Designing Meaningful Measures of Read Length for Data Produced by DNA Sequencers

David O. Nelson and Jane Fridlyand

Abstract

Nearly everyone uses "the number of Q20 bases" as a rough measure of the effective length of a given DNA sequence produced by the base-caller PHRED. This metric simply counts the number of bases in a read in which the PHRED quality score is at least 20. While the number of Q20 bases is a simple, easy to implement rule-of-thumb, it does not have much else going for it: it consistently underestimates the number of usable bases in the read. In this short paper, we develop and evaluate an alternative metric that uses more of the PHRED quality data in a read to predict how many bases from that read would make it into the eventual consensus sequence of an assembly. The metric was developed by evaluating a set of pre-existing, high-quality assembled contigs. The resulting predictor is a simple function of the histogram of PHRED quality values already produced by sequencing software and performs nearly as well as a more complex additive model that uses regression splines.

Keywords: DNA read length; genomics; predicting progress; PHRED; PHRAP

1 Introduction

Large-scale genome sequencing projects have become increasingly common over the last fifteen years. Many recent papers, starting with Lander and Waterman in 1988 [6], have described mathematical models for predicting the progress of such sequencing projects. These different "Lander-Waterman" analyses arise in response to different approaches to sequencing large genomes. They model the sequencing process as a coverage process like those described by Hall [5] and derive predictions of mean coverage, depth, expected number of gaps, and the like, as a function of the number of clones sequenced N, the genome size G, and the length of sequence L obtained from an individual clone chosen for sequencing. These predictions are then used to estimate the number of clones required to obtain an assembled genome to a given depth or coverage. Conversely, statistics on coverage and read length gathered during the sequencing effort are used with these models to track progress, detect problems, and refine estimates of the remaining work required.

The approximate genome size G can be determined in advance, and number of clones sequenced N is easy to obtain from daily production statistics. However, what

about L? The average number of bases sequenced from the end of a clone can be tuned by changing electrophoresis conditions, run-time, and the sequencer chosen to sequence the DNA. In fact, deciding on the desired length L is a major factor in planning sequencing projects, along with the size (or sizes) of sequencing clone and whether or not both ends are to be sequenced.

On one hand, most Lander-Waterman analyses assume L to be a constant (although Lander and Waterman do provide some guidance on the degradation in performance due to random clone sizes) and assume that any overlap between two clones is detected with probability one for overlaps of a certain size or greater. On the other hand, sequencing centers that use the most popular combination of base-caller and assembler, PHRED [4, 3] and PHRAP [7], are faced with a much more complex situation with respect to read length and overlap detection. PHRED can produce *extremely* long reads, but also throttles the process somewhat by providing a probability of error with each base read. This base-specific probability of error is expressed as a "quality value" for each base i: $q_i = -10\log_{10} p_i$, where p_i is (more or less) the probability that base i is called in error. PHRED produces integer quality values ranging from zero to approximately fifty, and those associated with bases at the ends of the read are typically much lower than those in the middle. PHRAP uses these quality values in the assembly process in a complex way. A byproduct of an assembly is a "trimmed" read for each read that entered the assembly, in which some number of bases at the start and end of each read are discarded during the alignment process.

Most sequencing centers finesse the problem of estimating L for a read by the simple expedient of counting the number of bases in a read for which $q_i \geqslant 20$. This Q_{20} rule arose during the initial phases of the Human Genome Project and was adopted by the public consortium as a common measure of read length. However, for planning future projects, it would be desirable to derive a better measure of read length, and preferably one that related to some measure of the useful size of a read.

In this paper, we define the "effective read length" of a read in an assembly as the length of the trimmed read produced by PHRAP. We believe that this definition of effective read length provides a more reasonable model for L in Lander-Waterman analyses of projects that use PHRED and PHRAP as a base-caller and assembler. In this paper, we explore some of the features of this distribution and build predictors of L as a function of the set of q_i. Our goal is to provide a simple algorithm to estimate L that is more accurate and precise than the Q_{20} rule currently in place.

