Published online September 30, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 17 ***5183–5191***
doi:10.1093/nar/gkh850

Adjust quality scores from alignment and improve sequencing accuracy

Ming Li, Magnus Nordborg and Lei M. Li*

Computational Biology, University of Southern California, Los Angeles, CA, USA

Received July 9, 2004; Revised and Accepted September 8, 2004

ABSTRACT

In shotgun sequencing, statistical reconstruction of a consensus from alignment requires a model of measurement error. Churchill and Waterman proposed one such model and an expectation–maximization (EM) algorithm to estimate sequencing error rates for each assembly matrix. Ewing and Green defined *Phred* quality scores for base-calling from sequencing traces by training a model on a large amount of data. However, sample preparations and sequencing machines may work under different conditions in practice and therefore quality scores need to be adjusted. Moreover, the information given by quality scores is incomplete in the sense that they do not describe error patterns. We observe that each nucleotide base has its specific error pattern that varies across the range of quality values. We develop models of measurement error for shotgun sequencing by combining the two perspectives above. We propose a logistic model taking quality scores as covariates. The model is trained by a procedure combining an EM algorithm and model selection techniques. The training results in calibration of quality values and leads to a more accurate construction of consensus. Besides *Phred* scores obtained from ABI sequencers, we apply the same technique to calibrate quality values that come along with Beckman sequencers.

INTRODUCTION

Shotgun sequencing is the standard methodology for genome sequencing (1). Starting with a whole genome, or a large genomic region, short random fragments are generated and sequenced. Enough fragments are sequenced so that almost all positions in the genome or region are covered multiple times just by chance. The standard sequencing procedure is exemplified by the commonly used *Phred*/*Phrap* suite of software (2,3). First, each fragment is base-called from its chromatogram, i.e. a vector times series of four fluorescence intensities, and a sequence of A, C, G and T, is inferred. The commercial producers of sequencing devices include Applied Biosystems, Inc., Beckman Coulter, Inc., etc. Typically each base is accompanied by a quality value that is meant to convey some idea of how likely the base-calling is correct. Second, the base-called sequences are assembled into a contig using an ad hoc alignment algorithm that compares both strands and overlap between fragments. Quality values are taken into account during the alignment to eliminate low quality reads. Third, a consensus sequence is constructed from this alignment by comparing different reads for each position.

The accuracy of the consensus sequence depends on the coverage (i.e. how many independent observations we have for each nucleotide base pair in the genome) and the performance of the base-calling algorithm. The quality values of base-calling play a crucial role in the construction of consensus. If they are misleading or interpreted incorrectly, the consensus sequence will be less reliable. The *Phred* quality scores for base-calling are defined from sequencing traces in such a way that they have a probabilistic interpretation. This is achieved by training a model on a large amount of data. However, in practice, sample preparations and sequencing machines may work under different conditions and therefore quality scores need to be adjusted. Moreover, the information given by quality scores is incomplete in the sense that they do not describe error patterns. We observe that each nucleotide base has its specific error pattern that varies across the range of quality values.

Churchill and Waterman (4) proposed another model to define a consensus. It is based on an assembly without assuming the availability of quality values. The parameters in the model include composition probabilities and sequencing error rates and are estimated by an expectation–maximization (EM) algorithm based on the alignment. The consensus is defined by the probability of the target sequence conditional on observations. This offers an evaluation of reliability.

In this article, we combine quality scores of base-calling and the idea in Churchill and Waterman's model (4), to improve sequencing accuracy. Specifically, we start with assembled contigs and quality scores to build up complete probabilistic error models. One option is to represent the error pattern of each nucleotide by a multinomial model. Since the true sequence is unknown, we develop an EM algorithm to deal with missing data. In a more sophisticated logistic model, we take quality scores as covariates. To parsimoniously represent the non-linear effect of quality scores, we adopt simple piecewise linear functions in the regression model. The model is trained by a procedure combining an EM algorithm, the Bayesian information criterion (BIC) criterion and backward deletion. The training results in calibration of quality values and leads to a more accurate consensus construction.

*To whom correspondence should be addressed. Tel: +1 213 740 2407; Fax: +1 213 740 2437; Email: lilei@usc.edu

MATERIALS AND METHODS

Sequencing data

The first source of data in this article comes from the *Campylobacter jejuni* whole-genome shotgun sequencing project (5). The raw data, generated on ABI 373 and 377 automated sequencers were downloaded from the Sanger Center (ftp.sanger.ac.uk/pub/pathogens/cj). The total length of the genome sequence is 1 641 481 bp. There are 33 824 reads and the average coverage is 10-folds. The sequence assembly was obtained by *Phrap* (see http://www.phrap.org). We tested our methods on the first 100 kb of the reference sequence and the corresponding reads. The coverage of the *C.jejuni* sequencing project is unusually high, so we randomly removed some reads to decrease the average coverage from 10- to 6-fold. Since the reference sequence was obtained on reads of 10-fold, we will assume that it is close to the true sequence later when we calculate single base discrepancy (SBD).

To test our methods on data obtained using another sequencing technology, we analyzed data from an *Arabidopsis thaliana* re-sequencing project carried out at USC (http://walnut.usc.edu/2010) using Beckman Coulter CEQ automated sequencers. These data were obtained as part of a polymorphism survey and thus contain different haplotypes. Since we are interested only in sequencing error, in this paper, we selected ~500 kb of raw data from non-polymorphic regions.

Setup

Throughout the article, we represent random variables by capital letters and their values by small letters. First, reads are aligned into an assembly matrix. We introduce two alphabets: $\mathcal{A} = \{A, C, G, T, -\}$ and $\mathcal{B} = \{A, C, G, T, -, N, \phi\}$, where $-$ denotes an internal gap, N denotes any ambiguous determination of a base and the null symbol ϕ is for non-aligned positions beyond the ends of a fragment. Each fragment is either in direct or in reverse complemented orientation. To deal with the issue of orientation, we introduce a complementary operation $\sim$ on the alphabet $\mathcal{B}$ as follows: $\tilde{A} = T$, $\tilde{T} = \mathrm{A}$, $\tilde{G} = \mathrm{C}$, $\tilde{C} = \mathrm{G}$, $\tilde{\phi} = \phi$ and $\tilde{-} = -$. An illustrative example of assembly matrices is shown in Figure 1.

