BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 22 2011, pages 3085–3092
doi:10.1093/bioinformatics/btr537

Genome analysis

Advance Access publication October 5, 2011

Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection

Michiaki Hamada[1,2,*], Edward Wijaya[1,2], Martin C. Frith[2] and Kiyoshi Asai[1,2]

[1]Graduate School of Frontier Sciences, University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa 277–8562 and [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2–41–6, Aomi, Koto-ku, Tokyo 135–0064, Japan

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Recent studies have revealed the importance of considering quality scores of reads generated by next-generation sequence (NGS) platforms in various downstream analyses. It is also known that probabilistic alignments based on marginal probabilities (e.g. aligned-column and/or gap probabilities) provide more accurate alignment than conventional maximum score-based alignment. There exists, however, no study about probabilistic alignment that considers quality scores explicitly, although the method is expected to be useful in SNP/indel callers and bisulfite mapping, because accurate estimation of aligned columns or gaps is important in those analyses.

Results: In this study, we propose methods of probabilistic alignment that consider quality scores of (one of) the sequences as well as a usual score matrix. The method is based on posterior decoding techniques in which various marginal probabilities are computed from a probabilistic model of alignments with quality scores, and can arbitrarily trade-off sensitivity and positive predictive value (PPV) of prediction (aligned columns and gaps). The method is directly applicable to read mapping (alignment) toward accurate detection of SNPs and indels. Several computational experiments indicated that probabilistic alignments can estimate aligned columns and gaps accurately, compared with other mapping algorithms e.g. SHRiMP2, Stampy, BWA and Novoalign. The study also suggested that our approach yields favorable precision for SNP/indel calling.

Availability: The method described in this article is implemented in LAST, which is freely available from: http://last.cbrc.jp.

Contact: mhamada@k.u-tokyo.ac.jp

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on July 8, 2011; revised on August 24, 2011; accepted on September 20, 2011

1 INTRODUCTION

A number of recent studies have revealed the importance of considering quality scores of (short) reads [produced by next-generation sequencing (NGS) platforms (Ansorge, 2009) e.g. Illumina GA, Roche 454 FLX and ABI SOLiD] in various fundamental analyses on NGS data: mapping (alignment) of reads to reference genomes (Frith *et al.*, 2010a; Lunter and Goodson, 2011; Smith *et al.*, 2008), error correction of reads (Kelley *et al.*, 2010), insertion and deletion (indel) detection from mapped reads (Albers *et al.*, 2011) and gene prediction from reads derived from metagenomic data (Rho *et al.*, 2010). This is because error rates of current NGS platforms are relatively high and the errors cannot be ignored in those analyses.

Several recent studies also suggest that probabilistic alignment techniques estimate more accurate alignment in terms of aligned columns or gaps than conventional maximum score-based alignments. For example, γ-centroid alignment (Frith *et al.*, 2010b) is theoretically suitable for accuracy measures such as the sensitivity and positive predictive value (PPV) with respect to aligned bases (Hamada *et al.*, 2011) (we call it 'Aligned column accuracy', formally defined later). In γ-centroid alignment, the marginal probability of aligned bases is employed. On the other hand, Lunter *et al.* (2008) and Schwartz *et al.* (2005) proposed probabilistic alignments, which use the marginal probabilities of gaps as well as the marginal probabilities of aligned columns. Those probabilistic alignments are, therefore, appropriate for accuracy measures in terms of prediction of gaps (we call it 'Gap accuracy').

When quality scores are given to both (or one of) the sequences to be aligned, probabilistic alignments that consider the quality scores are useful for estimating accurate alignments. There do not exist, however, any studies about this. [Note that Frith *et al.* (2010a) utilized (conventional) maximum score-based alignment, *not* probabilistic alignments, while considering quality scores; (Frith *et al.*, 2010b) employed a probabilistic (γ-centroid) alignment *without* considering quality scores.] In this article, we therefore propose two probabilistic alignment methods that consider quality scores explicitly. The methods are based on *posterior decoding* techniques, which use marginal probabilities that incorporate all the possible alignments, with quality scores. To demonstrate the effectiveness of the proposed methods, we apply them to short read mapping (alignment) to reference genomes.

Mapping short reads is one of the most fundamental information analyses of NGS data (Bao *et al.*, 2011), and a number of mapping algorithms have been proposed including Bowtie (Langmead, 2010; Langmead *et al.*, 2009), SHRiMP2 (David *et al.*, 2011), BWA (Li and Durbin, 2009), Stampy (Lunter and Goodson, 2011), novoalign, MAQ (Li *et al.*, 2008), PerM (Chen *et al.*, 2009) and others (Homer *et al.*, 2009; Jiang and Wong, 2008; Rizk and Lavenier, 2010; Smith *et al.*, 2009). Those mapping tools mainly focus on accurate prediction of the genomic location of mapped reads, and in the evaluation, a mapped read is considered to be correct if it overlaps

*To whom correspondence should be addressed.

the true mapping (we call it 'Mapping accuracy'). This means that the detailed alignment between the mapped read and the reference genome, for example, 'Aligned column accuracy' or 'Gap accuracy', is not always correct even if the mapping is correct. It is possible that probabilistic alignment improves those accuracy measures.

Moreover, in this personal genome era, the importance of accurate detection of SNPs and/or indels has been increasing (Nielsen *et al.*, 2011), because not only SNPs but also indels have been implicated in a number of diseases e.g. Chuzhanova *et al.* (2003). In the 1000 genome project, a number of novel indels was reported (Durbin *et al.*, 2010). There are several studies about detecting SNPs and/or indels: e.g. Dindel (Albers *et al.*, 2011), VarScan (Koboldt *et al.*, 2009), SAMtools (Li *et al.*, 2008, 2009) and others (Krawitz *et al.*, 2010). All the tools require the result of mapping of short-reads as the input. Especially, because SAMtools and VarScan directly employ the mapping results (without re-alignment), the accuracy of read alignments is important to the performance. In this article, we also demonstrate that the proposed probabilistic alignments improve the accuracy of SAMtools and VarScan.