2 Methods

All analyses were done using the statistical computing environment *R* [8].

Source of Reads

We analyzed assemblies from fifty-one sequencing projects produced by the Joint Genome Institute (JGI) in Walnut Creek, California during two time periods spanning the 1998–2002. Forty-eight of the sequencing projects were cosmids completed during the period of November 1998–April 1999. These projects were sequenced on ABI 377 slab gel sequencers [2]. Three of the sequencing projects were bacterial artificial chromosomes (BAC's) completed during the period June 2002–August 2002. These projects were sequenced on a combination of Molecular Dynamics Megabace 1000 [1] and ABI 3700 class capillary sequencers.

All projects were base-called and assembled using the current versions of PHRED and PHRAP with default parameters. From each project, we selected only those contigs which contained at least 300 reads and had coverage between 5 and 60. Five of the projects had two such contigs; the rest had just one "main" contig.

Data Gathering

We excluded reads that contained vector, as they would need a special treatment in order to remove the vector sequence and calculate effective read length. In addition, we excluded those reads that had ends extending outside the trimmed part of the final contig. We obtained statistics on each read from the output files produced by PHRED, as well as the standard output file produced by PHRAP. Data obtained for each read included the length of the untrimmed read; the length of the trimmed read; the insertion, deletion, and substitution error rates in the trimmed part of the read; the expected number of correct bases in the read, defined as $n - \sum_i 10^{-q_i/10}$, where n is the number of bases read and the q_i are the corresponding quality values; the number of bases in the read with PHRED quality values in five different histogram bins (0–9, 10–19, 20–29, 30–39, 40 and above); and the expected number of correct bases in each of those histogram bins.

3 Results

Distribution of Percent Trimmed

Table 1 summarizes, by quintile of average depth, the characteristics of the 52,097 reads from fifty-one contigs obtained from the forty-eight slab gel projects. Each row of the table shows summary statistics for one of five quintiles of depth of coverage in the slab data set. Summary statistics for each quintile include the number of contigs, the median number of reads in the contigs, and the median length of the contigs. Cosmids are around 40,000 bases long, approximately the same size as the median size of each contig in all five depth quintile.

Table 2 describes the characteristics of the 13,539 reads from five BAC contigs obtained from the three capillary electrophoresis projects. BAC's are considerably longer

Table 1: Characteristics of 51 cosmid contigs by depth quintile

Quintile	Depth	Number of Contigs	Median Number of Reads	Median Length of Contig
1	[12,14]	14	772	43,458
2	(14,16]	9	882	39,670
3	(16,18]	10	945	41,473
4	(18,21]	8	1,064	38,790
5	(21,46]	10	1,326	39,890

Table 2: Characteristics of five BAC contigs sequenced by capillary electrophoresis

Project	Contig	Depth	Number of Reads	Contig Length
THW	I	37	3,934	177,944
TKM	I	50	1,967	67,818
	II	50	2,729	96,546
TKP	I	39	1,012	40,502
	I	45	3,897	142,666

than cosmids, ranging from 150,000 to 200,000 bases in length. Adding up the contig sizes, we see that the selected contigs represent approximately the length of their respective clones, and are all sequenced to high depth.

Figure 1 shows the relationship between raw and trimmed read length as a function of sequencing technology and quartile of raw read length within sequencing technology. Note that the read length quartile values differ in slab and capillary technologies, largely because of the superior read length obtained with current capillary machines. The quartile bins of raw read length for slab reads were 107–607, 608–657, 658–832, and 833–2149. For capillary reads, the quartile bins were 187–795, 796–1114, 1115-1213, and 1214–1578. There is considerable similarity between the distributions of proportion trimmed, as a function of read length. The longer the raw read is, the larger the proportion trimmed. Hence, the number of bases trimmed goes up dramatically as the raw read length increases. The larger spread of proportion trimmed in the highest quartile of slab gel reads is likely due to the extremely large size of the bin (833–2149, versus 1214–1578 for the capillary reads). Overall, the median percentage trimmed was slightly over seven percent, approximately twenty-five percent of the reads had less than five percent trimmed, seventy-five percent of the reads had less than fifteen percent of the raw read length trimmed, and 456 reads had over 90 percent trimmed.

