base calls with the known phiX174 sequence, detected errors, and subsequently evaluated their new scoring method with the Illumina HiSeq built-in method. The Zhang et al. method used several features from the raw data used by the on-instrument base calling process, which is not available in our datasets, as we only receive the base calling results.

As GHOST reads purposefully overlap, every sample sequenced by the GHOST protocol provides a natural indirect experiment in the calibration of the PHRED scores. Analyzing the intentional overlap of paired reads allows for the isolation of sequencing errors from errors incurred during library construction. This is a significant advantage over existing methods that conflate the two error sources. GHOST is accepting sequences from multiple MiSeq instruments and has accumulated many runs using the overlapping protocol. These datasets provide an opportunity to measure the distributions of many features of the performance of the Illumina MiSeq instrument under varying conditions.

As an automated system making important public health recommendations, GHOST must be able to determine the quality of the data that motivates these recommendations. The authors have published two articles describing the specifics of sequencing quality monitoring Longmire et al. (2017) and Sims et al. (2018). At multiple points along the quality control pipeline, a sample can be rejected, and a suggestion is returned to the user on what likely caused the rejection and how to improve the sequencing quality. The reads are subjected to steps such as verifying sequence identifiers, verifying the expected primers, merging reads, checking for open reading frames, and a genotyping step, which rejects samples that cannot be assigned clearly to a subtype.

In this study, we intend to assess how accurate the MiSeq's internal error model, its calibration, and its usefulness for discrimination of sequencing errors using observations of overlapped paired-end reads. A maximum likelihood (ML) method of estimating the long-run error rate of sequencing errors indexed by machine-assigned PHRED scores is developed. We find that the PHRED scores are not calibrated, but the size of the error is small in most cases. However, the MiSeq error model does not separate sequencing errors from adjacent correct base calls, even though the assigned PHRED scores reflect the long-term error rate of the base call.

2. MATERIALS AND METHODS

2.1. Data

Sequence data used in our analyses were deposited in the NCBI Short Read Archive (SRA) under BioProject accession PRJNA580030 during 2019–2023 by the GHOST

Team, Division of Viral Hepatitis, CDC. These data were generated using blood samples positive for HCV RNA tested by the GHOST team during 2016–2019 while operating the GHOST system. Blood specimens submitted by various GHOST system users (e.g., state health departments) were sequenced to generate HVR1 gene sequencing data using the Illumina MiSeq sequencing platform and the standard GHOST HVR1 sequencing protocol (Longmire et al., 2017).

2.2. Data cleaning

The reads were filtered by discarding pairs where either read was shorter than 200 bp, contained ambiguous bases, or did not have the expected amplification primer sequences. The paired reads were then aligned to each other with a simple alignment as outlined in Algorithm 1. The minimum overlap length accepted was 10 bp to avoid misassemblies. Read pairs that did not have at least 85% matching base pairs in the overlapping region were discarded. Runs that did not have at least 10,000 reads assembled were discarded.

Algorithm 1: AlignRead, returns the overlap position of a read-pair that minimizes the overlap score.

Input: A read pair produced by the MiSeq sequencer and a minimum allowed overlap length.

Result: The offset defining the overlap with the smallest fraction of mismatches.

```

minOverlap = 10;
bestOverlap = ∞; bestOffset = 0;
for o = 0 to Length(f) - minOverlap do
    h = Hamming(f[o...], r[-o...]);
    c =  $\frac{h}{(l-o)}$ ;
    if c < bestOverlap then
        bestOffset = o;
        bestOverlap = c;
    end
end
return bestOffset, bestOverlap;

```

We ignore indel errors in our alignment of the paired reads, as the MiSeq instrument's technology produces these only rarely. These filtering steps parallel the error correction and filtering methods used in the GHOST pipeline (Longmire et al., 2017).

2.3. Forecast of match from PHRED scores

A probability of base calls matching can be derived given PHRED scores of two overlapping reads, as in Edgar and Flyvbjerg (2015):

$$P(\text{Match} | p_i, p_j) = (1 - p_i)(1 - p_j) + \frac{p_i p_j}{3} \quad (2)$$

where p_i is the probability associated with the forward PHRED score and p_j is the probability of the reverse score associated with the reverse PHRED score. The number of matching versus mismatching base calls can be directly observed from the data. We counted

the fraction of base calls that match across all of our datasets, both globally and collated with forward and reverse PHRED scores.

2.4. Measuring the calibration of MiSeq instrument PHRED scores

PHRED scores serve as indicators of the error rates associated with base calls in DNA sequencing. These scores allow for the calculation of matching probabilities, as described in Equation 2. Consequently, they can also be used to assess the likelihood of overlapping base calls matching. To evaluate the quality of such probabilistic binary forecasts, classification loss functions like the Brier score can be employed (Brier, 1950).

$$BS = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2 \quad o_i = \begin{cases} 0 & \text{basecall pair } i \text{ mismatch} \\ 1 & \text{basecall pair } i \text{ match} \end{cases} \quad (3)$$

where f_i is the forecast probability of matching, o_i is an indicator function, and N is the number of forecasts made. The value of a Brier score does not have a particular meaning in isolation; it is only useful in comparison of different methods to predict a particular system (Schmid and Griffith, 2005). Forecasting methods that produce a smaller Brier score on the same sample are more accurate for the prediction of that system. Given that PHRED scores are discrete and that matching probabilities depend solely on these two scores, our data can be neatly organized into equal bins based on paired PHRED scores. By grouping similar forecasts into mutually exclusive and exhaustive bins, the Brier score can be decomposed into three components, called reliability (REL), resolution (RES), and uncertainty (UNC).

REL is the difference between the forecast in the bin and the observed probability in that bin. It represents the error added by the miscalibration of the forecasting method itself. RES measures how well the forecast method matches events with different probabilities to bins. This is the reduction of the Brier score value by using the forecast over just using the average error of all events. UNC is related to the overall predictability of events. It is the Brier score for the baseline if the long-term frequency of the events is used rather than the forecasting method under study (Murphy, 1973).

