Chapter 3

On the Accuracy of Short Read Mapping

Peter Menzel, Jes Frellsen, Mireya Plass, Simon H. Rasmussen, and Anders Krogh

Abstract

The development of high-throughput sequencing technologies has revolutionized the way we study genomes and gene regulation. In a single experiment, millions of reads are produced. To gain knowledge from these experiments the first thing to be done is finding the genomic origin of the reads, i.e., mapping the reads to a reference genome. In this new situation, conventional alignment tools are obsolete, as they cannot handle this huge amount of data in a reasonable amount of time. Thus, new mapping algorithms have been developed, which are fast at the expense of a small decrease in accuracy. In this chapter we discuss the current problems in short read mapping and show that mapping reads correctly is a nontrivial task. Through simple experiments with both real and synthetic data, we demonstrate that different mappers can give different results depending on the type of data, and that a considerable fraction of uniquely mapped reads is potentially mapped to an incorrect location. Furthermore, we provide simple statistical results on the expected number of random matches in a genome (*E*-value) and the probability of a random match as a function of read length. Finally, we show that quality scores contain valuable information for mapping and why mapping quality should be evaluated in a probabilistic manner. In the end, we discuss the potential of improving the performance of current methods by considering these quality scores in a probabilistic mapping program.

Key words Mapping, Short reads, High-throughput sequencing

1 Introduction

1.1 The Read Mapping Problem

Because of the tremendously increased throughput of current sequencing technologies, we are often faced with hundreds of millions of short reads for a single experiment. In most experiments, we want to know the genomic origin of those reads and thus we need to align them to a reference genome. This step is called mapping, because the reads are usually almost identical to their origin in the genome apart from errors introduced by the experimental protocol or the sequencing technology.

Peter Menzel and Jes Frellsen have contributed equally.

Noam Shomron (ed.), *Deep Sequencing Data Analysis*, Methods in Molecular Biology, vol. 1038,
DOI 10.1007/978-1-62703-514-9_3, © Springer Science+Business Media New York 2013

The problem of mapping a read to a genome is an instance of the standard pairwise alignment problem. However, old and proven alignment methods, such as BLAST [1], are almost useless for this task—they are just too slow—and therefore much faster "next-generation" alignment programs are required. Many new algorithms and programs (called mappers) have been developed, which gain their speed from an advanced indexing of the reference genome or the reads and from the assumption of high similarity between the reads and the genome. However, while the success of BLAST was due to a careful consideration of the statistical significance of a match, many of these mappers lack sophistication in the statistical interpretation of the results. This leaves researchers in the unsatisfying situation of not knowing how trustworthy the mappings produced by the various programs really are, especially in the case of very short reads.

In this chapter we:

- Introduce the current state of short read mapping, exemplified on real data sets.
- Present a statistical analysis on the probabilities of short read mapping with mismatches.
- Discuss how to make best use of the available base quality scores in sequencing data in a probabilistic model for read mapping.

In this chapter, we focus on the human genome, although the general considerations are valid for any genome. Most problems with wrongly mapped reads are expected to occur for short reads, that is, reads shorter than 30–40 nt, which is below the maximum read length of several modern sequencing platforms. However, many types of experiments produce short sequences, such as short RNA sequencing and fragmented ancient DNA, or short tags coming from CAGE and ChIP-Seq. Thus, even with an increase of read length in the sequencing instruments, the task of accurately mapping short sequences remains.

1.2 The Current State of Short Read Mapping

Sequencing machines introduce distinct error patterns and experimental protocols add sequential noise from adapters, barcodes, and other sources. On top of that, a mismatch in a read compared to the reference genome can be due to the type of experiment or biological variation like SNPs, RNA editing, and DNA damage. All these error sources result in a certain divergence between the read sequence and its genomic origin, which has to be accounted for by the mapping program. If a read maps to several locations in the genome, we ideally want to know which of these locations is the true origin. Thus, the mapper uses some criterion for this decision or, if it has multiple indistinguishable mappings, reports a read as multiple mapped. For example, the mappers BWA [2], MAQ [3], SOAP2 [4], and Bowtie2 [5] report possible alternative matches.

Often the mapping of a short read is accepted if it is unique, and nonunique matches are discarded. This view is problematic because it is an insufficient criterion for accuracy and it does not allow for a detailed assessment of the quality of the mapping.