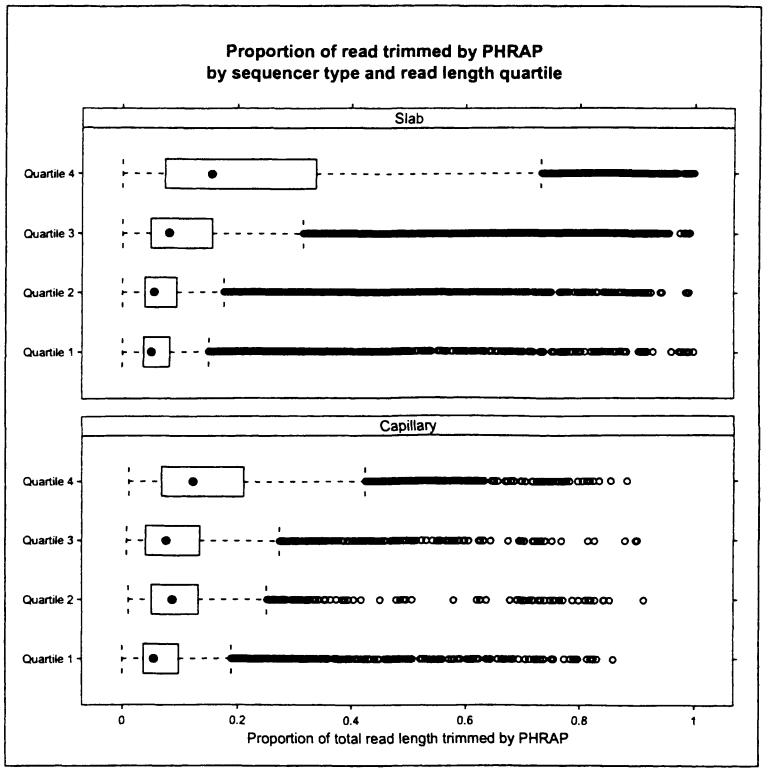

Figure 1: Boxplots showing the distribution of of the proportion of the read trimmed by PHRAP, as a function of the type of sequencer (slab gel vs. capillary) used and the quartile of raw read length. Within each panel, four boxplots are shown: one for each quartile of raw read length. The top panel shows the distribution for the reads in the data set that were produced by a slab gel sequencer, while the bottom panel shows the distribution for reads in the data set that were produced by capillary sequencer. Note that the average proportion trimmed increases with increased read length, but is relatively stable across sequencing technologies.

Current Measures of Effective Read Length

We now examine how well two common measures of read length predict the actual number of bases used by PHRAP: the Q_{20} rule and the "expected correct", or E_c rule. First, we examine the Q_{20} rule.

Recall that the Q_{20} rule simply counts the number of bases with a PHRED quality score of 20 or more. Figure 2 shows a scatter plot of the relationship between Q_{20} and the number of bases actually used. This plot shows clearly the extent to which Q_{20} dramatically underestimates the actual number of bases used by PHRAP. Superimposed on the scatter plot is a scatter plot smoother fit and a line dividing the region where Q_{20} overestimates from the region where it underestimates. Note the almost total lack of