Examining the magnitude of the three components can give an impression of the source of error in a forecast. In the case where $RES < REL$, nothing is gained from using the forecast method, and it would be better to simply use the long-term frequency of the events. To calculate the decomposition, events are grouped into k bins with similar forecast probability and sample size N .

$$BS = REL - RES + UNC \quad (4)$$

$$REL = \frac{1}{N} \sum_k n_k (f_k - \bar{o}_k)^2$$

(5)

$$\text{RES} = \frac{1}{N} \sum_k n_k (\bar{o}_k - \bar{o})^2$$

(6)

$$\text{UNC} = \bar{o}(1 - \bar{o})$$

(7)

with $\bar{o}$ being the fraction of times the forecast event occurred for the entire sample, $\bar{o}_k$ the observed fraction of events for the bin k , n_k the number of items in the bin k , and f_k as the forecast probability for the bin.

Spiegelhalter (1986) developed a z-statistic for the Brier score (Eq. 8)). This statistic is used to detect miscalibration of a particular forecasting method.

$$\text{E}[Brier] = \frac{1}{N} \sum_{i=1}^N f_i(1 - f_i)$$

(8a)

$$\text{Var}[Brier] = \frac{1}{N^2} \sum_{i=1}^N f_i(1 - f_i)(1 - 2f_i)^2$$

(8b)

$$Z = \frac{O(Brier) - \text{E}[Brier]}{\sqrt{\text{Var}[Brier]}}$$

(8c)

The Poisson binomial distribution is a discrete probability distribution of the number of independent Bernoulli trials when the probability of success at each trial varies (Wang, 1993). The Poisson binomial distribution is used to estimate the expected number of matches predicted by the PHRED scores of a particular database of reads using equation 9 (Edgar and Flyvbjerg, 2015).

$$\text{E[Match]} = \sum_{i,j} P(\text{Match} | p_i, p_j)$$

(9)

2.5. Maximum likelihood calibration of PHRED probabilities

The method of ML was used to estimate the true value of the PHRED score probabilities. Each pair of PHRED scores observed was assumed to follow an independent binomial distribution, with the probability of success being $P(\text{Match}|p_i, p_j)$ (see Eq. 2). The following likelihood was optimized for each dataset, keeping the best of 300 replicates of the SciPy optimize L-BGFS-B implementation with random starting values (Virtanen et al., 2020).

$$f(\mathbf{K}, \mathbf{R}, \mathbf{P}) = \prod_{i,j \in \mathbf{I}} \binom{k_{i,j} + r_{i,j}}{k_{i,j}} (1 - P(\text{Match}|p_i, p_j))^{r_{i,j}} (P(\text{Match}|p_i, p_j))^{k_{i,j}}$$

where $\mathbf{I}$ is the set of observed PHRED scores, $\mathbf{K} = \{k_{i,j}: i, j \in \mathbf{I}\}$ is the number of matches indexed by paired PHRED scores, and $\mathbf{R} = \{r_{i,j}: i, j \in \mathbf{I}\}$ is the number of mismatches indexed by paired PHRED scores.

The behavior of this estimator was investigated by attempting to recover the known ground truth of the simulated data.

Algorithm 2: MLE_Calibrate

Result: Fit the vector of PHRED probability scores

```

f[] = forwardReadPairs;
r[] = reverseReadPairs;
defaultPhred[] = PhredDefinitionValues();
counts = MultiSet()
for i=0 to Length(f) do
    bestOffset, bestOverlap = AlignRead(f[i], r[i]);
    if bestOverlap ≥ 0.85 then
        CountOverlap(counts, f[i], r[i], bestOffset);
    end
end
DropLowCountCells(counts);
brierScoreInitial = BrierScore(counts, defaultPhred);
bestRes = 0;
BestTheta = [];
for i=0 to 300 do
    res, theta = ArgMin(likelihood, counts);
    if res > bestRes then
        bestRes = res;
        bestTheta = theta;
    end
end
recalibratedBrier = BrierScore(counts, bestTheta);
if recalibratedBrier < brierScoreInitial then
    return bestTheta;
end
return defaultPhred;

```

The average of the estimates and their 95% confidence intervals were plotted for all datasets.

2.6. Generation of simulated data

To assess the performance of the methods developed in this article, we created a mock data tool. The tool uses real datasets as input and creates mock datasets, taking every $k_{i,j}$ and $r_{i,j}$ and drawing from a binomial distribution $k'_{i,j} \sim \text{Bin}(k_{i,j} + r_{i,j}, p_{i,j})$ and $r'_{i,j} = (k_{i,j} + r_{i,j}) - k'_{i,j}$. Where $p_{i,j}$ is the probability of the paired base calls with PHRED scores i and j matching.

2.7. Plotting and analysis

All plots and analyses were generated using Python 2.7, Matplotlib (Hunter, 2007), SciPy 0.19.0, NumPy 1.13.3, Pandas 0.20.2 (McKinney et al., 2010), and Seaborn 0.7.1 (Waskom et al., 2017).

3. RESULTS

3.1. Statistics of MiSeq instrument sequencing

We examined 91 MiSeq instrument runs produced by 17 different MiSeq instruments that contain $N = 9.79 \times 10^9$ paired base calls. Each run consisted of between 8 and 32 multiplexed HCV samples that belonged to different genotypes of HCV. The distribution of PHRED values showed a common distribution expected from the MiSeq instrument (Fig. 1). Base calls in the forward read received higher PHRED scores, which decreased as the length of the sequence increased. The reverse read followed a similar pattern but showed much more variability in the scores assigned. In general, 8.6×10^8 reads passed the filtering steps. Most of the reads discarded were rejected due to the stringent exact match requirement of the Molecule Identifier filter (Table 1).