In this context, a unique match means a match with fewer mismatches and insertions/deletions (indels) than any other possible match in the genome. More precisely, a score or a distance is calculated for a match by summing integer numbers for matches/mismatches and indels. Once a score or a distance is defined, a unique match is defined as one which is better than any other match, meaning that it has a higher score or a lower distance. One scheme (a distance) is to count the number of mismatches and add a penalty (e.g., 2) for each indel. Another scheme (a score) is to add a number for a match and subtract penalties for mismatches and indels. These schemes do the same: the fewer mismatches and indels, the better; and the main challenge is setting the relative cost of an indel versus a mismatch. `BWA` calculates a distance between read and genome by using a default penalty of 3 for a mismatch, and 11 for gap open and 4 for gap extension. `Bowtie2` uses a scoring scheme, which depends on quality scores (see below). Therefore, the "degree of uniqueness" can be quantified by a mapping quality score as discussed later.

When comparing mappers, their performance is usually assessed using simulated data, where the correct location of the read is known in advance (e.g., [6]). On real data, it is difficult to assess the mapping accuracy, because the correct origin is unknown. We can, however, measure the difference between mapping programs. If they differ in their results, it indicates that some reads are wrongly mapped by at least one of the programs. We decided to compare `BWA` and `Bowtie2` on several publicly available data sets. We selected data sets from various types of experiments in order to investigate the differences in the performance of mappers according to the type of data used. The different data sets had different lengths (ranging from 36 to 70 nt), which also allowed us to compare the performance of the alignment methods on different length ranges. In all cases, the sequencing was done on an Illumina Genome Analyzer platform. Before mapping, we removed barcodes and adapter sequences, if necessary, and trimmed low-quality nucleotides and unknown nucleotides (`N`) at the ends of the reads. *See* **Note 3** for the accession numbers and the detailed mapping protocol. The different data sets used in our mapping example are of the following types:

Ancient DNA sequencing is a type of sequencing where specialized techniques and protocols are used to extract DNA fragments from ancient tissue samples, which have characteristic damage patterns in the DNA [7] and can have large variation in fragment length. Contamination from other species is another important aspect for the mapping problem [8]. This particular sample [9] is from a hair extract and is quite clean with very low levels of damage and contamination.

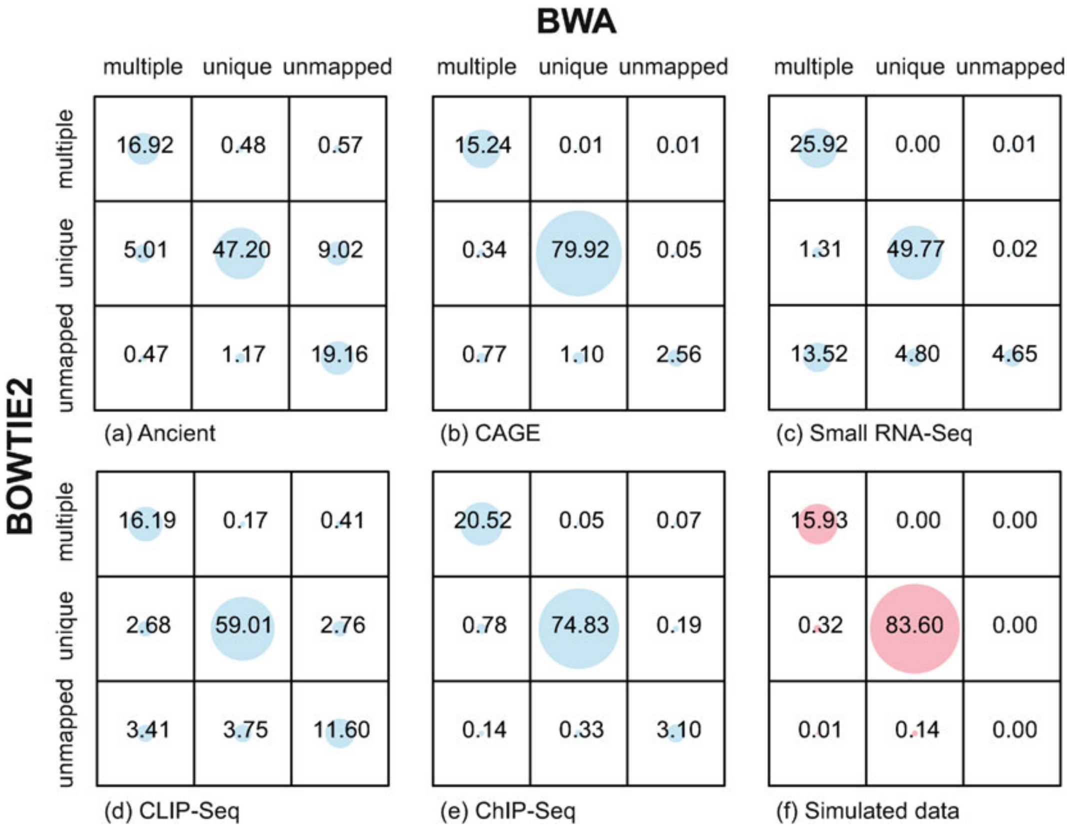
Cap Analysis of Gene Expression (CAGE) is a technology developed to identify transcription start sites (TSS) and measure gene expression. In the CAGE protocol, mRNAs are reversely transcribed to produce cDNAs. Those cDNAs that reach the 5′-end of the gene are selected by cap-trapping and small fragments of ~20 nt from the beginning of the gene are produced by using specific restriction enzymes [10].