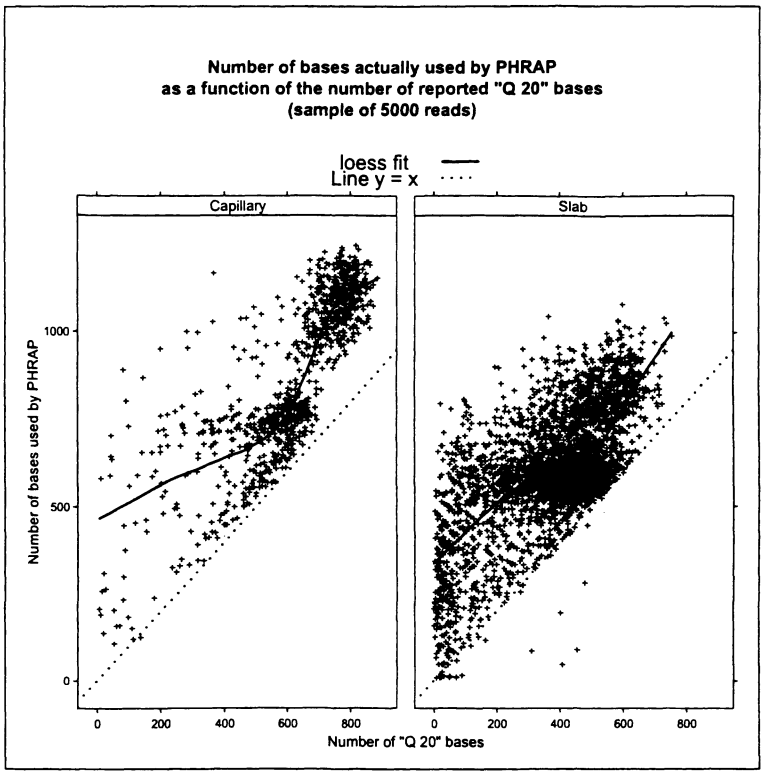

Figure 2: The relationship between the number of "Q 20" bases in a read and the actual number of bases used in the assembly. The Q_{20} rule almost universally underestimates the number of usable bases in a read, irrespective of sequencing technology. Note that for ease in graphing, only 5000 randomly sampled points are plotted.

points where Q_{20} underestimates the read length. This result is not too surprising, as PHRAP goes to great lengths to use the bases at the ends of the read, where the quality scores are typically low. In addition, the graph does indicate that some other metric that uses more of the information in the PHRED histogram cannot help but improve the performance of a read length predictor.

A second obvious choice for read length estimator is the expected number of correct bases, which can be written as

$$E_c = n - \sum_{i=1}^{n} 10^{-q_i/10},$$

where n is the number of bases in the untrimmed read. This estimator subtracts off a read-specific constant from the untrimmed read length in an attempt to estimate the number of trimmed bases.

Figure 3 shows a scatter plot much like Figure 2, only plotting the number of bases