We denote the target sequence by $S = S_1 S_2 \cdots S_n$, where S_j takes any value from the alphabet $\mathcal{A}$. Random fragments generated from the template are aligned by an assembler. This results in an assembly matrix $\{X_{ij}\}_{m\times n}$. The elements of the fragment assembly matrix, denoted by x_{ij}, take values from the alphabet $\mathcal{B}$. Each row in $\{X_{ij}\}$ contains the ordered sequence of bases and possible gaps in a particular fragment. The column index $j = 1, \ldots, n$ runs from the leftmost base in the assembly to the rightmost. We represent the orientation of the ith fragment in the assembly by

$$r_i = \begin{cases} 0 & \text{fragment } i \text{ is in direct orientation,} \\ 1 & \text{fragment } i \text{ is in reverse orientation.} \end{cases}$$

The observations X_{ij} are subject to measurement error. We denote the true base of fragment i at position j by $Y_{ij} \in \mathcal{A}$. Therefore

$$Y_{ij} = \begin{cases} S_j & \text{if } r_i = 0, \\ \tilde{S}_j & \text{if } r_i = 1. \end{cases}$$

We denote the compositional probability by $\alpha_a = \mathbf{Pr}(S_j = a)$, $a \in \mathcal{A}$.

Phred quality scores

After appropriate preprocessing, each sequencing chromatogram contains a series of peaks of four colors. The rationale of base-calling is that each peak represents one base, and the order of peaks from the four channels is consistent with the order of nucleotide bases on the underlying DNA fragment. In addition to base-calling, *Phred* also assigns each base-call a quality score q, which takes integer values from 0 to Q (Q is 64 for *Phred* scores) (2). Quality scores are based on trace features such as peak spacing, uncalled/called peak ratio and peak resolution. The model that defines quality scores was so trained, on a large amount of sequencing traces, that the scores could be interpreted as probabilities. Mathematically, the score is defined by

$$q_{ij} = -10\log_{10}\varepsilon_{ij}, \quad \text{where } \varepsilon_{ij} = \mathbf{Pr}\left(Y_{ij} \neq x_{ij} \,|\, X_{ij} = x_{ij}\right), \qquad \mathbf{1}$$

where, ε_{ij} is the error probability of base-calling. We randomly select one position from an assembly and let Y and X be its true base and called base, respectively. Let $\mathcal{E}$ denote the event that the base-calling is incorrect, namely, $\mathcal{E} = \{X \neq Y\}$. Then the correct calling probability given base a is: $1 - \varepsilon = \mathbf{Pr}(Y = a \,|\, X = a)$, where $a \in \mathcal{A}$. Notice that

$$\begin{aligned} \mathbf{Pr}(X = a \,|\, Y = a) &= \frac{\mathbf{Pr}(Y = a \,|\, X = a)\mathbf{Pr}(X = a)}{\mathbf{Pr}(Y = a)} \\ &= (1 - \varepsilon)\frac{\mathbf{Pr}(X = a)}{\mathbf{Pr}(Y = a)}. \end{aligned}$$

If the assumption of unbiased base-calling is valid, namely, $\mathbf{Pr}(X = a) = \mathbf{Pr}(Y = a)$, then we have $\mathbf{Pr}(X = a \,|\, Y = a) = \mathbf{Pr}(Y = a \,|\, X = a) = 1 - \varepsilon$. Consequently, we are able to interpret the *Phred* scores as probabilities by

$$\mathbf{Pr}\left(X_{ij} = x_{ij} \,|\, Y_{ij} = x_{ij}\right) = \mathbf{Pr}\left(Y_{ij} = x_{ij} \,|\, X_{ij} = x_{ij}\right) = 1 - 10^{-q_{ij}/10}.$$

Even though *Phred* scores are valuable information for the construction of consensus, they are not the complete picture of measurement error. In general, for $a \neq b \in \mathcal{A}$, we have

$$\begin{aligned} \mathbf{Pr}(X = b \,|\, Y = a) &= \mathbf{Pr}(X = b \,|\, Y = a, \mathcal{E})\mathbf{Pr}(\mathcal{E} \,|\, Y = a) \\ &= \mathbf{Pr}(X = b \,|\, Y = a, \mathcal{E}) \cdot \varepsilon. \end{aligned}$$

Chromosome	A	G	C	C	T	A	G	A	T	T	C
direct	A	G	C	C	C	A	G	A	ϕ	ϕ	ϕ
direct	A	G	C	C	T	A	G	N	T	$-$	ϕ
reverse	$\tilde{N}$	$\tilde{G}$	$\tilde{C}$	$\tilde{C}$	$\tilde{T}$	$\tilde{A}$	$\tilde{G}$	$\tilde{A}$	$\tilde{T}$	$\tilde{T}$	$\tilde{G}$
reverse	$\tilde{\phi}$	$\tilde{G}$	$\tilde{C}$	$\tilde{C}$	$\tilde{T}$	$\tilde{A}$	$\tilde{G}$	$\tilde{A}$	$\tilde{-}$	$\tilde{T}$	$\tilde{C}$
direct	ϕ	ϕ	N	C	T	A	G	A	T	T	C

Figure 1. An illustrative example of the problem. The bases with a $\sim$ sign represent their complementary bases.

We denote sequencing error rates, conditional on event $\mathcal{E}$, by $w(b\,|\,a) = \mathbf{Pr}(X = b\,|\,Y = a, \mathcal{E})$ for $a \neq b$, and arrange them in the following table:

w	A	C	G	T	$-$
A		$w(C\,|\,A)$	$w(G\,|\,A)$	$w(T\,|\,A)$	$w(-\,|\,A)$
C	$w(A\,|\,C)$		$w(G\,|\,C)$	$w(T\,|\,C)$	$w(-\,|\,C)$
G	$w(A\,|\,G)$	$w(C\,|\,G)$		$w(T\,|\,G)$	$w(-\,|\,G)$
T	$w(A\,|\,T)$	$w(C\,|\,T)$	$w(G\,|\,T)$		$w(-\,|\,T)$
$-$	$w(A\,|\,-)$	$w(C\,|\,-)$	$w(G\,|\,-)$	$w(T\,|\,-)$	

where $\{w(b\,|\,a)\}$ satisfy

$$\sum_{b\in\mathcal{A},\, b\neq a} w(b\,|\,a) = 1 \text{ and } w(b\,|\,a) \geqslant 0 \quad \text{for } b \neq a.$$

The sequencing error rates relate to the conditional probabilities as follows.

$$\mathbf{Pr}(X = b\,|\,Y = a) = \begin{cases} \varepsilon w(b\,|\,a) & \text{if } a \neq b, \\ 1 - \varepsilon & \text{if } a = b. \end{cases} \qquad \mathbf{2}$$

Since *Phred* scores provide only partial information about sequencing error rates, we need to estimate the rest. For the sake of simplicity, we skip the issue of fragment orientation when we describe the sequencing error models.