Small RNA sequencing is used to measure the expression of micro-RNAs and other types of small RNAs in a sample. RNA is purified and only fragments with a length around 18–30 nt are selected for sequencing. Some of these RNAs can undergo modifications such as addition of nucleotides in the 3′-end [11], which increase the mapping difficulty.

CLIP-Seq was developed to identify the binding sites of RNA-binding proteins in RNAs. Protein–RNA complexes are UV-cross-linked and immunoprecipitated with specific antibodies. The cross-linking process can introduce errors in the reads such as deletions due to the cross-linking protocol [12].

ChIP-Seq is a similar technique to CLIP-Seq and it is used to study the binding of proteins to the DNA. This technique has been applied to identify the binding of proteins that recognize specific sequences in the DNA such as transcription factors or to characterize the location of proteins more loosely associated to the DNA such as Pol-II or histones.

To compare the performance of the two mappers, we calculated the overlap in the classification of reads (as uniquely mapped, multiple mapped, or unmapped). The overlaps for the five data sets are shown in Fig. 1a–e. For comparison, we also show the results obtained on simulated reads (Fig. 1f). Each of the bubble plots represents the agreement between the mappers for a given data set. Each cell shows the percentage of reads with a particular classification by `Bowtie2` and `BWA` (e.g., in Fig. 1a: 5.01 % of reads were classified as uniquely mapped by `Bowtie2` and as multiple mapped by `BWA`). The size of the bubbles represents these percentages. The diagonal (from top left to bottom right) shows the agreement between the two programs, whereas all the other cells show the disagreement. The first thing that we notice is that for different data sets and mappers the amount of reads that can be mapped to the genome varies from 71 to 97 %. This number is reduced to ~50 % in some cases if we only consider uniquely mapped reads. Using simulated data, less than 1 % of the reads are missed and up to 84 % of the reads can be uniquely mapped. When we look at the agreement between the two mappers, we see that it is ranging from 80 % in small RNA-Seq (Fig. 1c) to 98 % in the case of



(a) Ancient (b) CAGE (c) Small RNA-Seq

(d) CLIP-Seq (e) ChIP-Seq (f) Simulated data

Fig. 1 Bubble plots showing the overlap of uniquely mapped, multiple mapped, and unmapped reads between Bowtie2 and BWA in real (**a**)–(**e**) and simulated (**f**) data. The numbers in each cell show the percentage of reads with a particular classification given the mapping of Bowtie2 (*rows*) and BWA (*columns*). The size of each bubble represents the percentage of reads in the cell

ChIP-Seq (Fig. 1e). Interestingly, the agreement in the reads that are mapped uniquely varies from 47 % in ancient DNA sequencing to 80 % in CAGE (Fig. 1a, b). We also have to consider that if a read is reported as uniquely mapped by the two programs it does not mean that it is mapped to the same location in the genome. In the majority of the data sets the location in the genome of the uniquely mapped reads is the same in more than 99 % of the cases. But this may not always be the case, as for instance in the ancient DNA data set, ~8 % of uniquely mapped reads are mapped to different locations by the two programs.

2 Materials: Statistics of Read Mapping

In our example, we saw that the two mappers disagree on some fraction of the reads, and for some data sets the disagreement is quite substantial. This may of course be due to subtle differences in the algorithm or scoring scheme used by the mappers, but it may also be due to some reported matches simply being wrong.