The total proportions of matching base calls varied significantly by run, indicating different average qualities of each run. In general, $\mu = 0.970$ of events were matches. The proportions of matches correlate strongly with the forward PHRED scores. The reverse score did not show as much influence on the proportion of matches. Similar results were obtained by Schirmer et al. (2015).

3.2. PHRED score calibration

PHRED scores were found not to be calibrated ($p < 0.0001$), but the absolute difference between the expected proportion of matches ($E[\mu]/N = 0.981$) and the observed proportion of matches ($\mu/N = 0.970$) was small. Given the large number of observed paired base calls ($N = 9.79 \times 10^9$), even very small deviations from perfect calibration could be detected. Each PHRED pair in isolation showed little absolute deviation from calibration (Fig. 2). The largest deviations were associated with rare PHRED pairs. Figure 3 shows that each individual run had a statistically significant deviation from calibration. The smallest magnitude Z-score calculated for a run was $Z = 10.6$.

The Brier score decomposition showed the amount of improvement in the predictive accuracy of matching by using the PHRED scores. In all but two cases examined, the uncorrected quality scores are useful with $REL \leq RES$. We found that the datasets with the highest UNC gain the most benefit in the Brier score from using the MiSeq instrument

PHRED scores. The dataset with the largest UNC had an improvement of -9.38×10^{-3} compared to -1.64×10^{-4} for the dataset with the least UNC. However, the reduction in uncertainty was small in most cases, with the improvement averaging -2.91×10^{-3} .

3.3. Estimation of PHRED score true values

The described ML estimator was able to recover the ground truth of the simulated data with reasonable accuracy (Fig. 4). The 95% confidence intervals of each PHRED value were found to contain the true known value. Regression of real data showed several points of miscalibration in the MiSeq instrument's PHRED estimation (see Fig. 5). Brier scores improved by an average of 2.83% when using newly calibrated values.

4. DISCUSSION

We find that the PHRED scores produced are not calibrated, but the absolute size of the error is small. Our MLE calibration method successfully corrected the error as much as possible using the features examined. To improve calibration further, additional features such as per-nucleotide biases or raw intensities would need to be utilized. However, the MLE model does not always improve the calibration, but this can be detected by examining the change in the Brier score achieved.

Although quality scores are accurate to their definition, their resolution (RES) scores are small relative to the total Brier score. The quality scores assigned are relatively accurate in predicting the error rate of the base calls. The base call procedure cannot separate correct calls from substitution errors. It tends to assign substitution errors of similar quality to the surrounding correct base calls. Given the low substitution rate of the MiSeq instrument, this is a relatively calibrated methodology, but it means that the PHRED scores are uninformative of sequencing errors.

The MiSeq instruments use two different light colors with four filters to perform base calling. Base calls generated by the same laser channel are significantly more likely to be substituted for each other than base calls generated by different channels. We note that in $98.90\% \left(\frac{90}{91}\right)$ of the datasets have on-channel transitions greater than would be expected if all transitions were equally likely (see Fig. 6). This also mirrors previous results (Schirmer et al., 2015).

Most of the findings reported in this article were foreshadowed by Schirmer et al. (2015), Eren et al. (2013), and Zhang et al., (2017). Eren et al. (2013) addressed quality scores' effects on downstream analysis and not the quality scores themselves. They found that even reads with consistently high-quality scores contained many mismatched base calls in their overlap. Schirmer et al. (2015) and Zhang et al. (2017) analyzed the quality scores directly. No direct comparison was made between our method and the method of Zhang et al. because it uses raw instrument intensities, which are not provided to the GHOST system. Schirmer et al. use mock communities of known sequences to measure errors, as our dataset does not contain known sequences, we also did not do a direct comparison to this method. These methods could not separate errors in library preparation from sequencing errors. The greatest strength of our analysis is its ability to separate these two sources of error, leading to subtly

different results. We found greater agreement between the measured PHRED scores and their true values than those found by Schirmer et al. (2015). Contrary to their results, we found that the high-quality scores seemed to underestimate the true quality of the base calls. We attribute the errors they measured to library preparation. We agree with both Schirmer et al. (2015) and Eren et al. (2013) that PHRED scores are of limited value in identifying substitution errors.

The MLE model is capable of removing the majority of calibration errors from quality scores. In two runs, the prediction error was increased by the MLE estimator. We believe that this is due to the optimization function getting stuck in a local maximum of the likelihood function.

Recalculating the Brier scores with the newly calibrated values produced an average relative improvement of 2.83%. The maximum improvement that a run-specific calibration curve can achieve is to reduce the reliability measure to zero. This was achieved in most cases. However, prediction accuracy can only be improved by considering features beyond PHRED scores.

As part of the current quality control process, the CASPER method is used to merge the overlapping paired reads (Kwon et al., 2014). By collecting statistics during this step, the calibration of PHRED scores can be included with only minimal extra processing time of less than a second per sample. A new haplotyping method using calibrated PHRED scores is being developed to be published at a later date. The accuracy of PHRED score calibration is important to improve the calling of rare variants, to which GHOST clustering is very sensitive. Currently, GHOST employs a simple method of excluding low-reliability haplotype calls; haplotypes observed fewer than 10 times are excluded. Using the calibrated PHRED scores, we can reach more deeply into the low-frequency tail of haplotypes to include rarer haplotypes in our analysis.

We also intend to further develop the methods presented in this article to provide additional quality monitoring of the sequencing of GHOST submissions (Sims et al., 2018). The new quality monitoring will include new quality control thresholds and suggestions for the laboratory process. GHOST currently requires a variant to be seen 10 times or more in a dataset to be used in clustering.

5. CONCLUSION

PHRED scores are generally accurate on the basis of their ability to predict matching. Although the Spiegelhalter z -statistic showed that the scores were not calibrated ($p < 0.0001$), the absolute deviations from the expected values were small. However, the calibration of the scores varied significantly among runs. Generally, the raw PHRED scores accurately predicted the probability of a match, and our methods improved their calibration. Even when recalibrated, the PHRED scores were not very informative of sequencing errors, as the MiSeq instrument does not separate correct base calls from substitution errors and generally assigns the background error frequency of the cycle. Given the very low error rate of the MiSeq, this tends to correctly assign high-quality scores to most calls.