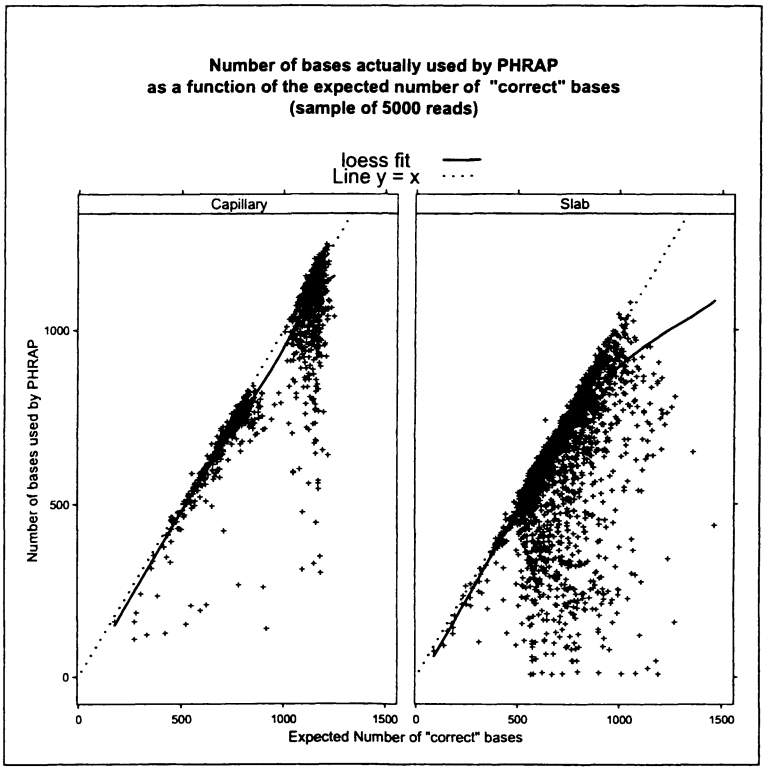

Figure 3: The relationship between the expected number of correct bases, as estimated by PHRED quality scores, and the actual number of bases used in the assembly. Note that the "expected number" rule performs better, but often overestimates the number of usable bases in a read.

used against E_c. Here we see the opposite effect of that observed with the Q_{20} rule: E_c tends to *overestimate* the number of bases used by PHRAP. Despite that overestimation, the tight clustering of points around the line $y = x$ indicates that this estimator is clearly superior to the Q_{20} rule, especially for the capillary reads. Perhaps some combination of the two estimators should be considered. We now examine that possibility.

Additive Combinations of Histogram Values

We now consider a simple generalization of the above two estimators. The Q_{20} rule can be written as a simple affine combination of the histogram counts produced as a byproduct of the sequencing process flow at the JGI:

$$Q_{20} = w_0 + w_1 N_1 + w_{10} N_{10} + w_{20} N_{20} + w_{30} N_{30} + w_{40} N_{40},$$

where $N_1, N_{10}, N_{20}, N_{30}$, and N_{40} are the number of bases in the read with PHRED quality values in $[0,9)$, $[10,19)$, $[20,29)$, $[30,39)$, and $[40,50]$, respectively, and w_0, w_1, w_{10},

w_{20}, w_{30}, and w_{40} are weights. For the Q_{20} rule, $w_0 = w_1 = w_{10} = 0$, and $w_{20} = w_{30} = w_{40} = 1$. The E_c estimator can also be easily approximated by linear combination of these histogram counts, with the weights w_j approximating error probabilities for bases in histogram bin j. We will now generalize to additive combinations of smooth functions of histogram counts as predictors of read length.