2 METHODS

2.1 Incorporating quality scores into a score matrix

Suppose that we have a (usual) score matrix (e.g. HOXD70/55), $\{S_{a,b}\}_{a,b\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}}$ (where $S_{a,b}$ means a score for nucleotide a aligning with b) and $p(c|b,q)$, a distribution on $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ given (b,q), where b is a nucleotide and q is its quality score. The distribution $p(c|b,q)$ is obtained by using a model of quality scores adopted by each NGS platform.

Then, the score between a nucleotide a and a nucleotide b with a quality score q (which is denoted as $S'_{a,(b,q)}$) is computed by the following formula:

$$S'_{a,(b,q)} = T \times \log(R'_{a,(b,q)}) \tag{1}$$

where $R'_{a,(b,q)}$ is an expected value,

$$R'_{a,(b,q)} = \sum_{c\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \left[\exp\left(\frac{S_{a,c}}{T}\right) P(c|b,q)\right]$$

and T is a scaling parameter, which is computed from the given scoring matrix $\{S_{a,b}\}$ (Yu and Altschul, 2005). The quality scores are usually given by *finite* integers and therefore all the values $\{S'_{a,(b,q)}\}_{a,b,q}$ can be pre-computed beforehand, which reduces the overhead of the computation of an alignment when considering quality scores, compared with the case not considering quality scores. [The score in Equation (1) is rounded to the nearest integer in our implementation.)

Note that this extension of a usual score matrix with quality scores was originally proposed by Frith *et al.* (2010a), and was justified from a Bayesian viewpoint. See also Section 4.3 in the case of considering quality scores of *both* sequences.

2.2 Probabilistic model for alignments

By using the transformed score matrix $\{S'_{a,(b,q)}\}$ and a specific score model (e.g. specified by gap open/extend scores), the score of an individual (local) alignment A between x and y, where each base in y has a quality score, is defined. We denote the score by $S(A)$ in this article. [In other words, $S(A)$ is computed by using a position-specific score matrix (PSSM) for a read, given by Equation (1).]

For a given (local) alignment A, the probability of the alignment A is naturally introduced,

$$p(A|x,y) = \frac{1}{Z}\exp\left(\frac{S(A)}{T}\right) \tag{2}$$

where Z is the partition function: $Z = \sum_A \exp\left(S(A)/T\right)$ in which the sum is throughout all the possible local alignments A between two sequences, which is specified by a local alignment model (e.g. Supplementary Fig. S2). Note that the maximum score alignment achieves the highest probability of the probabilistic model.

2.3 Various marginal probabilities

For the probabilistic distribution of Equation (2), the following marginal probabilities (with respect to the distribution) are efficiently computed by using the forward and backward algorithms e.g. Durbin *et al.* (1998):

- p_{ik} is the marginal probability that a base x_i (i-th base of x) aligns with a base y_k (k-th base of y).
- $q_i^{(x)}$ is the marginal probability that a base x_i aligns with a gap.
- $u_i^{(x)}$ is the marginal probability that x_i belongs to an un-aligned (outer gap) region that is not contained in the local alignment.

The key point is that these probabilities are based on all possible ways of aligning the two sequences, not just one maximum score alignment.

These probabilities are essential in the next section. Because the distribution Equation (2) includes information of the quality scores, all the marginal probabilities reflect the quality scores of sequence y.

2.4 Probabilistic alignments with quality scores

In this study, we propose two probabilistic alignment methods, both of which are able to consider quality scores of (one of) the sequences by using the marginal probabilities described in the previous section.

2.4.1 Probabilistic alignment (I) (γ-centroid alignment) The first method is based on the γ-centroid estimator (Frith *et al.*, 2010b) for the probabilistic model of Equation (2). The γ-centroid estimators are employed in many bioinformatics studies (Hamada *et al.*, 2011).

The γ-centroid alignment is an alignment that maximizes the sum of aligned column marginalized probabilities larger than $1/(\gamma+1)$, that is, it maximizes the following score for alignment A:

$$q(A) = \sum_{x_i\sim y_k\in A} [(\gamma+1)p_{ik} - 1] \tag{3}$$

where $x_i \sim y_k \in A$ means an aligned column (without gaps) in A. The alignment is therefore computed by a Needleman–Wunsch (Needleman and Wunsch, 1970) type dynamic programming (DP):

$$M_{i,k} = \max\begin{cases} M_{i-1,k-1} + (\gamma+1)p_{ik} - 1 \\ M_{i-1,k} \\ M_{i,k-1} \end{cases} \tag{4}$$

where $M_{i,k}$ stores the optimal value of the alignment between two subsequences, $x_1 \ldots x_i$ and $x'_1 \ldots x'_k$. In this case, a local alignment is obtained from the global alignment by removing unaligned regions at the edges. (In other words, the outer gaps in the global alignment are removed.)

To be clear, we *first* perform the forward and backward algorithms to calculate the marginal probabilities, and *then* perform the DP algorithm just described.

The parameter γ adjusts the sensitivity and PPV with respect to aligned columns. When γ is low, the method is conservative and only aligns bases that have high probability p_{ik}: this tends to reduce false positive aligned bases but increase false negatives. When γ is high, the method aligns bases more permissively, which tends to increase false positives while reducing false negatives.

It is known that γ-centroid alignment is theoretically appropriate for accurate prediction of aligned nucleotides (i.e. Aligned column accuracy). [see Hamada *et al.* (2011) for details.] A drawback of the γ-centroid alignment is, however, that it is not always appropriate for gaps. Actually, γ-centroid alignment with small γ tends to contain many gaps. We, therefore, introduce another probabilistic alignment method, which considers the marginal probabilities of both aligned columns and gaps explicitly, in the next subsection.