ACKNOWLEDGMENTS

The authors are deeply indebted to all members of the GHOST team who contributed to this project and to members of the GHOST Project laboratories for providing data. The authors thank Eldin Talundzic (Center for Global Health Office of the Director, CDC) for editing suggestions. The authors are also very grateful to the NCHHSTP Informatics Office for their constant help with the GHOST portal website. The authors are also very grateful to the CDC's ITSO DSO office for their indispensable help with the information technology infrastructure of the GHOST analysis platform.

FUNDING INFORMATION

The work was partially supported by the Advanced Molecular Detection program (Office of Infectious Diseases, Centers for Disease Control and Prevention). The GHOST project is also the recipient of the "2015 CDC Surveillance Strategy Innovation Project Award" from the CDC Health Information Innovation Consortium (CHIC, Office of Public Health Scientific Services). NCHHSTP Informatics Office provided funding and technical resources in support of the GHOST Web User Interface and the middle-tier processes that interface with Amazon's AWS and the GHOST computational platform.

AVAILABILITY OF DATA AND MATERIALS

The data used in our analyses are deposited in and can be downloaded from the NCBI SRA under BioProject accession PRJNA580030.

REFERENCES

- Brier GW. Verification of forecasts expressed in terms of probability. *Mon Wea Rev* 1950;78(1):1–3.
- Buckley GJ and Strom BL., editors. A National Strategy for the Elimination of Hepatitis B and C: Phase Two Report. The National Academies Press, Washington, DC, 2017; doi: 10.17226/24731
- Dawid AP. The well-calibrated bayesian. *J Am Stat Assoc* 1982;77(379):605–610; doi: 10.1080/01621459.1982.10477856
- Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* 2015;31(21):3476–3482; doi: 10.1093/bioinformatics/btv401 [PubMed: 26139637]
- Eren AM, Vineis JH, Morrison HG, et al. A filtering method to generate high quality short reads using Illumina paired-end technology. *PLoS One* 2013;8(6):e66643. [PubMed: 23799126]
- Ewing B, Green P. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res* 1998;8(3):186–194. [PubMed: 9521922]
- Ewing B, Hillier L, Wendl MC, et al. Base-calling of automated sequencer traces using Phred. I. accuracy assessment. *Genome Res* 1998;8(3):175–185. [PubMed: 9521921]
- Hunter JD. Matplotlib: A 2d graphics environment. *Comput Sci Eng* 2007;9(3):90–95; doi: 10.1109/MCSE.2007.55
- Kwon S, Lee B, Yoon S. Casper: Context-aware scheme for paired-end reads from high-throughput amplicon sequencing. *BMC Bioinformatics* 2014;15(Suppl 9):S10; doi: 10.1186/1471-2105-15-S9-S10
- Longmire AG, Sims S, Rytsareva I, et al. Ghost: Global hepatitis outbreak and surveillance technology. *BMC Genomics* 2017;18(Suppl 10):916; doi: 10.1186/s12864-017-4268-3 [PubMed: 29244005]
- McKinney W Data structures for statistical computing in Python. In van der Walt S and Millman J., editors, *Proceedings of the 9th Python in Science Conference*, 51–56, 2010.
- Murphy AH. A new vector partition of the probability score. *J Appl Meteor* 1973;12(4):595–600.
- Powdrill MH, Tchesnokov EP, Kozak RA, et al. Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance. *Proc Natl Acad Sci USA* 2011;108(51):20509–20513. [PubMed: 22135458]
- Schirmer M, Ijaz UZ, D'Amore R, et al. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* 2015;43(6):e37. [PubMed: 25586220]

- Schmid CH, Griffith JL. Multivariate Classification Rules: Calibration and Discrimination. John Wiley & Sons, Ltd, 2005; doi: 10.1002/0470011815.b2a13049
- Sims S, Longmire AG, Campo DS, et al. Automated quality control for a molecular surveillance system. BMC Bioinformatics 2018;19(Suppl 11):358; doi: 10.1186/s12859-018-2329-5 [PubMed: 30343674]
- Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5(5):421–433. [PubMed: 3786996]
- Virtanen P, Gommers R, Oliphant TE, et al. ; SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(3):261–272; doi: 10.1038/s41592-019-0686-2 [PubMed: 32015543]
- Wang YH. On the number of successes in independent trials. Stat Sin 1993;3:295–312.
- Waskom M, Botvinnik O, O’Kane D, et al. mwaskom/seaborn: V 0.8.1 (September 2017); 2017; doi: 10.5281/zenodo.883859
- Zhang S, Wang B, Wan L, et al. Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling. BMC Bioinformatics 2017;18(1):335–314. [PubMed: 28697757]