Conditional sequencing error model

Our perspective is to incorporate *Phred* quality scores into the Churchill–Waterman model (4). We first adopt the parameterization in Equation 2 to model sequencing error, and refer to it as the conditional sequencing error model. The parameters θ in this model include the composition probability $\{\alpha_a\}$ and the conditional sequencing error rates $\{w(b\,|\,a)\}$. The likelihood of the assembly and underlying sequence is given by

$$\begin{aligned} &\left[\prod_{j=1}^{n}\prod_{i=1}^{m}\mathbf{Pr}\left(X_{ij} = x_{ij}\,|\,S_j;\theta\right)\right]\cdot\prod_{j=1}^{n}\mathbf{Pr}\left(S_j;\theta\right) \\ &= \left[\prod_{j=1}^{n}\prod_{i=1}^{m}\left[\left(1-\varepsilon_{ij}\right)^{\left\{1\left(S_j = x_{ij}\right)\right\}}\cdot\left(w\left(x_{ij}\,|\,s_j\right)\cdot\varepsilon_{ij}\right)^{\left\{1\left(S_j \neq x_{ij}\right)\right\}}\right]\right] \\ &\quad\times\prod_{j=1}^{n}\mathbf{Pr}\left(S_j;\theta\right). \end{aligned}$$

Since $\{S_j\}$ are missing, we estimate the parameters by the EM algorithm. The following form of log-likelihood is easy for imputing the sufficient statistics.

$$\begin{aligned} &\sum_{a\in\mathcal{A}}\left\{\sum_{a\in\mathcal{A}/\{a\}}\left[\sum_{i=1}^{m}\sum_{j=1}^{n}1\left(x_{ij} = b, S_j = a\right)\cdot\log\left[w(b\,|\,a)\varepsilon_{ij}\right]\right]\right. \\ &\quad+\left[\sum_{i=1}^{m}\sum_{j=1}^{n}1\left(x_{ij} = b, S_j = a\right)\cdot\log\left(1-\varepsilon_{ij}\right)\right] \\ &\quad\left.+\sum_{j=1}^{n}1\left(S_j = a\right)\log\alpha_a\right\} \qquad \mathbf{3} \end{aligned}$$

Logistic model

From a regression perspective, we take *Phred* scores as a covariate. We denote

$$\mu\left(b\,|\,a, q_{ij}\right) = \mathbf{Pr}\left(X_{ij} = b\,|\,S_j = a;\, q_{ij}\right), \quad a, b \in \mathcal{A}. \qquad \mathbf{4}$$

We assume that base-calling error rates follow a logistic form:

$$\log\left(\frac{\mu(b\,|\,a,q)}{\mu(a\,|\,a,q)}\right) = \beta_{a,b,0} + \sum_{l=1}^{L}\beta_{a,b,l}h_l(q), \quad b \in \mathcal{A}/\{a\},$$

where each covariate $h_l(q)$ is a function of quality score q and takes the form

$$h_l(q) = (q - o_l)_+ = \begin{cases} 0, & \text{if } q \leqslant o_l, \\ q - o_l, & \text{otherwise.} \end{cases}$$

Notice that each function has a knot o_l, where $0 \leqslant o_1 < o_2, \ldots < o_L < Q$. Thus each regressor is a piecewise linear function of the quality score, which allows us to approximate any potential non-linear effect. Equivalently, base-calling rates can be represented as:

$$\begin{cases} \mu(b\,|\,a,q) = \dfrac{\exp\left\{\beta_{a,b,0} + \sum_{l=1}^{L}\beta_{a,b,l}h_l(q)\right\}}{1 + \sum_{c\in\mathcal{A}/\{a\}}\exp\left\{\beta_{a,c,0} + \sum_{l=1}^{L}\beta_{a,c,l}h_l(q)\right\}}, \\ \qquad b \in \mathcal{A}/\{a\}, \\ \mu(a\,|\,a,q) = \dfrac{1}{1 + \sum_{c\in\mathcal{A}/\{A\}}\exp\left\{\beta_{a,c,0} + \sum_{l=1}^{L}\beta_{a,c,l}h_l(q)\right\}}. \end{cases}$$

Similar to Equation 3, this parameterization leads to the following form of log-likelihood function for the assembly and the underlying sequence, up to a term only relating to parameters.

$$\begin{aligned} &\sum_{a\in\mathcal{A}}\left\{\left[\sum_{b\in\mathcal{A}/\{a\}}\sum_{i=1}^{m}\sum_{j=1}^{n}1\left(x_{ij} = b, S_j = a; q_{ij}\right)\right.\right. \\ &\qquad\times\log\left(\frac{e^{\left\{\beta_{a,b,0}+\sum_{l=1}^{L}\beta_{a,b,l}h_l\left(q_{ij}\right)\right\}}}{1 + \sum_{c\in\mathcal{A}/\{a\}}e^{\left\{\beta_{a,c,0}+\sum_{l=1}^{L}\beta_{a,b,l}h_l\left(q_{ij}\right)\right\}}}\right)\Bigg] \\ &\qquad+\left[\sum_{i=1}^{m}\sum_{j=1}^{n}1\left(x_{ij} = a, S_j = a;\, q_{ij}\right)\right. \\ &\qquad\left.\times\log\left(\frac{1}{1 + \sum_{c\in\mathcal{A}/\{a\}}e^{\left\{\beta_{a,c,0}+\sum_{l=1}^{L}\beta_{a,b,l}h_l\left(q_{ij}\right)\right\}}}\right)\right] \\ &\qquad\left.+\sum_{j=1}^{n}\mathbf{1}\left(S_j = a\right)\log\alpha_a\right\}, \end{aligned}$$

where θ represents all the unknown parameters.