2.4.2 Probabilistic alignment (II) (LAMA alignment) We propose a probabilistic alignment by maximizing the following score $q(A)$ of the alignment A, which contains the marginalized probabilities of gaps explicitly:

$$q(A) = \sum_{x_i \sim y_k \in A} [2\gamma p_{ik} - u_i^{(x)} - u_k^{(y)}] + \sum_{x_i \sim - \in A} [\gamma q_i^{(x)} - u_i^{(x)}] + \sum_{- \sim y_k \in A} [\gamma q_k^{(y)} - u_k^{(y)}] \quad (5)$$

where $x_i \sim - \in A$ and $- \sim y_k \in A$ mean alignment columns with deletion and insertion in the alignment A, respectively. This alignment can be computed by the following recursive equation:

$$M_{i,k} = \max \begin{cases} M_{i-1,k-1} + 2\gamma p_{ik} - u_i^{(x)} - u_k^{(y)} \\ M_{i-1,k} + \gamma q_i^{(x)} - u_i^{(x)} \\ M_{i,k-1} + \gamma q_k^{(y)} - u_k^{(y)} \\ 0 \end{cases} \quad (6)$$

In this case, a Smith–Waterman-like algorithm (Smith and Waterman, 1981) is utilized in order to obtain a local alignment.

It is noted that this probabilistic alignment is equal to the 'Alignment Metric Accuracy' (AMA) estimator for global alignments (Schwartz *et al.*, 2005), when a global alignment model is considered instead of a local alignment model (in that case, $u_i^{(x)} = u_k^{(y)} = 0$ for all i and k) in the above estimator. Obviously, the alignments given by the AMA estimator are suitable for the accuracy measure, AMA, which explicitly considers gaps.

2.5 Mis-mapping probability

When a DNA read is aligned to a genome sequence, it will often align to multiple locations. We assume, however, that it comes from one location. We therefore calculate a mis-mapping probability for each location, the probability that it is *not* the source of the read. According to Frith *et al.* (2010a), the mis-mapping probability of a read at location i is computed as follows:

$$1 - \frac{\exp(S(A_i)/T)}{\sum_j \exp(S(A_j)/T)} \quad (7)$$

where $S(A_i)$ is the (non-probabilistic) alignment score at location i, $\{A_j\}_j$ is a set of alignments/locations derived from the read and T is the same as in Equations (1) and (2).

The threshold δ for the mis-mapping probability can adjust the sensitivity and PPV of 'Mapping accuracy'. [A lower threshold will achieve higher PPV and lower sensitivity (Frith *et al.*, 2010a).]

2.6 Implementation

To demonstrate our method, we have incorporated it into a large-scale alignment tool, LAST [see Kielbasa *et al.* (2011) for details of the algorithm]. In brief, LAST follows these steps: (i) find seeds (initial matches); (ii) extend a gapless alignment from each seed; (iii) shrink each alignment to a 'core' (maximal run of identical matches); (iv) extend a gapped alignment from either end of each core, using an X-drop algorithm . Probabilistic alignment adds: (v) apply a forward–backward algorithm, within the DP region determined by the preceding X-drop algorithm (Altschul *et al.*, 1997). See Supplementary Section A for details. This implementation has a weakness: probabilistic alignment is not applied to the core. Nevertheless, the method gives useful results (see below). Note that, as the result of our implementation, the 'Mapping accuracy' of probabilistic alignments is expected to be the same as that of the original (non-probabilistic) LAST.

In this study, the gapless and gapped score thresholds of LAST (`-d` and `-e`) were set to 108 and 120, respectively (determined by a small dataset). Then, we calculated mismap probabilities. Finally, we discarded alignments with score <150, or mismap probability [computed by Equation (7)] larger than $\delta = 0.01$. (The match and mis-match scores are set to 6 and -18, respectively.)

3 EXPERIMENTS

3.1 Experimental settings

3.1.1 Accuracy measures In this study, we employed three types of accuracy measure as follows. (i) Mapping accuracy: a mapped read is deemed to be true positive when the read shares at least one alignment column (without gap) with the reference (correct) mapping (alignment); (ii) aligned column accuracy: a predicted aligned column (without gap) is deemed to be true positive when the aligned column is exactly identical to the reference alignment; (iii) gap accuracy: a predicted gap is deemed to be true positive if the gap matches a reference gap within a $+3/-3$ window. This is because the position of gaps in the reference can be ambiguous. [For example, a gap in a homo-polymer has several equivalent placements (Krawitz *et al.*, 2010).] In each evaluation described above, the sensitivity (SEN) and positive predictive value (PPV) are employed: SEN = TP/(TP+FN) and PPV = TP/(TP+FP) where TP, FP and FN are the number of true positives, false positives and false negatives, respectively. It should be emphasized that the 'Mapping accuracy' has been utilized in most of the previous studies about short-read mapping, in which we cannot evaluate the accuracy of the aligned columns and gaps.

3.1.2 Simulated data for read mapping We employ several simulation experiments in this study. To generate simulated reads, we utilized Stampy (v1.0.11) (http://www.well.ox.ac.uk/project-stampy) (Lunter and Goodson, 2011). As the reference genome, human chromosome 1 (hg19) was used. We generated reads of length 76 bp. (This length is after simulating indels.)

The dataset labeled 'SNPn-INDELl' (for $n, l = 0, 1, 2, 3, 4$) consists of 100 000 (single-end) reads, each of which contains n SNPs and one insertion or deletion (continuous gaps) whose length is l. [Note that actual reads given by sequencers are expected to resemble a (weighted) mixture of these datasets.] The ratio between insertions and deletions for simulated reads is set to 1:1.

In addition, we simulated sequencing errors in the reads. To do this we obtained real, non-simulated reads in FASTQ format; extracted their quality scores; attached them to our simulated reads; and then mutated each simulated base with probability implied by the attached quality score. We used the FASTQ dataset from SRR005802, whose quality scores imply a mean error probability of 0.13 and median of 0.03.