In this section we discuss in more detail the problem of wrongly mapped reads by giving some simple examples, show the statistics, and elucidate the problem with simulated reads.

2.1 Unique Matches Are Not Necessarily Correct

We are going to start with a very simple example showing that a unique match is not necessarily the *correct* match. Imagine that we have an Illumina read originating from an Alu repeat. This is a common situation, since there are around one million Alu repeats in the human genome [13]. It happens that this read has a perfect unique match in the genome, i.e., a single match with no mismatches. In this example, we also assume that there are a relatively large number (n) of matches in the genome for this read that have one mismatch. Now, we want to know the probability that the one perfect match is correct. It is lower than one would probably expect!

Let us assume that there is some probability p that there is either a sequencing error or a polymorphism at any given base (ignoring the fact that the quality depends on the position in the read). With a few other simplifying assumptions we can show (*see* **Note 1**) that the probability for a perfect match being correct is approximately equal to $1/(1 + np/3)$. For example, if there is a 2 % error rate ($p = 0.02$) and there are $n = 150$ genomic matches with exactly one mismatch, then there is a probability of 50 % that the perfect match is correct. If $n = 1,000$, it is as low as 13 %.

Whether the above example is realistic or not depends entirely on the repeat structure of the genome in question. The human genome, on which we focus in this chapter, is quite repetitive, but the repeat structure is complex ranging from ancient duplications over transposable elements to simple repeats or low-complexity regions. It is therefore difficult to theoretically quantify the impact of genomic repeats on the mapping problem, and therefore we address it empirically.

What is the probability that a unique match is wrong? To address this question we generated a large number of reads from the human genome by randomly sampling short sequences from all chromosomes and adding errors. They were generated with different lengths and different error rates. In order to simplify the example, the error rates were independent of the position in read and neither insertions nor deletions were introduced. We then mapped the reads back to the genome using exhaustive mapping with up to three mismatches and without insertions and deletions. This means that no heuristics were used in the mapping and all possible matches in the genome were found.

As can be seen from Fig. 2, a considerable fraction of the reads have unique matches to an incorrect position in the genome. The phenomenon is particularly pronounced for short reads and the largest error is observed for reads of length 18, where nearly 10 % of the reads with error rate 0.05 are uniquely mapped to wrong positions. Even for reads of length 50, more than 1 % of the reads

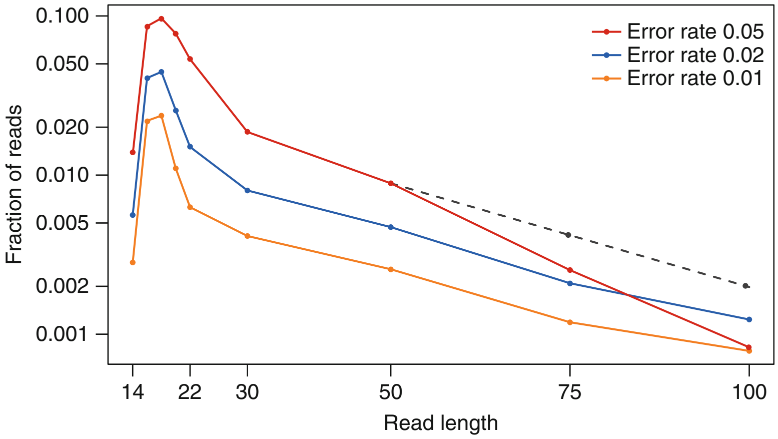


Fig. 2 Fraction of reads from the human genome that map uniquely to an incorrect position in the human genome. We generated three sets of reads from the human genome (hg19), with position-independent error rates of 0.01, 0.02, and 0.05. Every set contains one million reads for each of the lengths: 14, 16, 18, 20, 22, 30, 50, 75, and 100. The reads were mapped to the human genome using exhaustive mapping with up to three mismatches and without insertions and deletions. For every set, the fraction of all reads that have unique matches and are mapped back to an incorrect position in the human genome is shown as a function of the read length. The reads of length 75 and 100 with error rate 0.05 were also mapped using exhaustive mapping with up to six mismatches; these results are shown with grey dashes

with error rate 0.05 are uniquely and wrongly mapped. Although the rates are small for long reads, it is surprising that more than 1 in 1,000 reads of length 100 are uniquely, but wrongly, mapped.