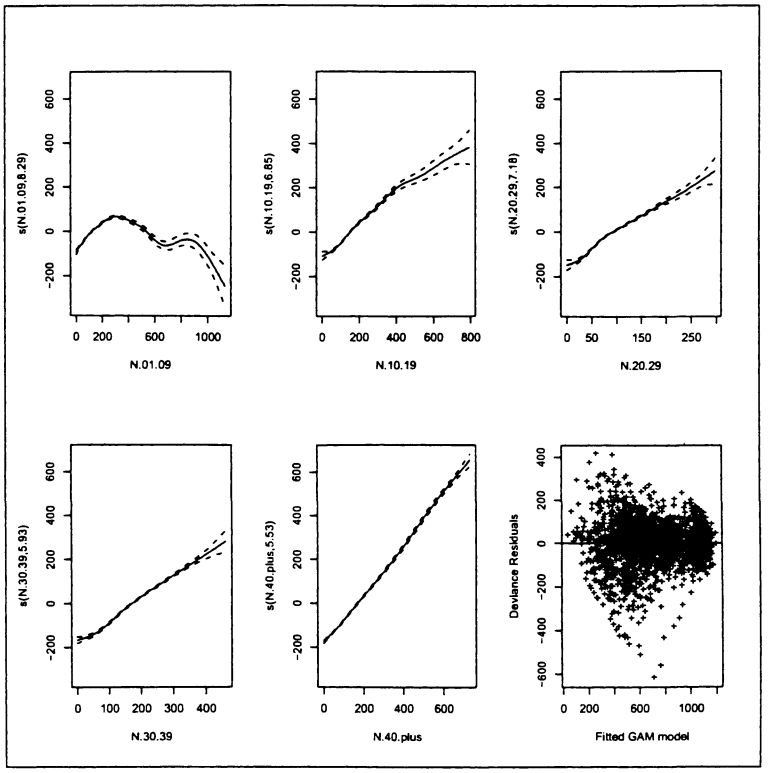

Figure 4: The five smooth terms in a GAM estimate of the number of bases used by PHRAP, along with a plot of the deviance residuals versus fit. The five terms correspond to smooth functions of the number of bases in each of the five histogram bins described in **Methods**. In addition, a two-factor term adjusting for sequencing technology was also included. The labels "N.01.09" through "N.40.plus" correspond to the five histogram bins of PHRED values produced by local sequencing software (0–9, 10–19, 20–29, 30–39, and 40 or more). Each y axis label is of the form $s(x,d)$, where $s()$ is an estimated smoothing spline, x is the count in the corresponding histogram bin, and d is the approximate degrees of freedom in the estimated spline. Only the 5000 random points plotted in Figures 2 and 3 were used in the fit.

Figure 4 shows the result of fitting a generalized additive model to the read length data described above. The model fit was of the form

$$y_i = \alpha_0 + \alpha_1 x_i + \sum_j f_j(n_{ij}) + \varepsilon_i$$

where y_i is the number of bases used by PHRAP for read i, x_i is an indicator variable that is one whenever the read was performed on a slab instrument, n_{ij} is the number of bases in histogram bin j for read i, the f_j are penalized regression splines [9] with knots to be determined by cross-validation, and ε_i is Gaussian error. Five of the panels in Figure 4 show estimates of the f_j, along with standard errors, while the lower right panel shows a plot of deviance residuals versus fitted values. All of the terms in the model were highly significant, and the adjusted R^2 of the fit was 0.88.

From Figure 4 we see that, except for the histogram bin corresponding to bases with $q < 10$, the resulting splines are reasonably linear. The spline for the $q < 10$ bin, on the other hand, looks more complex. The conclusion we draw is that more detail about the structure of the $q < 10$ quality values will be needed to make a substantial improvement. However, as a first approximation, the spline looks like it might be adequately approximated by a quadratic.

Consequently, we refit the data to a linear model with a quadratic term for the $q < 10$ histogram term:

$$y_i = \alpha_0 + \alpha_1 x_i + \beta_0 \frac{n_{i0}}{100} + \beta_1 \left(\frac{n_{i0}}{100}\right)^2 + \sum_{j>0} \gamma_j n_{ij} + \varepsilon_i \tag{1}$$

(In order to keep the quadratic coefficient to a reasonable size, we scaled the $q < 10$ histogram value by dividing by 100). The results of the fit are shown in Table 3. We

Table 3: Results of linear model (quadratic term for N.01.09)

Term	Estimate	S.E.	t Ratio	$\Pr(>\|t\|)$
Intercept	-101.90	7.58	-13.45	$< 10^{-15}$
Slab	35.00	4.36	8.03	$< 10^{-14}$
N.01.09/100	79.00	2.19	36.07	$< 10^{-15}$
$(N.01.09/100)^2$	-9.59	0.29	-32.98	$< 10^{-15}$
N.10.19	0.79	0.02	40.04	$< 10^{-15}$
N.20.29	1.57	0.03	51.83	$< 10^{-15}$
N.30.39	1.01	0.02	61.52	$< 10^{-15}$
N.40.plus	1.15	0.01	107.61	$< 10^{-15}$
Residual S.D.	76			
Adjusted R^2	0.87			

see that the linear coefficients for bases with $q < 20$ are around 0.79, while the bases with $q \geqslant 20$ have coefficients somewhat above one. We also see that a large number of bases with low quality decreases the effective read length.