3.1.3 Real data for read mapping In these experiments, we used the entire human genome (hg19) as a reference genome, and 1 million reads with quality scores from the short read archive (SRA): SRR003994 (the read length is 36 bp) and SRR005802 (the read length is 76 bp) derived from the 1000 genome project (Durbin *et al.*, 2010).

3.1.4 Simulated data for variant calling We created simulated datasets for evaluating variant calls as follows. The reference genome is a 5 Mb region (chr17:11,200,001-16,200,000) of the human genome (hg19). First we created a variant list by using *dwgsim* (http://sourceforge.net/projects/dnaa/). The total variant rate was set to 0.001, where the SNP and indel rates are 0.0009 and 0.0001, respectively. The probability of indel extension is 0.3. (As a result, the indel length ranges from 1 to 8 bp (average = 1.4 bp, SD = 0.78).) Second, we generated diploid genomes from this variant

list (where SNPs were simulated according to genotypes generated by dwgsim, and indels were simulated in both genomes). Finally, using the diploid genomes we generated reads with sequencing errors using Stampy (Lunter and Goodson, 2011). Similarly to the previous section, the quality scores for the simulated 76 bp reads are taken from real reads (SRR005802) for generating sequencing errors. These settings were used to create 10×, 20× and 40× coverage simulated datasets.

3.1.5 Compared methods We chose the following state-of-the-art mapping tools for comparison with LAST: BWA (ver. 0.5.9rc1) (Li and Durbin, 2009), Bowtie (ver. 0.12.7) (Langmead, 2010) (only for the real dataset because it cannot handle indels in read mapping), Novoalign (in novocraftV2.07.06) (http://www.novocraft.com/), SHRiMP2 (ver. 2.1.1b) (David *et al.*, 2011) and Stampy (ver. 1.0.11) (Lunter and Goodson, 2011).

We tried the `-sensitive` and `-n 10` options in Stampy and BWA, respectively, as well as default options. For the other tools, the default parameters were employed in our experiments.

3.2 Results for simulated data

3.2.1 Considering quality scores substantially improves accuracy of read alignment In Figure 1, we showed comparisons between LAST considering quality scores and LAST without considering quality scores, with respect to the accuracy measures described in the previous section. The results clearly indicate that considering quality scores improved *every* accuracy measure for LAST with/without probabilistic alignments. For both 'Mapping accuracy' and 'Aligned column accuracy', considering quality scores substantially improved the sensitivity (∼15%), although the PPVs were decreased slightly (∼0.5%). For 'Gap accuracy', both the sensitivity and PPV were improved by considering quality scores (∼10% for the sensitivity). Note that this is a simulation where the quality scores are *perfectly* accurate and actual quality scores produced by NGS instruments might be less accurate.

3.2.2 Probabilistic alignments improve 'Aligned column accuracy' and 'Gap accuracy' In Figure 2 (and Supplementary Figs S3–S6), we compared LAST without probabilistic alignment and LAST with probabilistic alignments. Figure 2 shows that the probabilistic alignments improve the sensitivity (∼2 % with respect to Aligned column accuracy; ∼6 % with respect to Gap accuracy) compared with the non-probabilistic alignment, while both achieve similar PPVs.

3.2.3 Usefulness of the γ parameter in probabilistic alignments For Aligned column accuracy, both the probabilistic alignments (I) and (II) with larger γ gradually increase the SENs while slightly decreasing the PPVs (Fig. 2). From a theoretical viewpoint, the probabilistic alignment (I) of Equation (4) (i.e. γ-centroid alignment) is more suitable for 'Aligned column accuracy' than the probabilistic alignment of Equation (6) and maximum score alignment [see Hamada *et al.* (2011) for details]. In our computational experiments (cf. Fig. 2), we observed that the probabilistic alignment (I) achieved slightly higher PPVs than the probabilistic alignment (II).

As mentioned in Section 2, the probabilistic alignment (I) with small γ is not appropriate for 'Gap accuracy'. Actually, γ-centroid alignment with small γ decreases the PPV of gap accuracy,

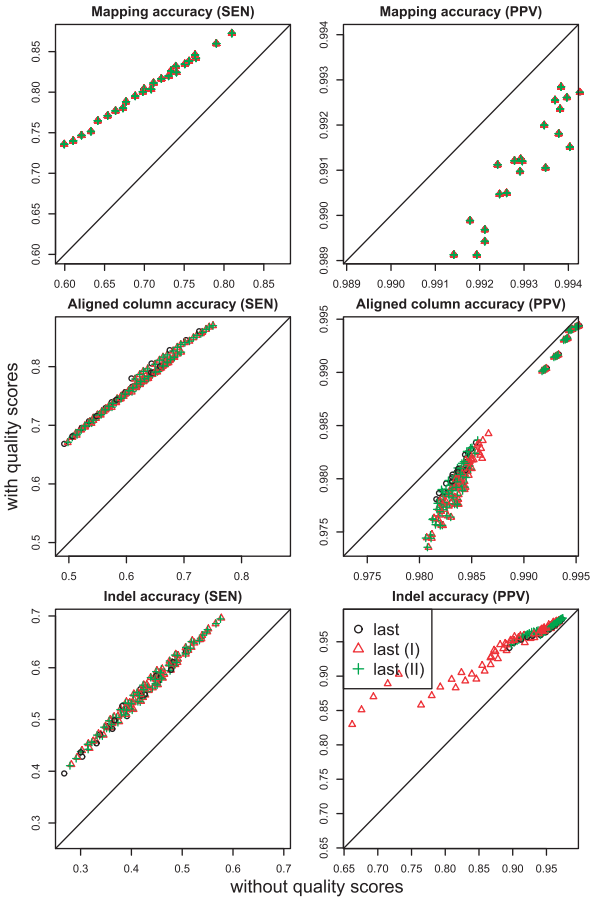

Fig. 1. Comparison of performance when considering and not considering quality scores. The horizontal axis and the vertical axis indicate the performance when not considering quality scores and considering quality scores, respectively. The black circles, red triangles and green crosses show LAST without probabilistic alignment, LAST with probabilistic alignment (I) and LAST with probabilistic alignment (II), respectively. We used dataset of SNPn-INDELm for $n,m=0,1,2,3,4$ (Each point corresponds to one of the dataset). The γ is fixed to 4 in the probabilistic alignments.

because it tends to produce many gaps. On the other hand, the probabilistic alignment (II) of Equation (6) is more appropriate for 'Gap accuracy'. In our computational experiments, we observed those theoretical properties, especially, the probabilistic alignment of Equation (6) achieved slightly better accuracy with respect to 'Gap accuracy' than the γ-centroid alignment.