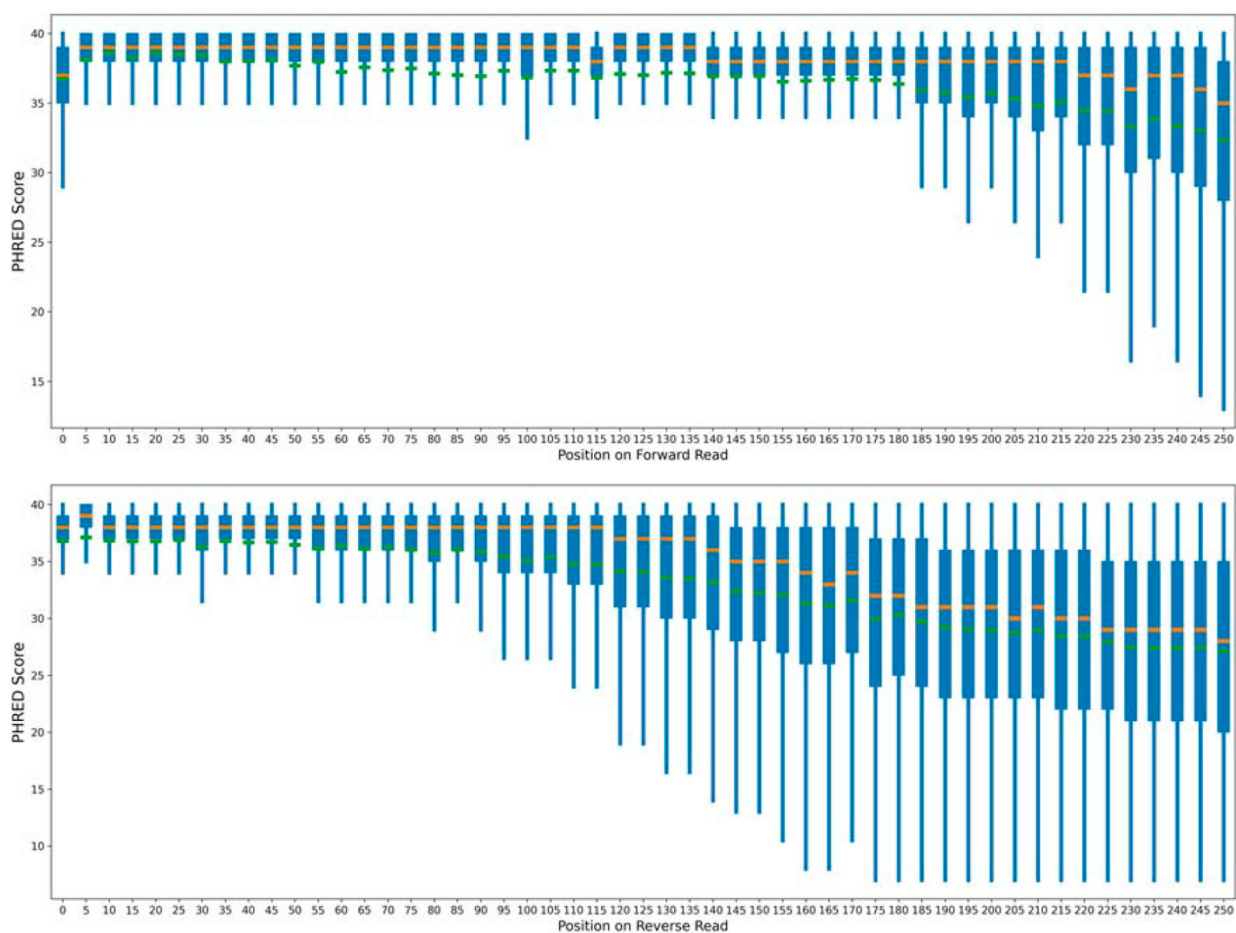


FIG. 1.
Quality score distribution by position. Orange bars are the median PHRED score, and green bars are the mean score.