Parameter estimation and EM training algorithm

In both the conditional sequencing error model and the logistic model, the underlying sequence $\{s_j\}$ is unknown. Its reconstruction relies on the estimates of parameters in the models. On the other hand, algorithms of estimating parameters are

well established when $\{s_j\}$ are known. Thus we use the EM algorithm to train the model iteratively. In the E-step, we impute the sufficient statistic from observations at the current parameter value and in the M-step, we update the maximum likelihood estimate using the current imputed values of missing data. In the case of the conditional sequencing error model, the parameters are estimated by counting frequencies. In the case of the logistic model, the likelihood can be decomposed into five independent logistic regression models (6). Consequently, we run re-weighted least squares to estimate the parameters (7). The mathematical and technical details are rather lengthy and tedious, and will be published in our technical report (8).

Consensus and quality values

According to the logistic model, the distribution of nucleotides at each position is given by:

$$\mathbf{Pr}\left(S_j = a \mid \{X_{ij} = x_{ij}\}; \{q_{ij}\}\right) = \frac{\alpha_a \prod_{i=1}^{m} \left[(1 - r_i) \cdot \mu\left(x_{ij} \mid a, q_{ij}\right) + r_i \cdot \mu\left(\tilde{x}_{ij} \mid \tilde{a}, q_{ij}\right)\right]}{\sum_{b \in \mathcal{A}} \alpha_b \prod_{i=1}^{m} \left[(1 - r_i) \cdot \mu\left(x_{ij} \mid b, q_{ij}\right) + r_i \cdot \mu\left(\tilde{x}_{ij} \mid \tilde{b}, q_{ij}\right)\right]}.$$

As shown in the formula, the issue of orientation can generally be dealt with by the orientation indicators $\{r_i\}$ and the complementary operator $\sim$. In our convention, we observe $\tilde{x}_{ij}$ directly when a fragment is in reverse orientation. After we plug in the estimated value of θ, we define the consensus at one position and its quality score by maximizing the above probability.

Parsimonious representation and model selection

Although we can include piecewise linear functions at all possible knots in the logistic regression model (Equation 4), we seek a parsimonious model for several purposes. First, we would like to avoid potential overfitting, especially when the size of assembly is not large. Second, a parsimonious model may give us insights into quality scores.

The selection of knots is nothing but a model selection problem. To compare different models, we need an evaluation criterion. Based on quality scores, each fitted model defines a set of error rates, which in turn can be used to construct a consensus. If the truth is known, we can calculate SBD for a model (9). SBD is thus one criterion for model comparison.

A practical solution to model selection ought to be self-evident from data. One such criterion is BIC (10). It is defined as

$$\text{BIC} = -\text{log-likelihood of assembly} + \frac{1}{2}(\#\text{ parameter})\log(\#\text{ observation}).$$

That is, BIC penalizes log-likelihood by model complexity in terms of the number of parameters. For a logistic model with L knots, we have $20(L + 1)$ parameters. We calculate the BIC score for each model and choose the one that minimizes the quantity. The idea is to trade off goodness of fit and model complexity. Computationally, it is intensive to evaluate every model. We adopt the backward deletion strategy used in regression analysis to search for the optimal model (7).

RESULTS

Bias of quality scores

If we do not otherwise specify the data source, the results reported hereafter are based on the *C.jejuni* sequencing data explained earlier. In the conditional sequencing error model, quality scores are interpreted as error probabilities of base-calling. The model that defines the *Phred* scores is determined from a training data set (2,3). When the model is applied to sequencing traces obtained under different working conditions, scores may deviate from probabilities to some extent. We examine this issue on sequencing reads from one BAC. After alignment, we count incorrect base-calls for each value of quality scores—*Phred* scores take integer values from 0 to 64. The observed score for the predicted quality score q is calculated from the assembly by:

$$q_{\text{obs}}(q) = -10 \cdot \log_{10}\left(\frac{\text{Err}_q}{\text{Err}_q + \text{Corr}_q}\right),$$

where Err_q and Corr_q are, respectively, the number of incorrect and correct base-calls at quality score q. In Figure 2, we plot the observed scores against predicted *Phred* scores. When scores are >55, essentially no error is observed. When scores are <20, the prediction is fairly consistent. When scores are between 20 and 55, *Phred* scores overestimate probabilities. Thus calibration is desired for the purpose of improving accuracy of base-calling.

Next we apply the logistic model to the data. Let

$$\varepsilon'_{ij} = \mathbf{Pr}\left(S_j \neq x_{ij} \mid X_{ij} = x_{ij}, q_{ij}; \theta\right),$$

where θ represents all the parameters. Under the assumption of unbiased base-calling, we have:

$$\varepsilon'_{ij} = 1 - \mathbf{Pr}\left(X_{ij} = x_{ij} \mid S_j = x_{ij}, q_{ij}; \theta\right) = 1 - \mu\left(x_{ij} \mid x_{ij}, q_{ij}\right).$$

Then we can assign a new quality score to each base-call x_{ij}:

$$q'_{ij} = -10 \cdot \log_{10} \varepsilon'_{ij}.$$

The bias of this adjusted quality score can be examined by:

$$q_{\text{obs}}(q') = -10 \cdot \log_{10}\left(\frac{\text{Err}_{q'}}{\text{Err}_{q'} + \text{Corr}_{q'}}\right),$$

where $\text{Err}_{q'}$ and $\text{Corr}_{q'}$ are respectively the number of incorrect and correct base-calls at the adjusted quality score q'. We plot the observed against the corrected quality score in Figure 3. Compared with Figure 2, we see that the corrected quality score is more consistent with the observed quality score. After adjustment, no error occurs above score value 42.

The CEQ software that comes along with Beckman sequencers offers quality values similar to *Phred* scores (11). However, their scores are trained from a smaller data set when compared with *Phred*. As in Figure 2, we plot the observed scores against the predicted CEQ scores in Figure 4. The overestimate pattern is seen across almost the entire region. Then we apply the adjustment procedure to correct for the obvious bias. The training data set is ~500 kb. In Figure 5 we plot the