It should be emphasized that by using smaller δ described in the previous section (which can adjust 'Mapping accuracy') and γ (that adjusts 'Aligned column accuracy' and 'Gap accuracy') values, the probabilistic alignments are expected to achieve arbitrarily low PPVs, which is useful in a number of downstream analyses, such as SNP/indel detection.

3.2.4 Comparison with state-of-the-art methods In Figure 3 (and Supplementary Figs S7–S12), we show the comparison between LAST and existing state-of-the-art methods. The main result is that LAST with or without probabilistic alignment, and Novoalign, performed much better than the other methods. When there are

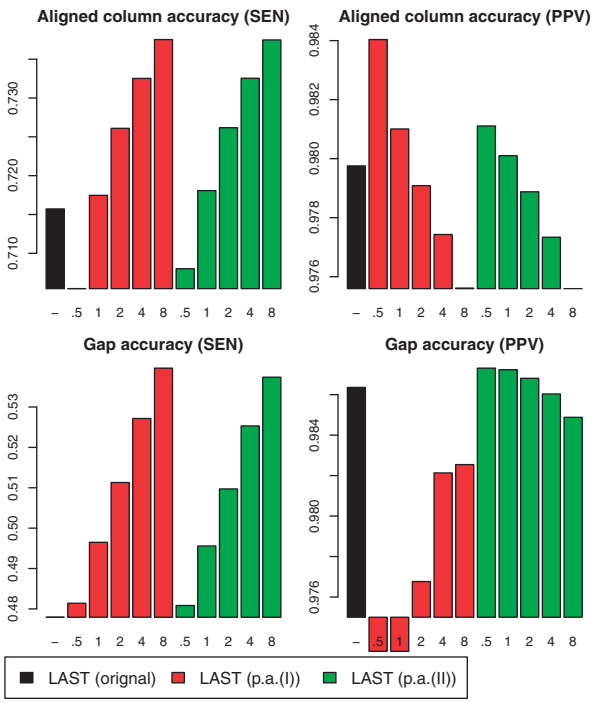

Fig. 2. Comparison between LAST without probabilistic alignment and LAST with probabilistic alignments, for the dataset SNP3-INDEL3 of 76 bp reads. The top row shows aligned column accuracy and the bottom row shows Gap accuracy. The left and right columns show the sensitivity and PPV, respectively. The black bar indicates the performance of LAST without probabilistic alignment, the red bars and the green bars indicate performance with probabilistic alignment of Equations (4) and (6), respectively. The γ has values 0.5, 1, 2, 4 and 8 from left to right in the red/green bars. The values of the 2nd and 3rd columns in Gap accuracy (PPV) are 0.6956 and 0.9010, respectively. The complete results for SNP*n*-INDEL*l* are shown in Supplementary Figures S3–S6.

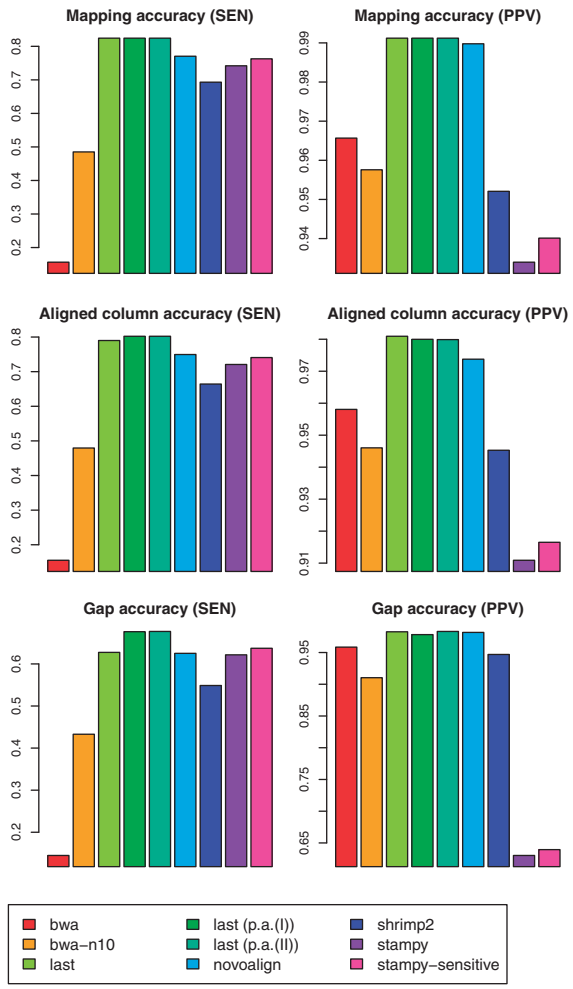

Fig. 3. Comparison between LAST and state-of-the-art mapping tools. Mapping accuracy (top), Aligned column accuracy (middle) and Gap accuracy (bottom) for SNP1-INDEL1 (i.e. *every* read contains 1 SNP and 1 indel of length 1 in addition to sequencing errors) are shown. The first column and the second column indicate the sensitivity and PPV, respectively. We used $\gamma = 4$ for probabilistic alignments in LAST. 'Mapping accuracy' of the original LAST is expected to be almost identical to that of LAST with probabilistic alignments. Note that the scale of vertical axis in PPV is quite different from the one in SEN. The complete results for SNP*n*-INDEL*l* with $n, l = 0, 1, \ldots, 4$ are shown in Supplementary Figures S7–S12.

no indels and few SNPs, Novoalign performed slightly better, otherwise LAST performed slightly better. LAST is much faster than Novoalign (Table 1). It is vital to bear in mind that these methods have parameters that can be varied to increase accuracy at the expense of run time (e.g. lastal's `-m` parameter). Thus, it is hard to draw fundamental conclusions about the methods: we can only say that Novoalign's default parameters are tuned for longer run time in the hopes of higher accuracy, compared with LAST's.