The fit in Table 3 was based on a training of 5000 points. In order to evaluate the prediction error, we examined the distribution of the absolute value of the difference

between the predicted number of bases and actual number of bases on a test set consisting of the other 60,636 reads in the data set. Table 4 summarizes that distribution, and compares it with the prediction error for four other estimators: the generalized additive model described above, a linear model without a quadratic term for the $q < 10$ bases, E_c, and Q_{20}.

Table 4: Absolute prediction error quantiles for estimators of effective read length

	Quantile of Prediction Error				
Model	25%	50%	75%	95%	99%
Additive Model	10	24	52	159	298
Linear Model					
(with quadratic term)	12	26	56	161	302
(no quadratic term)	19	39	71	173	323
E_c	10	22	62	365	703
Q_{20}	115	201	309	462	594

We see that, except for extremely large errors, the Q_{20} estimator is dominated by each of the other estimators analyzed. We see that the E_c estimator is quite competitive with the linear model, at least until the read length gets extremely large.

4 Conclusions

We can draw several conclusions from the above analyses. First, the Q_{20} predictor grossly underestimates the effective length of a sequencing read. Except for the extreme cases, all of the other predictors discussed dominate it under all circumstances in which they were compared. Second, the E_c and the linear model predictors have comparable prediction error: on average, about sixty bases. However, the E_c estimator has two disadvantages when compared to the estimators derived from a linear model. First, the errors for E_c appear to be biased: on average, E_c overestimates the read length. Second, E_c requires the entire set of PHRED quality scores. If we restrict our attention to estimators based on histograms only, Table 4 shows that the best estimator based only on a linear combination of histogram values is dominated by the linear model with an added quadratic term, as expected. Finally, we note that the appropriate simple linear combination of PHRED histogram bins is quite competitive with the much more complex generalized additive model. The main benefit of adding a quadratic term seems to be to decrease prediction error in the extreme case of a long, low-quality read.

The boxplots in Figure 1 show a considerable amount of skewness in the distribution of percent trimmed. Although this skewness is transmitted somewhat to the number of bases in the consensus, log-transforming the outcome y_i in Equation 1 does not improve the prediction error at all.

In these analyses, we have not explored any effects due to mis-called bases. Other statistics gathered for these analyses include the percent of indels (insertions/deletions) and substitution errors in the trimmed read. Our analysis (not shown) indicates that this component of effective read length is small (under a few percent), and is dominated by PHRAP's trimming process.

It seems clear that the relationship between the PHRED quality values and the size of the region PHRAP trims from the raw read is both simple and quite complex. It is simple in the sense that, in most cases, the expected number of bases E_c closely matches what PHRAP uses. However, an examination of Figure 3 shows that as the raw read gets longer, the situation becomes quite complex, and the size of the region trimmed becomes more of a function of serial correlations between quality values. This situation is exactly what various kinds of "moving window" trimming algorithms try to capture. It would be interesting to explore the extent to which statistically-based moving window algorithms might outperform the marginal approach outlined above in the situation of long, low-quality reads.

Acknowledgments

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract number W–7405–ENG–48.

David O. Nelson, Joint Genome Institute, Lawrence Livermore National Laboratory,
daven@llnl.gov

Jane Fridlyand, Comprehensive Cancer Center, University of California, San Francisco,
janef@cc.ucsf.edu

References

[1] Amersham Biosciences. Megabace web site. http://www.megabace.com.

[2] Applied Biosystems. http://www.appliedbiosystems.com/products.

[3] Brent Ewing and Phil Green. Basecalling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research*, 8:186–194, 1998.

[4] Brent Ewing, LaDeana Hillier, Michael C. Wendt, and Phil Green. Basecalling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8:175–185, 1998.

[5] Peter Hall. *Introduction to the Theory of Coverage Processes*. John Wiley and Sons, 1988.

[6] Eric S. Lander and Michael S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2:231–239, 1988.

[7] University of Washington. Phrap web site. `http://www.phrap.org`.

[8] R Project. R web site. `http://www.r-project.org`.

[9] Simon N. Wood. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society B*, 62(4):413–428, 2000.