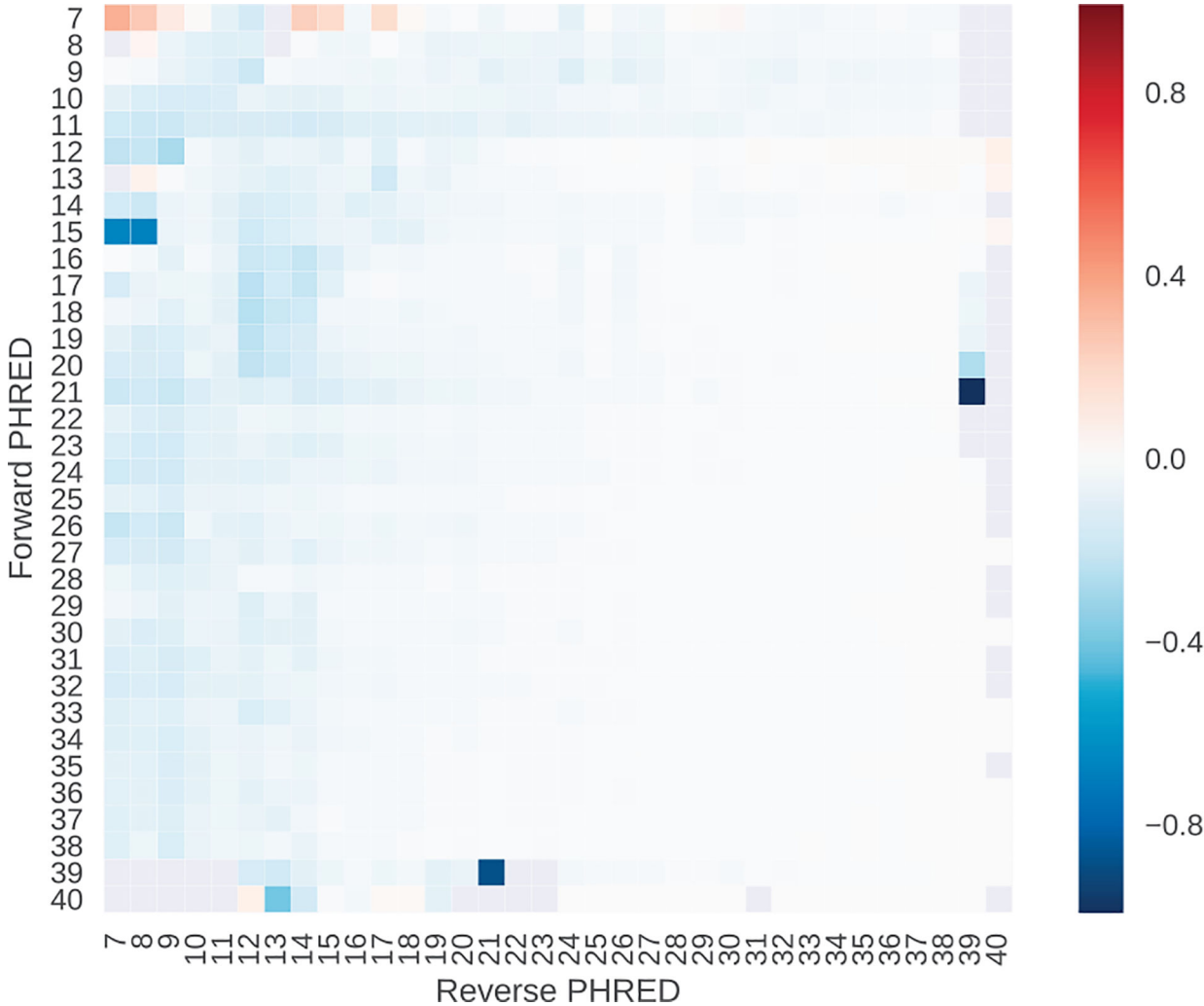


FIG. 2.
Difference between measured probability of matching ($\bar{o}_k$) and $P(\text{Match}|p_i, p_j)$ of definition of PHRED values.

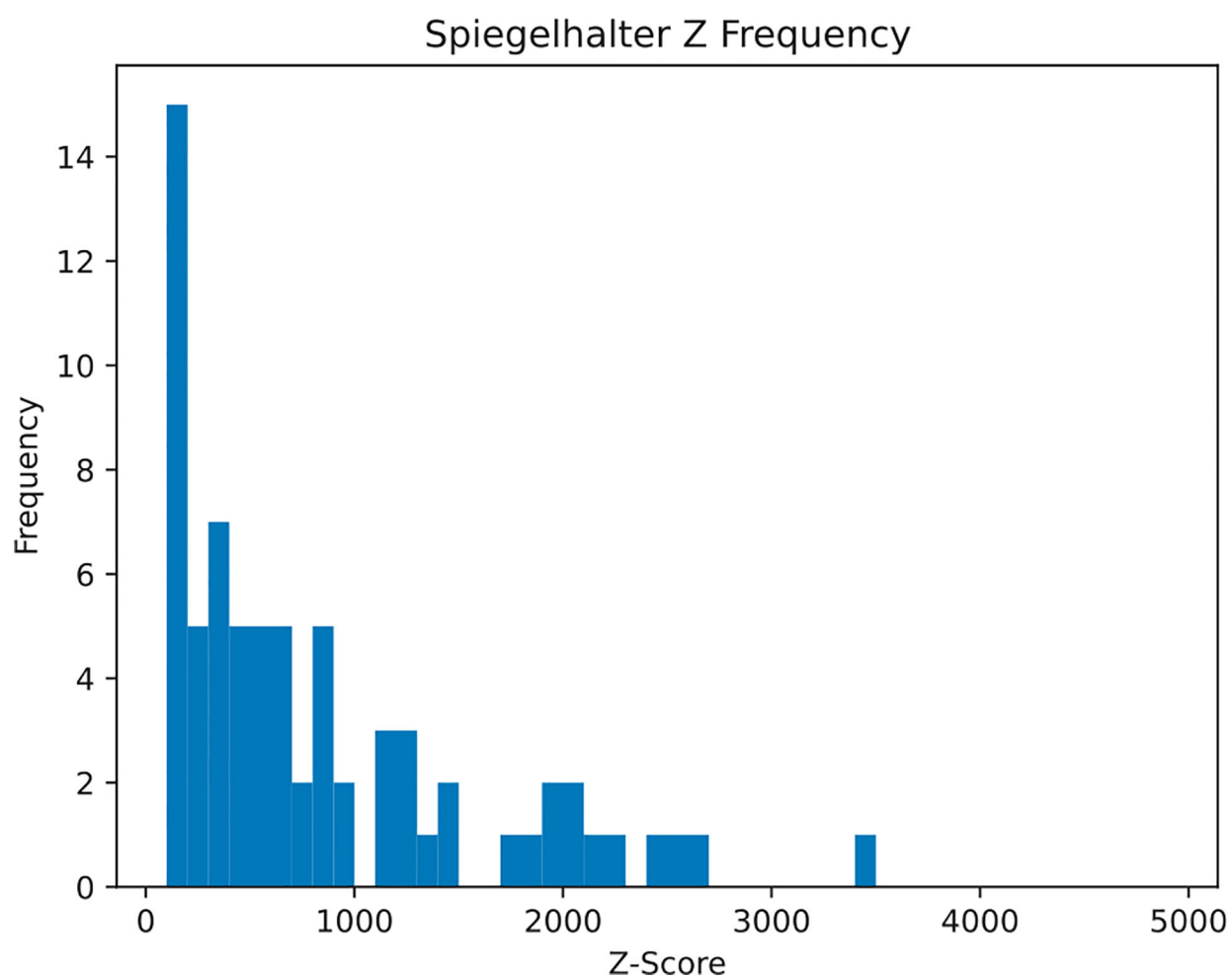


FIG. 3.
Histogram of Spiegelhalter's z-score across all datasets.