LAST with probabilistic alignment (LAMA) achieved better Gap accuracy than Novoalign in all cases. Stampy sometimes achieved higher sensitivity, but at the expense of poor PPV.

In these tests, BWA exhibited low sensitivity (but somewhat better PPV), because it is designed to be more accurate and faster on queries with low error rates ($\sim$3%) or few differences (cf. Bao *et al.* (2011); BWA achieved good performance with low error rates (Li and Durbin, 2009).).

It is hard to be certain why LAST performs better, because these methods differ in many details. Likely one reason is the efficiency of LAST's adaptive spaced seed algorithm (Kielbasa *et al.*, 2011). Another reason is that LAST models both sequencing errors and real differences in a rigorous way (Frith *et al.*, 2010a), which may give it an advantage when there are real differences.

3.3 Results for real data

In Table 1, the results of real read mapping (alignment) are shown. All methods find fewer alignments for shorter reads, but this effect is greatest for LAST. This may indicate that the sensitivity of LAST, or the PPV of the other methods, suffers for shorter reads. We note that Novoalign has the second-greatest decrease in alignments for shorter reads. Stampy predicted more indels than the other methods. This is because Stampy employs relatively small gap open/extension costs and aims at sensitive prediction of indels. [Actually, the authors mainly evaluate the sensitivity (recall) in their article.] However, considering not only sensitivity but also PPV is important to a number of downstream analyses e.g. SNP/indel calls from the mapping results. Note that LAST with probabilistic

Table 1. Results for real data (1 million reads)

Tool	Mapped	Aligned	Indel (gap)	Time (min)	Mem (GB)
	SRR003994 (36 bp)				
last	341 305	11 741 829	202 (235)	4	15
last(p.a.(I))	341 305	11 927 480	289 (330)	12	15
last(p.a.(II))	341 305	11 927 761	289 (329)	11	15
bowtie	447 671	16 116 156	0 (0)	2	2.2
stampy	561 938	20 078 888	34945 (170938)	51	2.7
stampy(sen)	579 706	20 718 128	34930 (171607)	109	2.7
novoalign	374 023	13 280 253	175 (194)	52	7.8
bwa	237 974	8 566 290	1135 (1163)	5	2.3
bwa(n = 10)	403 341	14 516 588	3647 (4082)	10	2.6
	SRR005802 (76 bp)				
last	768 359	57 280 135	9044 (14417)	41	15
last(p.a.(I))	768 359	57 643 075	10706 (17100)	194	15
last(p.a.(II))	768 359	57 644 959	10580 (16716)	184	15
bowtie	741 446	56 349 896	0 (0)	3	2.2
stampy	836 342	63 341 594	51420 (261364)	72	2.7
stampy(sen)	845 938	64 070 592	51435 (261714)	248	2.7
novoalign	782 291	59 036 774	9047 (14811)	518	7.8
bwa	591 232	44 929 711	6652 (8161)	16	2.3
bwa(n = 10)	726 853	55 230 254	14085 (20427)	67	2.6

Each dataset (SRR003994 of 36 bp; SRR005802 of 76 bp reads) contains 1 million reads with quality scores. The columns 'Mapped', 'Aligned' and 'Indel (gap)' indicate the number of mapped reads, the number of aligned columns and the number of indels (gaps), respectively. The columns 'Time' and 'Mem' shows the elapsed computation time in minutes (not including the indexing time of every tool) and used memory in gigabytes, respectively. The computation was conducted using a machine with Intel(R) Xeon(R) CPU X5550 2.67 GHz and 24 GB memory. We used $\gamma = 4$ for probabilistic alignments. Bowtie cannot handle gaps in read mapping. Due to the huge memory requirement, we gave up executing SHRiMP2 in this experiment.

alignments easily adjusts SEN and PPV with respect to 'Aligned column accuracy' and 'Gap accuracy' by using the γ parameter. ('Mapping accuracy' is adjustable by using the parameter δ in LAST with/without probabilistic alignments.)

One drawback of probabilistic alignment is computational time. Actually, Table 1 indicated that LAST with probabilistic alignments is 3–5 times slower than LAST without probabilistic alignment. However, it is still faster than Stampy (sensitive) and Novoalign, both of which achieved a comparable performance to LAST.

Bowtie was much faster than the other programs, although it could not predict any indels in mapped reads, because the algorithm of Bowtie cannot treat indels. BWA is one of the fastest tools, but it achieves less sensitive predictions than LAST, Stampy and Novoalign. (Note that most of these tools have tunable parameters that enable them to be arbitrarily fast, at the expense of accuracy.)

3.4 Toward accurate SNP/indel detection

Most SNP/indel callers take the result of short-reads mapping as input (Nielsen *et al.*, 2011). In this section, we apply and examine our proposed alignment method as a pre-processing step for SNP/indel calling.

We employed a simulated dataset described in Section 3.1.4. For comparison with the proposed methods, we applied BWA (with default options and the '`-n 10`' option), Novoalign and Stampy (default options). As variant callers, we used SAMtools(-Pileup) (Li *et al.*, 2008) and VarScan (Koboldt *et al.*, 2009), which were widely used in many studies, and directly employ the information of input read alignments. An estimated SNP is considered a true positive if it is in the correct variant list. An estimated indel is deemed to be true positive if it is in the correct variant list within a $+5/-5$ window (we require that the length of the indel must be equal to the length of the reference indel).