This example clearly illustrates that although all reads are originating from the human genome and match a position uniquely, there is a considerable risk that this position is wrong for short reads. The fact that the fraction of wrongly mapped reads drops faster for 5 % error rate than for 2 % is caused by the reads having more mismatches than accepted in the mapping, as can be seen from the dashed line where we allow up to five mismatches.

For very short reads, there can also be multiple random matches even in a non-repetitive genome. The chance of such a random match is another important consideration, also relating to contamination in the sample, etc. This is the subject of the next section.

2.2 Significance of a Match

In all sequencing experiments some of the reads originate from DNA, which is not present in the reference genome. This may be contamination from other sources of DNA, such as primer/adapter sequences and carrier DNA, or contamination from other species. It may also be regions in the sample DNA which are not present in the reference genome because they have not been sequenced or because of differences between individuals or strains. In metagenomics, for instance, reads may be mapped to all known microbial genomes, but it is likely that a large fraction of the DNA does not originate from any of these. While sample contamination is likely to be most common, one should also keep in mind that a contamination of the reference sequence is also plausible due to misassembly [14].

Therefore, one might expect that some fraction of the reads should not be mapped to the reference genome(s); and if they are mapped, it is a source of error. Here we consider the chance that such foreign sequence maps to a genome. This problem is not a new one. The *E*-value returned by the BLAST program is an estimate of the number of random matches expected in a database of the same size as the reference database. Fortunately, we can estimate the probability that a match is random under reasonable assumptions.

Let us assume that we are mapping reads of length l, which are randomly composed of the four bases with equal probability 1/4. Then the probability of a match to a specific location with exactly k mismatches is given by the binomial distribution

$$f\left(k; l, \frac{1}{4}\right) = \binom{l}{k}\left(\frac{1}{4}\right)^{l-k}\left(\frac{3}{4}\right)^{k}. \tag{1}$$

The probability of a match with up to m mismatches is then just the sum $F\left(m; l, \frac{1}{4}\right) = \sum_{k=0}^{m} f\left(k; l, \frac{1}{4}\right)$. Assuming independence between positions in the genome, the expected number of matches in a genome of length L is

$$E(m, l, L) = L \times F\left(m; l, \frac{1}{4}\right). \tag{2}$$

This is the *E*-value for short read mapping under these assumptions. Note that what we call the length L is really the number of possible match positions, and since a match can be on both strands, it is twice the length of the genome. One can show that the corresponding probability of a random match (the *p*-value) is well approximated by (*see* **Note 2**)

$$p(m, l, L) = 1 - \mathrm{e}^{-E(m,l,L)}. \tag{3}$$

For $L = 6 \times 10^9$, which corresponds to a random "genome" with three billion bp (approximately the same size as the human genome), this probability is shown in Fig. 3 as a function of read length and for a few different values of m. Notice that the read length should be around 28 nt to map with a *p*-value below 0.01 when allowing up to three mismatches. For shorter reads, such as miRNA-sized reads or CAGE tags, it is probably wise to only accept perfectly matching reads, for which we get a *p*-value of less than 0.01 at read length of at least 20 nt.

In the calculations above we assumed independence of positions in the genome, which will not hold in a real repetitive genome. However, the *p*-value is a good indication of what to expect when