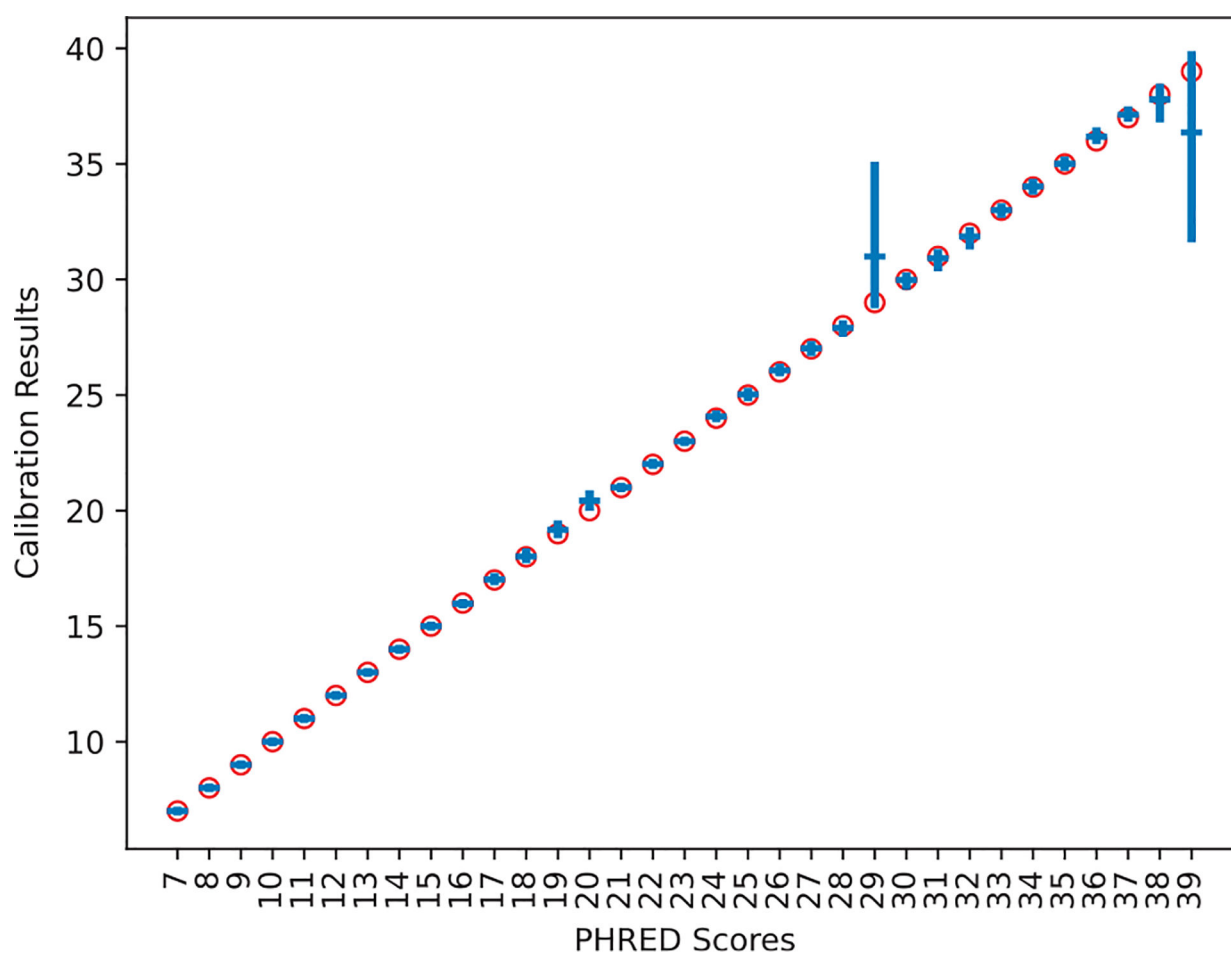


FIG. 4. Average regressed values of simulated perfectly calibrated data with 95% confidence intervals. The red circles are the correct values defined for the PHRED score.