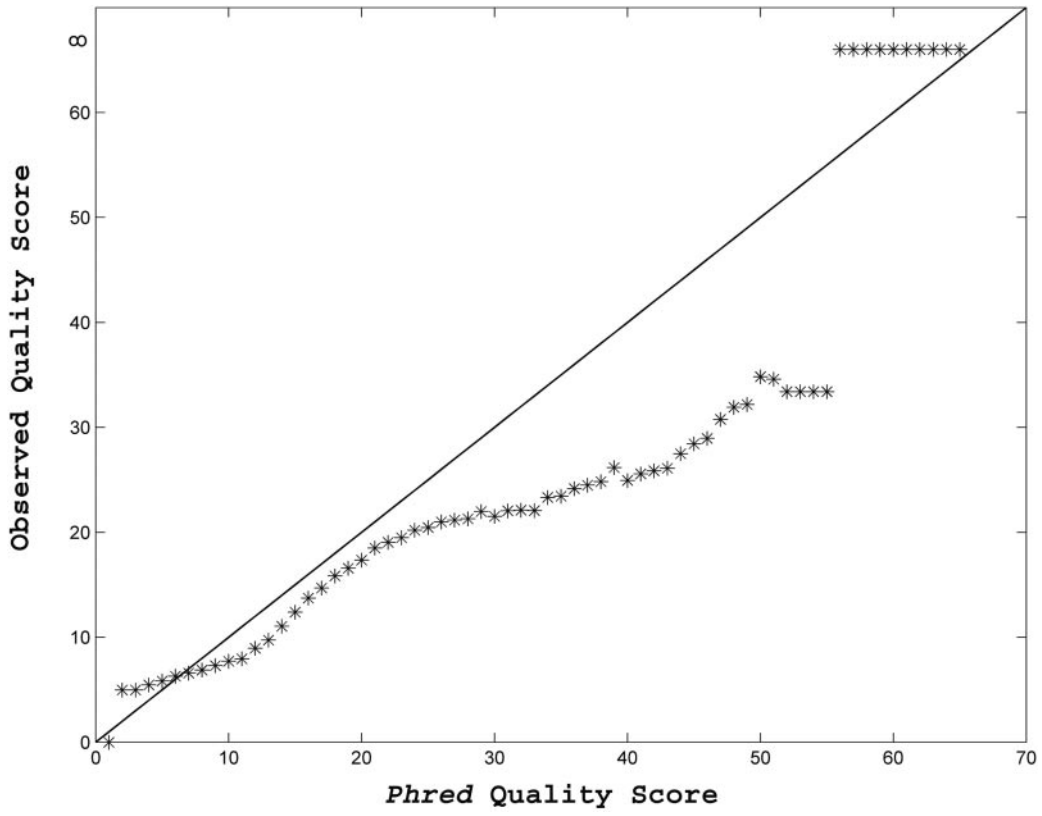

Figure 2. Observed sequencing error rates versus predicted error rates by *Phred* quality score.

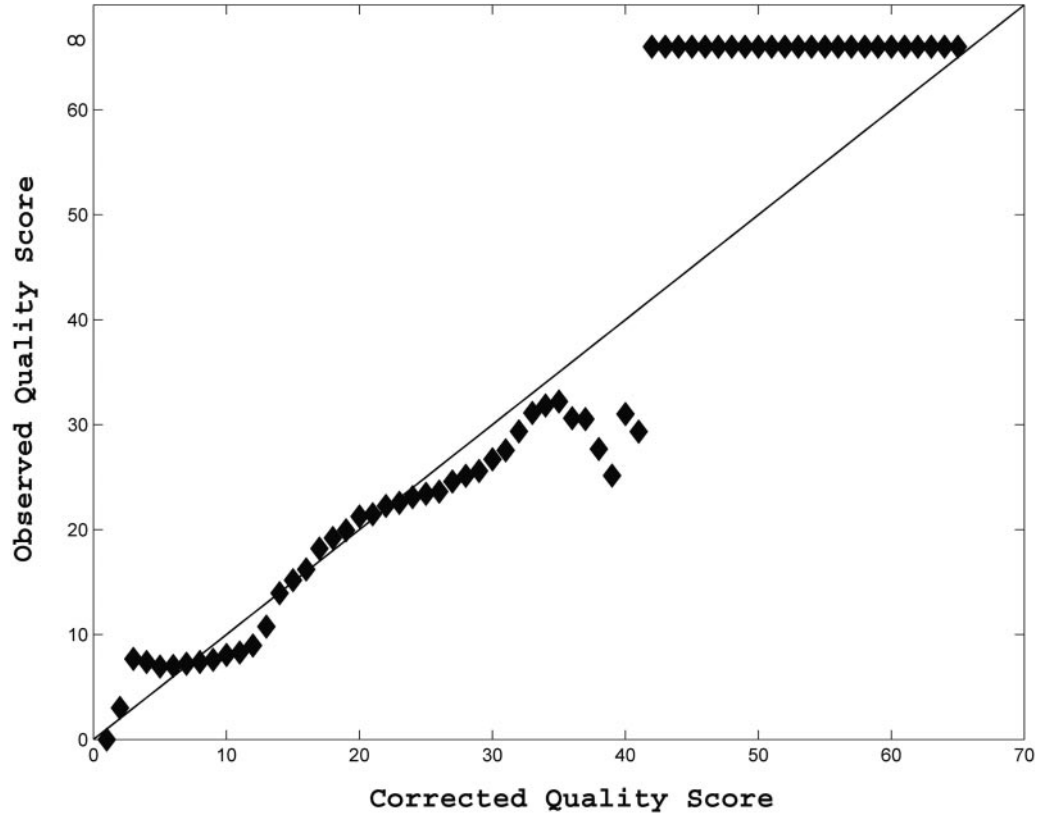

Figure 3. Observed sequencing error rates versus corrected error rates by a logistic model.