In Table 2 and Supplementary Table S2, we show the performance of sequence alignment tools with application to SAMtools and VarScan, respectively. For SNP detection, LAST consistently outperforms BWA, Novoalign and Stampy. LAST also achieves favorable accuracy at low coverage ($10\times$) for indel detection. The results also indicate that both probabilistic alignment (I) and (II) of LAST improve the sensitivity and PPV, compared with LAST without probabilistic alignment in many cases (for predicting both SNPs and indels).

We further observed that indel detection accuracy depends on the treatment of ambiguous indels (e.g. in homopolymer runs). SAMtools and VarScan detect such indels more accurately if the alignment tool places them consistently when aligning different reads. Unfortunately, alignment using quality scores tends to place insertions inconsistently, because the quality scores vary stochastically. Probabilistic alignment can also place gaps inconsistently. To partly compensate for this problem, we tried left-justifying all gaps in the LAST alignments. (That is, we slid each gap to the left, one base at a time, so long as the number of mis-matches in aligned columns did not increase.) Gap justification made almost no difference to SNP prediction but it improved indel prediction (Supplementary Tables S1 and S2). In summary, our alignment methods are clearly promising for SNP/indel detection, but would benefit from better integration with the downstream variant caller.

4 DISCUSSION

4.1 Selecting γ for probabilistic alignments

Both methods (γ-centroid and LAMA) in Section 2.4 have one parameter γ. The γ of the γ-centroid alignment can trade-off between sensitivity and PPV with respect to aligned and unaligned bases. On the other hand, the γ in LAMA alignment can trade-off between aligned regions and unaligned regions in local alignment.

There exist accuracy measures, which consider a balance between Sensitivity and PPV: Mathews correlation coefficient (MCC) and F-score. In RNA secondary structure prediction, Hamada *et al.* (2010) proposed an approximate method to maximize expected MCC/F-score by combining the γ-centroid estimator and *pseudo*-expected MCC. The method is also applicable to the proposed probabilistic alignments.

However, we believe that the optimal γ setting depends on the application in many cases: if algorithms in downstream analysis are robust to false-positive prediction, we can use relatively large γ to obtain sensitive predictions, while a small γ value should be used when we would like to avoid false-positive predictions. A merit of using probabilistic alignments is that we can arbitrarily trade-off sensitivity and PPV.

4.2 Accuracy of quality scores produced by sequencers

Although we assume that quality scores are completely accurate in Equation (1), quality scores produced by current sequencers are

Table 2. Comparison of the effect of mapping tool with SAMtools

Tools	SNPs		INDELs		
	SEN	PPV	SEN	PPV	% mapped
	10× coverage				
bwa	0.0592	0.9044	0.0417	1.0000	14.8
bwa (n = 10)	0.3506	0.9018	0.4093	1.0000	41.5
novoalign	0.5016	0.9541	0.7549	0.9904	89.8
stampy	0.4665	0.9401	0.7500	0.9903	76.4
last(orig) [js]	0.5275	0.9696	0.7353	0.9967	94.0
last(p.a.(I)) [js]	0.5447	0.9686	0.7892	0.9969	94.0
last(p.a.(II)) [js]	0.5447	0.9690	0.7917	0.9969	94.0
	20× coverage				
bwa	0.2327	0.9613	0.2623	1.0000	17.4
bwa (n = 10)	0.5462	0.9931	0.8529	1.0000	45.9
novoalign	0.7224	0.9991	0.8431	0.9942	90.0
stampy	0.6628	0.9990	0.8873	0.9891	78.2
last(orig) [js]	0.7733	0.9994	0.8284	0.9912	93.9
last(p.a.(I)) [js]	0.7856	0.9994	0.8799	0.9917	93.9
last(p.a.(II)) [js]	0.7854	0.9994	0.8652	0.9916	93.9
	40× coverage				
bwa	0.4694	0.9934	0.6936	0.9965	18.8
bwa (n = 10)	0.6523	0.9990	0.9583	0.9949	42.8
novoalign	0.8618	0.9997	0.9069	0.9920	86.5
stampy	0.7724	1.0000	0.9314	0.9819	76.2
last(orig) [js]	0.9080	0.9995	0.9216	0.9895	91.1
last(p.a.(I)) [js]	0.9158	0.9995	0.9436	0.9897	91.1
last(p.a.(II)) [js]	0.9158	0.9995	0.9412	0.9897	91.1

In this experiment, SAMtools (Li *et al.*, 2008) was used as a variant caller. The threshold for consensus quality score was 20. We used $\gamma=4$ for LAST with probabilistic alignments in this table. [js] indicates the results for LAST with gap justification. Complete results (including LAST without gap justification) are shown in Supplementary Table S1. See Supplementary Table S2 for results using VarScan.

sometimes not so accurate. For example, in Illumina sequencers, once the quality score hits the lowest value ('B') in a read, the quality scores of subsequent bases stay the lowest value, even if the intensity values regain higher quality for a base-call. Other sequencer specific errors are also known (Nakamura *et al.*, 2011), which might lead to inaccurate quality scores. Obviously, improvement of base-callers and quality estimation would benefit our methods. Fortunately, pessimistic qualities ('B') are fairly benign because they simply make our methods attach less weight to those bases. The effect is similar to, but less drastic than, trimming those bases from the read. Applying a base quality score recalibration (http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration) is also useful to our methods.

4.3 Extensions of the proposed methods

A possible extension of the probabilistic alignments is probabilistic alignments that consider quality scores of *both* target and query sequences by extending $S_{a,(b,q)}$ of Equation (1) to $S_{(a,q_1),(b,q_2)}$, the score between a nucleotide a with a quality score q_1 and a nucleotide b with a quality score q_2:

$$S'_{(a,q_1)(b,q_2)} = T \times \log(R'_{(a,q_1),(b,q_2)})$$

where $R'_{(a,q_1),(b,q_2)}$ is computed by

$$R'_{(a,q_1),(b,q_2)} = \sum_{c,d \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \left[\exp\left(\frac{S_{c,d}}{T}\right) P(c|a,q_1) P(d|b,q_2) \right].$$

This equation is derived from a Bayesian formula similar to Frith *et al.* (2010a). If the set of quality scores is *finite* (like quality scores produced by current NGS instruments), the overhead of computation (compared with the case without considering quality scores) is expected to be small because $S_{(x,q_1),(y,q_2)}$ for all combinations of x, q_1, y and q_2 can be pre-computed beforehand. Probabilistic alignments considering quality scores of both sequences will be useful in assembly of reads (Paszkiewicz and Studholme, 2010).