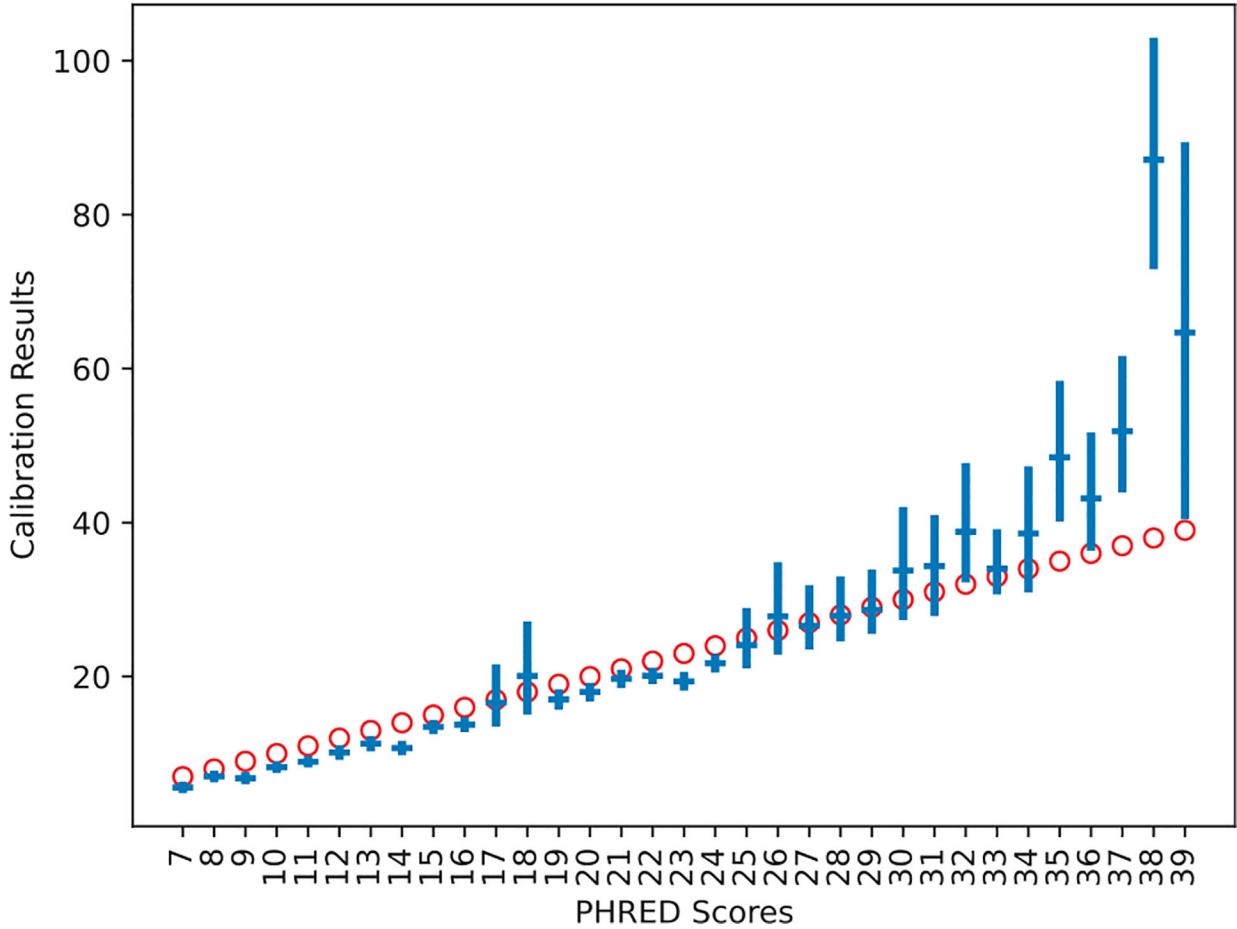


FIG. 5. Average regressed values of data with 95% confidence intervals. The red circles are the correct values defined for each PHRED score.

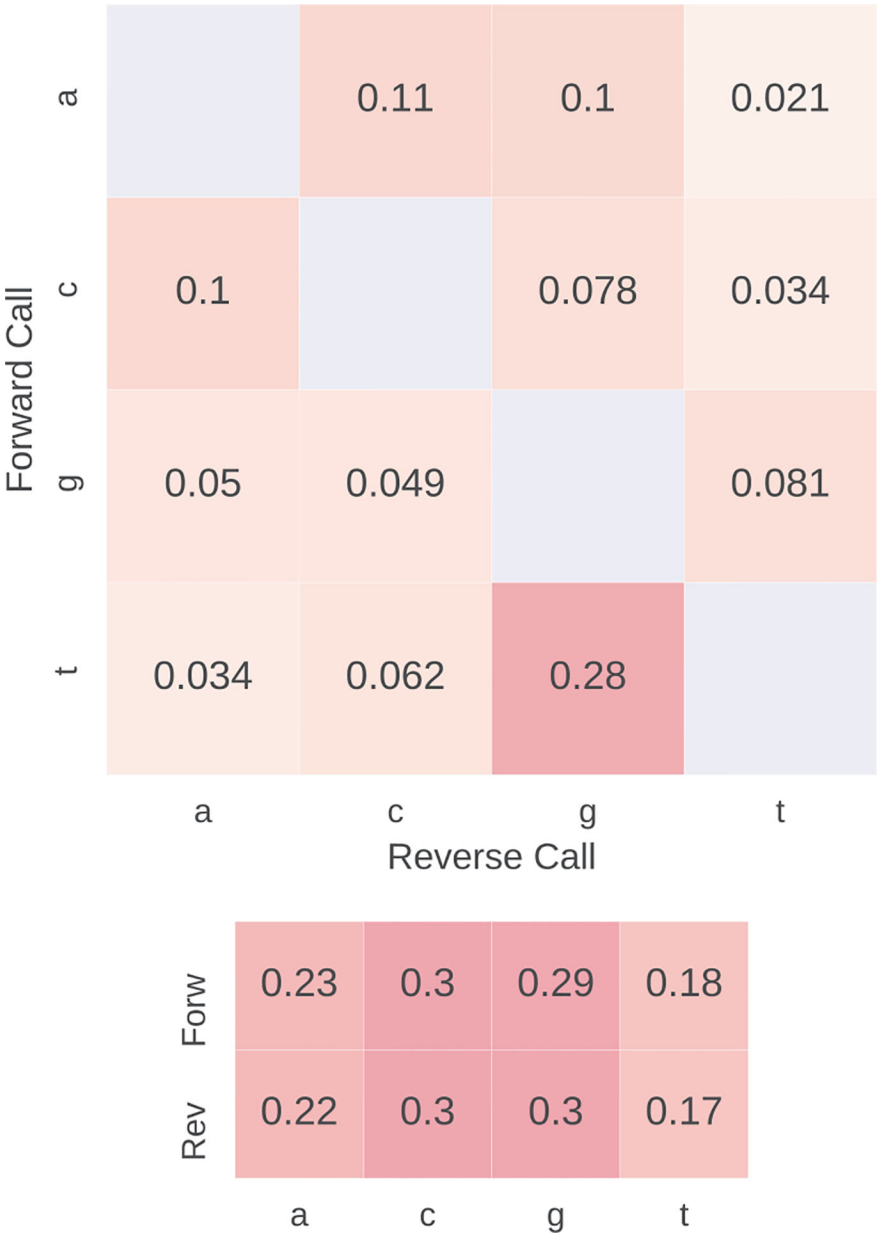


FIG. 6. The large matrix is the proportion of each mismatched paired base calls. The smaller matrix is the nucleotide composition of each read. The high GC content of the nucleotide composition is similar to that reported by Powdrill et al. (2011) for hepatitis C virus.

Table 1.

Filter Statistics

	Length dropped	N dropped	No primers	MID error	No overlap	Passed
Mean	931,415.5304	453,827.3652	1,563,563.217	2,432,329.783	638,479.287	7,474,166.548
Standard deviation	1,823,711.84	2,102,838.347	3,114,852.383	3,154,940.167	804208.6185	6,363,754.282
Minimum	0	0	2879	9672	1438	10,904
First quartile	7410.5	1	67,595.5	187,545	74,285.5	711,593
Median	166,371	12	578,217	1,774,128	392,095	7,699,798
Third quartile	1,236,604.5	12,612	1,427,689.5	2,943,832.5	825,187	13,138,532.5
Maximum	10,746,614	15,525,011	17,478,841	17,479,145	5,354,004	21,010,761
IQR	1,229,194	12,611	1,360,094	2,756,287.5	75,0901.5	12,426,939.5

IQR, interquartile range; MID, Molecule Identifier.