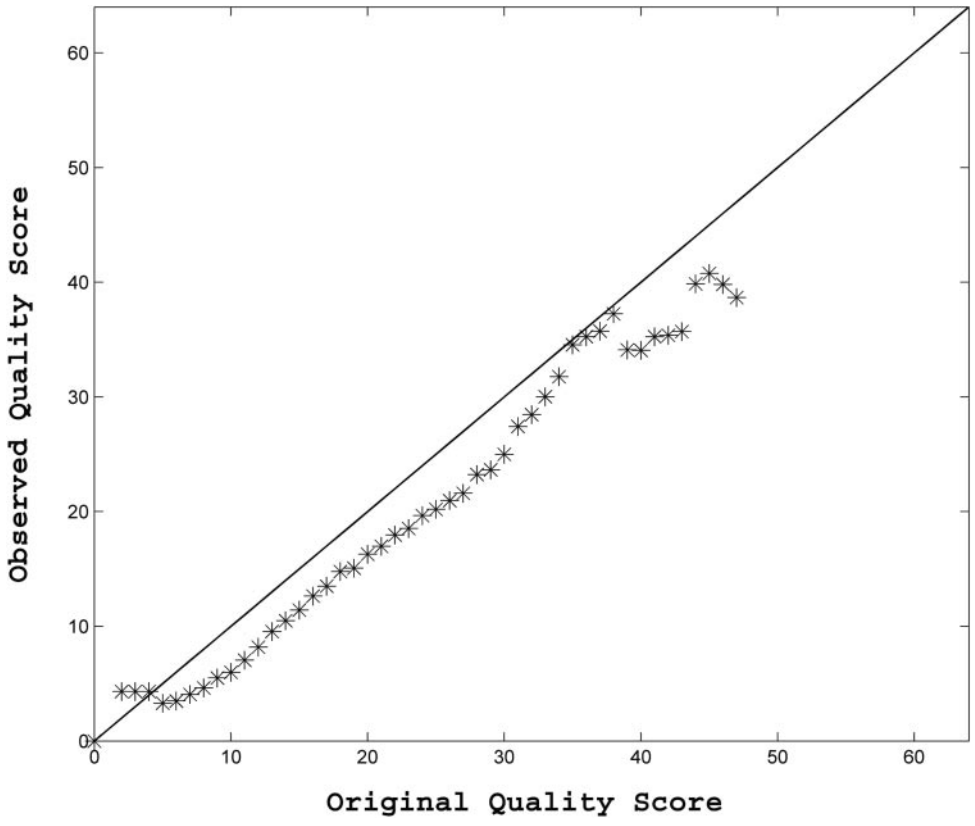

Figure 4. Observed sequencing error rates versus predicted error rates by CEQ Quality Score.

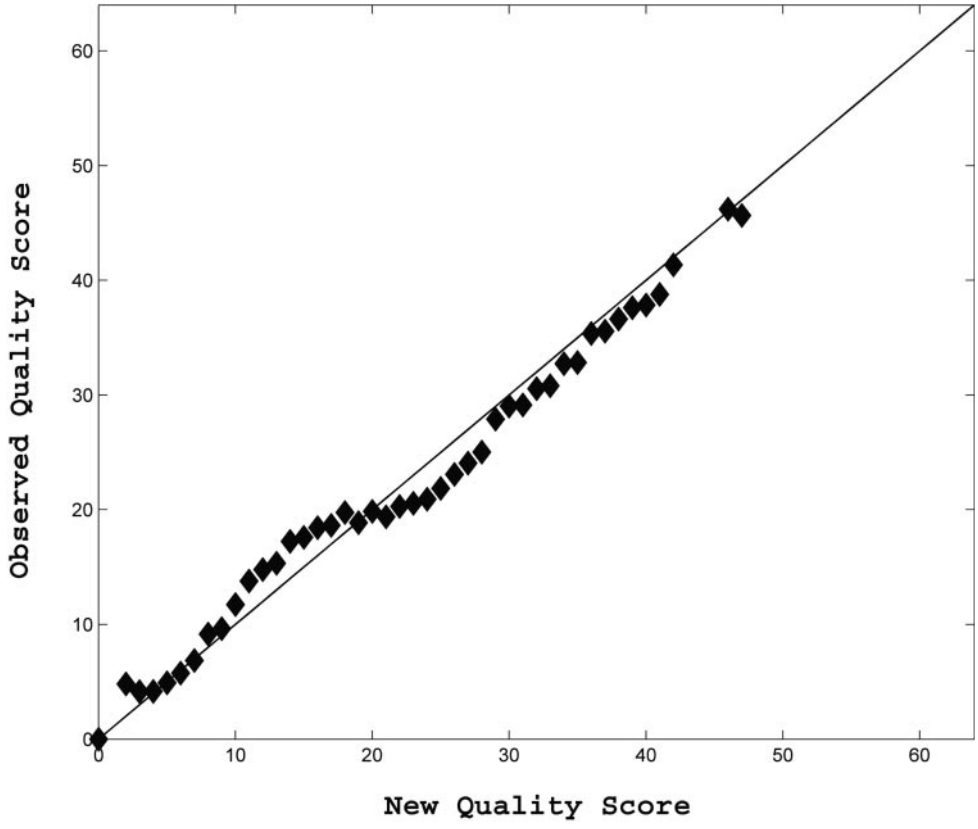

Figure 5. Observed sequencing error rates versus predicted error rates by a logistic model.

observed against the corrected quality scores. This calibration can benefit our haplotype construction project at USC.

Score-dependent error patterns

In the conditional sequencing error model, we assume that the error patterns, or the conditional error rates, are constant regardless of quality scores. To check the assumption, we compare frequencies of each type of sequencing errors at each quality value ranging from 0 to 64. That is, given an assembly, we calculate the empirical conditional error rates as follows:

$$w_{\text{obs}}(b\,|\,a;q)=\frac{\sum_{i,j}\mathbf{1}\left(x_{ij}=b,s_j=a,q_{ij}=q\right)}{\sum_{c\in\mathcal{A}/\{a\}}\sum_{i,j}\mathbf{1}\left(x_{ij}=c,s_j=a,q_{ij}=q\right)},\quad a,b\in\mathcal{A}.$$

When the true base is $\mathcal{A}$, we plot these conditional error rates against quality scores in Figure 6. It indicates that error patterns do depend on quality scores. After we fit a logistic model to the assembly, the conditional error probabilities as a function of quality scores are shown in Figure 7. When quality scores are >55, no sequencing error is observed. Thus conditional error patterns make sense only for scores <55. Many sequencing projects use the Q_{20} rule as a rough measure of the effective length of a DNA read (12). Scores <20 indicate low quality regions. As we can see, error patterns change significantly at scores ~20–24. Since we do not have many bases with high scores, the inference in the high quality range is less reliable than that in the low quality range. When quality scores are <20, C and G are similar to each other; when the scores are >24, a totally different error pattern is observed. This by-product of the parsimonious model offers another perspective of the Q_{20} rule.