Another extension of the proposed probabilistic alignments is to extend them to multiple alignment (Phuong *et al.*, 2006), which can be applied to the multiple alignment of several reads mapped to a similar region. Re-alignment of multiple reads and the reference genome is a more promising approach to SNP/indel detection than re-alignment between each read and the reference genome separately (Homer and Nelson, 2010).

4.4 Further applications of the proposed methods

Recently, Li (2011) has shown that marginal probabilities of aligned columns [they call it 'base alignment quality (BAQ)'] in a read alignment improve the accuracy of SNP detection. That method, however, does not use the marginal probabilities to create the alignment. The marginal probabilities with quality scores proposed in this article will be directly applicable to their method.

Another application of the proposed probabilistic alignment is to map short reads derived from *bisulfite* sequencing (Lister *et al.*, 2008; Meissner *et al.*, 2008), because accurate estimation of aligned columns is important in this case.

5 CONCLUSION

We proposed two probabilistic alignment methods (γ-centroid alignment and LAMA alignment) in which quality scores are naturally considered from a Bayesian viewpoint. To the best of our knowledge, this is the first article that combines probabilistic alignments and quality scores. We implemented the methods in LAST (http://last.cbrc.jp/) and applied them to read mapping and variant calling. Compared with the original LAST, probabilistic alignment slightly improves both 'Aligned-column accuracy' and 'Gap accuracy' in read alignment. (The probabilistic alignment also outperformed existing algorithms.) Moreover, by utilizing a parameter γ, probabilistic alignments can trade-off sensitivity and PPV with respect to aligned columns and/or gaps. This property is useful for various downstream analysis (e.g. variant calling or detection of (un)methylated nucleotides) after read mapping.

ACKNOWLEDGEMENTS

M.H. is grateful to Drs Kana Shimizu and Raymond Wan for useful comments. Thanks are also due to anonymous reviewers for useful suggestions: an example of inaccurate quality scores was provided by one of the reviewers.

Funding: Grant-in-Aid for Scientific Research on Innovative Areas, in part.

Conflict of Interest: none declared.

REFERENCES

Albers,C.A. *et al.* (2011) Dindel: accurate indel calls from short-read data. *Genome Res.*, **21**, 961–973.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ansorge,W.J. (2009). Next-generation DNA sequencing techniques. *Nat. Biotechnol.*, **25**, 195–203.

Bao,S. *et al.* (2011) Evaluation of next-generation sequencing software in mapping and assembly. *J. Hum. Genet.*, **56**, 687.

Chen,Y. *et al.* (2009) PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*, **25**, 2514–2521.

Chuzhanova,N.A. *et al.* (2003) Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.*, **21**, 28–44.

David,M. *et al.* (2011) SHRiMP2: sensitive yet practical short read mapping. *Bioinformatics*, **27**, 1011–1012

Durbin,R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Durbin,R.M. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Frith,M.C. *et al.* (2010a) Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res.*, **38**, e100.

Frith,M.C. *et al.* (2010b) Parameters for accurate genome alignment. *BMC Bioinformatics*, **11**, 80.

Hamada,M. *et al.* (2010) Prediction of RNA secondary structure by maximizing pseudo-expected accuracy. *BMC Bioinformatics*, **11**, 586.

Hamada,M. *et al.* (2011) Generalized centroid estimators in Bioinformatics. *PLoS One*, **6**, e16450.

Homer,N. and Nelson,S.F. (2010) Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol.*, **11**, R99.

Homer,N. *et al.* (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, **4**, e7767.

Jiang,H. and Wong,W.H. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, **24**, 2395–2396.

Kelley,D.R. *et al.* (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.*, **11**, R116.

Kielbasa,S.M. *et al.* (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.*, **21**, 487–493.

Koboldt,D.C. *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

Krawitz,P. *et al.* (2010) Microindel detection in short-read sequence data. *Bioinformatics*, **26**, 722–729.

Langmead,B. (2010) Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics*, **Chapter 11**, Unit 11.7.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,H. (2011) Improving SNP discovery by base alignment quality. *Bioinformatics*, **27**, 1157–1158.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lister,R. *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Lunter,G. and Goodson,M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.*, **21**, 936–939.

Lunter,G. *et al.* (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, **18**, 298–309.

Meissner,A. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.

Nakamura,K. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.

Needleman,S. and Wunsch,C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Nielsen,R. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**, 443–451.

Paszkiewicz,K. and Studholme,D.J. (2010) De novo assembly of short sequence reads. *Brief. Bioinformatics*, **11**, 457–472.

Phuong,T.M. *et al.* (2006) Multiple alignment of protein sequences with repeats and rearrangements. *Nucleic Acids Res.*, **34**, 5932–5942.

Rho,M. *et al.* (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, **38**, e191.

Rizk,G. and Lavenier,D. (2010) GASSST: global alignment short sequence search tool. *Bioinformatics*, **26**, 2534–2540.

Schwartz,A.S. *et al.* (2005) Alignment metric accuracy. arXiv.org:q-bio/0510052.

Smith,A.D. *et al.* (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, **9**, 128.

Smith,A.D. *et al.* (2009) Updates to the RMAP short-read mapping software. *Bioinformatics*, **25**, 2841–2842.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J .Mol. Biol.*, **147**, 195–197.

Yu,Y.K. and Altschul,S.F. (2005) The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, **21**, 902–911.