Comparison of different methods

We have introduced a conditional sequencing model and a logistic model. In the literature, two different methods exist to estimate sequencing error rates. On the one hand, the method proposed by Churchill and Waterman (4) relies only on assembly but not on quality scores, and we refer to it as the simple probability model. On the other hand, we can use *Phred* scores and simply assign equal error chances among bases. Hereafter, we refer to it as the simple quality score method. In Table 1, we compare the performance of these methods using the *C.jejuni* data set. The majority rule defines the consensus by choosing the most frequent nucleotide at each position. Compared with the majority rule, the simple probability model reduces errors by one-quarter, not resorting to any other information other than the assembly itself. The simple quality score method cuts errors by more than half. The gain is from the training data set that defines *Phred* scores. The conditional sequencing error model reduces errors further. The best result, 346 SBDs, is achieved by the logistic model with five knots. BIC selects a logistic model with three knots that has 348 SBDs. The likelihood scores for these

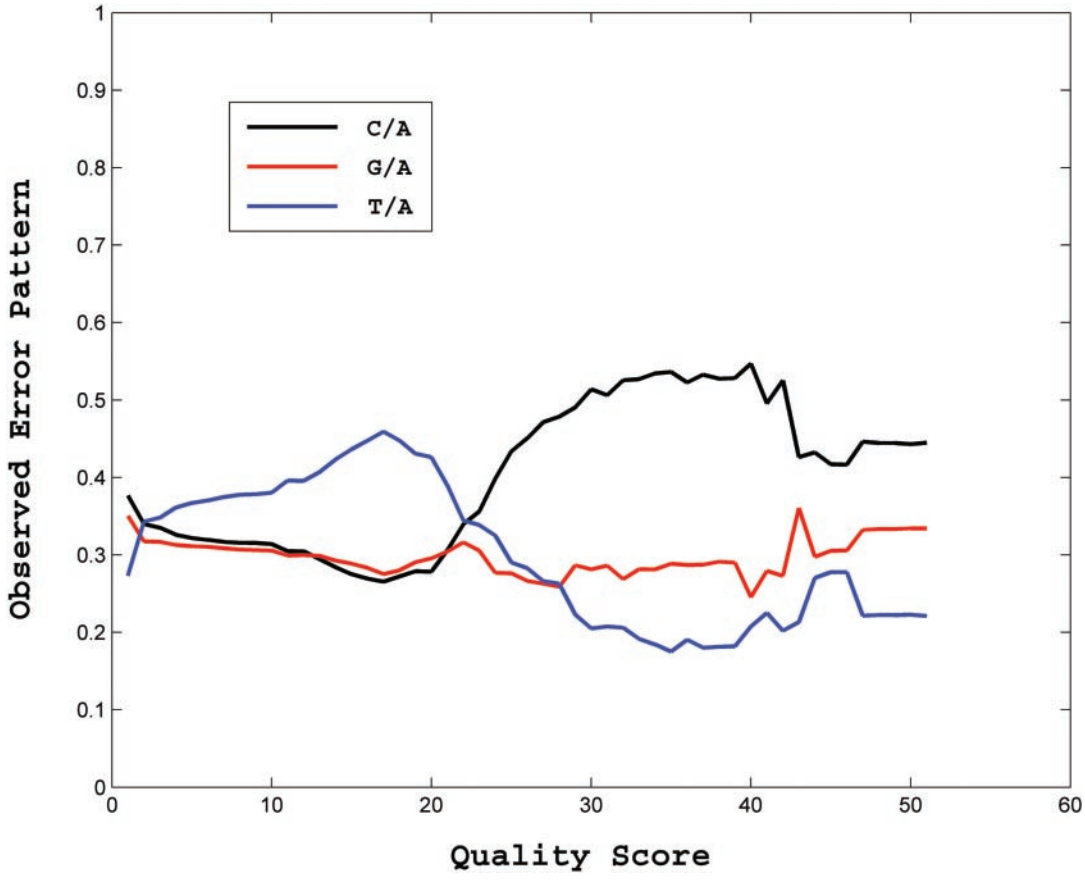

Figure 6. Observed score-wise conditional error rates. The true base is A.

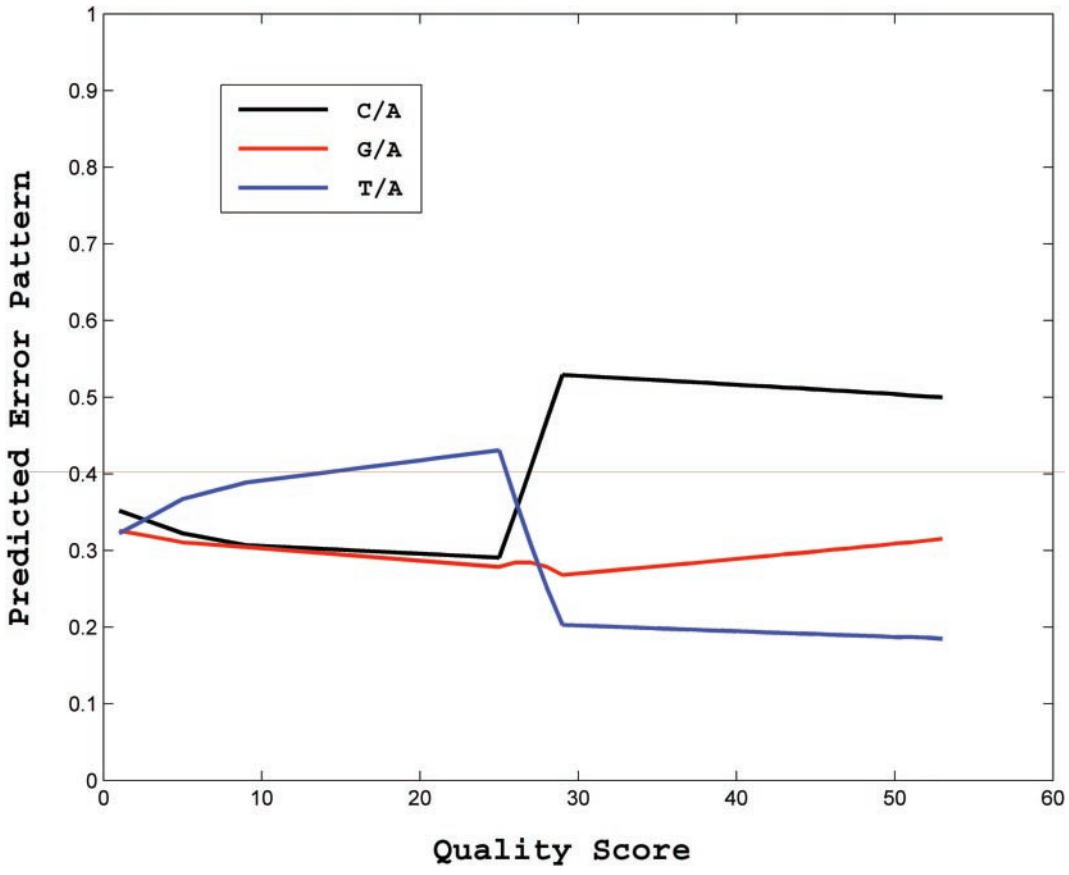

Figure 7. Conditional error rates predicted from a logistic model. The true base is A.

Table 1. Comparison of different methods

Method	SBD	Log-likelihood of assembly
Majority rule	810	
Simple probability	591	−339704
Simple quality score	367	−292781
Conditional sequencing error	358	−281411
Logistic (five knots)	346	−272341

The majority rule is a straightforward counting strategy; the simple probability model is the method proposed by Churchill and Waterman (2); the simple quality score method uses *Phred* scores and assigns equal error chances among bases; the conditional sequencing model uses *Phred* scores and estimates error pattern from data by an EM algorithm; the logistic model predicts sequencing errors by *Phred* scores.

models are also shown in Table 1. The likelihood of a model measures its goodness of fit to the data. For the same data set, we slightly perturb the *Phred* scores associated with the called bases, and errors resulting from the simple quality score method increase substantially from 367 to 517 while the performance of the logistic method remains almost the same.

DISCUSSION

Alignment algorithm

We have observed that different alignment algorithms may produce slightly different assembly matrices. Our adjustment of scores is adaptive to alignment in the sense that it optimizes performance based on each assembly. When a new alignment procedure is used, adjustment may change correspondingly.

Phrap (see http://www.phrap.org) examines all individual sequences at a given position, and generally uses the highest quality sequence to build the consensus. *Phrap* also uses the quality information of individual sequences to estimate the quality of the consensus sequence. By comparison, our method can be used with any other assembly algorithms. The reconstruction of consensus and definition of quality value are based on a probabilistic model. It can adjust potential bias of quality scores in base-calling.

Computing complexity

In the logistic model, the inner loop computes the parameters in logistic regressions by either the Fisher scoring method or by the Newton–Raphson method. Both methods converge quadratically. The outer loop is an EM procedure, and in general an EM algorithm converges at a linear rate. Thus, the computing complexity hinges on the EM algorithm. Specifically, let κ, L, D_1 and D_2 be the coverage, number of knots, number of Newton iterations and number of EM iterations, respectively; then the complexity is about $O[(n\kappa + L^3D_1)D_2]$, where n is the size of the target DNA. Similarly, the complexity for the conditional sequencing error model is $O(n\kappa D)$, where D is the number of EM iterations.

Repeat patterns

We have checked the errors that are left uncorrected by the procedure described in this article. Almost all of them are from regions with repeat patterns. They can be single-, di- or tri-nucleotide repeats. Situations become even more subtle if two repeats are next to one another. In one example, an A is in the middle of four Cs and is missed. In these cases, it is not appropriate to assume that the sequencing error pattern is independent of local contexts. We are considering more sophisticated models to deal with regions with repeats. Li and Speed (13,14) proposed a parametric deconvolution procedure to improve accuracy of sequencing for regions with repeats.

Size of training data set

We have applied our method to data sets of different sizes. The larger the data set, the greater the number of knots selected in the optimal model. We can achieve satisfactory training with an assembly of size 30 kb and a coverage of six. The result is not sensitive to the 'quality' of the quality scores. In comparison, the training of *Phred* scores requires several hundred million base-calls, and it has been carried out on the sequencing traces generated from ABI sequencers. It is difficult to obtain reliable quality scores for other sequencers by the *Phred* training method if only limited base-calling data are available. In this situation, we can apply the method proposed in this article to adjust the preliminary quality scores obtained under roughly the same conditions and obtain probabilistically meaningful quality scores. Earlier, we reported one such example that calibrates Beckman CEQ quality scores using 500 kb from an *Arabidopsis* re-sequencing project.

ACKNOWLEDGEMENTS

We thank Prof. Michael Waterman for his various suggestions. This work is supported by a NIH CEGS grant to University of Southern California.

REFERENCES

1. Adams,M.D., Fields,C. and Ventor,J.C. (eds). (1994) *Automated DNA Sequencing and Analysis*. Academic Press, London, San Diego.
2. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using *phred*. 2. Error probabilities. *Genome Res*., **8**, 186–194.
3. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using *phred*. 1. Accuracy assessment. *Genome Res*., **8**, 175–185.
4. Churchill,G.A. and Waterman,M.S. (1992) The accuracy of DNA sequences: estimating sequence quality. *Genomics*, **14**, 89–98.
5. Parkhill,J., Wren,B.W., Mungall,K., Ketley,J.M., Churcher,C., Basham,D., Chillingworth,T., Davies,R.M., Feltwell,T., Holroyd,S., *et al*. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.
6. McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Model*, 2nd edn. Chapman and Hall, London.
7. Venables,W.N. and Ripley,B.D. (1994) *Modern Applied Statistics with S-plus*, 2nd edn. Springer, New York.
8. Li,M. and Li,L.M. (2004) *Estimate Sequencing Error Patterns by Logistic Regression Models with Missing Responses*. Technical report, Computational Biology, University of Southern California.
9. Felsenfeld,A., Peterson,J., Schloss,J. and Guyer,M. (1999) Assessing the quality of the DNA sequence from the human genome project. *Genome Res*., **9**, 1–4.
10. Schwarz,G. (1978) Estimating the dimension of a model. *Ann Stat*, **6**, 461–464.
11. Winer,R., Yen,G. and Huang,J. (2002) *Call Scores and Quality Values: Two Measures of Quality Produced by the CEQ® Genetic Analysis Systems*. Beckman Coulter, Inc, Fullerton, CA.
12. Nelson,D.O. and Fridlyand,J. (2003) Designing meaningful measures of real length for data produced by DNA sequencers. In Goldstein,D. (ed.), *Science and Statistics: A Festschrift for Terry Speed*, Vol. 40 of Lecture Notes—Monograph series. Institute of Mathematical Statistics, Beachwood, OH, pp. 295–306.
13. Li,L.M. (2002) DNA sequencing and parametric deconvolution. *Stat Sin*, **12**, 179–202.
14. Li,L.M. and Speed,T.P. (2002) Parametric deconvolution of positive spike trains. *Ann Stat*, **28**, 1279–1301.