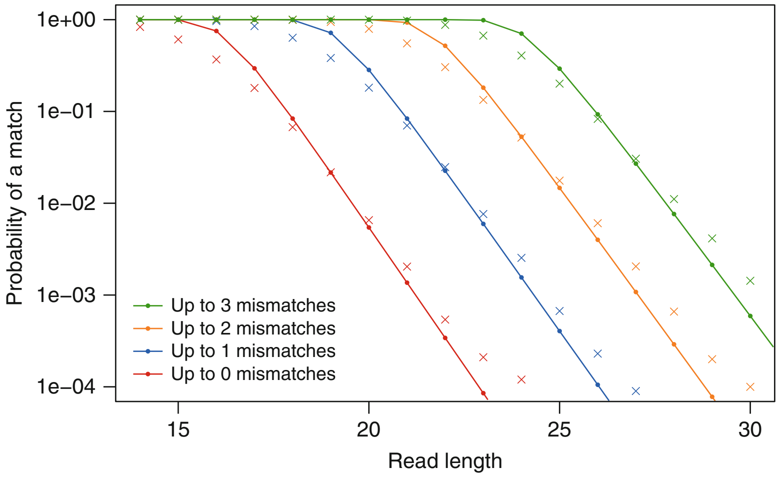


Fig. 3 The probability of a random match. The full lines show the probability of a random match, given by equation (3), as a function of the read length for $L = 6 \times 10^9$, which correspond to a genome of three billion bp. Individual lines are shown for 0, 1, 2, and 3 mismatches in *red*, *blue*, *orange*, and *green*, respectively. The crosses are obtained by generating one million errorless reads from the *E. coli* genome for each of the read lengths: 14, 16, 18,$\cdots$, 30. These reads were mapped to the human genome (hg19) using exhaustive mapping with up to three mismatches and without insertions and deletions. The *red crosses* show the fraction of reads that has a match with up to 0 mismatches for each read length, while the *blue*, *orange*, and *green crosses* show the fractions for up to 1, 2, and 3 mismatches, respectively

dealing with DNA of homogeneous composition. For genomes with nonuniform base composition, such as high or low GC content, the approximation of the *p*-value will be worse in the sense that longer reads are required in order to have significant matches.

It is also of interest to ask for the probability of obtaining a random *unique* match. Although this can be calculated for random sequences, here it suffices to observe that if the probability of obtaining any match in a random sequence is small, then almost all matches will be unique.

We can illustrate the problem of obtaining random matches by mapping reads from the *E. coli* genome (GenBank acc. FM180568) to the human genome (hg19). For practical purposes we can consider these two genomes to be unrelated, and accordingly all matches will be random. From the *E. coli* genome we sampled a large amount of errorless reads with different lengths, and mapped the reads to the human genome, using exhaustive mapping with up to three mismatches. The fraction of these reads with at least one match in the human genome is shown in Fig. 3. There is a remarkable agreement between these points and the theoretical lines in the logarithmic plot. This is partially coincidental, and a close

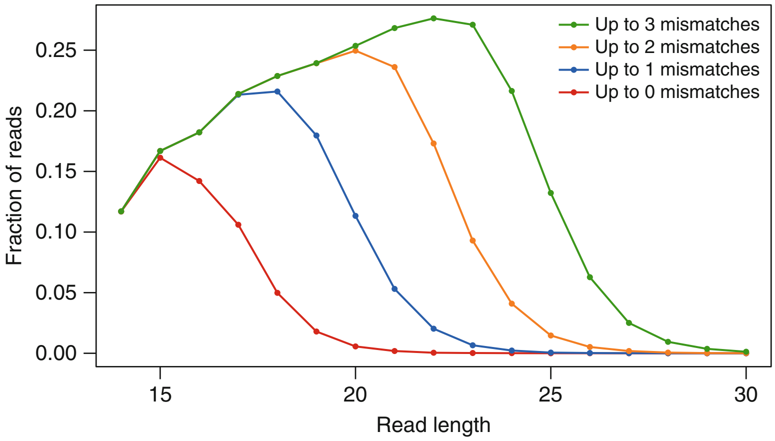


Fig. 4 Fraction of reads from *E. coli* that have unique matches in the human genome. The fraction of all reads that map uniquely to the human genome is shown as a function of the read length. Individual curves are shown for matches with up to, respectively, 0, 1, 2, and 3 mismatches. The same reads and procedure as in Fig. 3 were used in this figure

inspection shows deviations that are due to dependencies between positions in the human genome, and perhaps real similarities between the two genomes.

The fraction of *E. coli* reads that have unique matches in the human genome is shown in Fig. 4. Initially, this fraction increases with read length, because for very short reads there are so many random matches that it is unlikely to find just one unique match. The curves peak at read lengths 15–22 depending on the number of mismatches allowed. From the maximum they decrease towards zero and as the fraction of mapped reads becomes smaller, the curves approximately coincide with the curves in Fig. 3, as we discussed above. For instance, we observe that more than 25 % of all the length 22 nt reads match the human genome uniquely with up to three mismatches.

3 Methods: Probabilistic Read Mapping

We have seen that uniqueness is not a good measure of mapping quality on its own. Ideally, we would like to know the probability that a match is correct, and in this section, we derive such a probability. For this, we can use the per base quality scores.

3.1 Quality Scores

High-throughput sequencing (HTS) data normally include an error probability for each base in a read. Typically, the precision of base calls drops towards the end of the read (Fig. 5b) and the error probabilities allow us to estimate how large the uncertainty for correct base calls becomes. Error probabilities are converted into a range of discrete integers called quality scores or Phred scores,

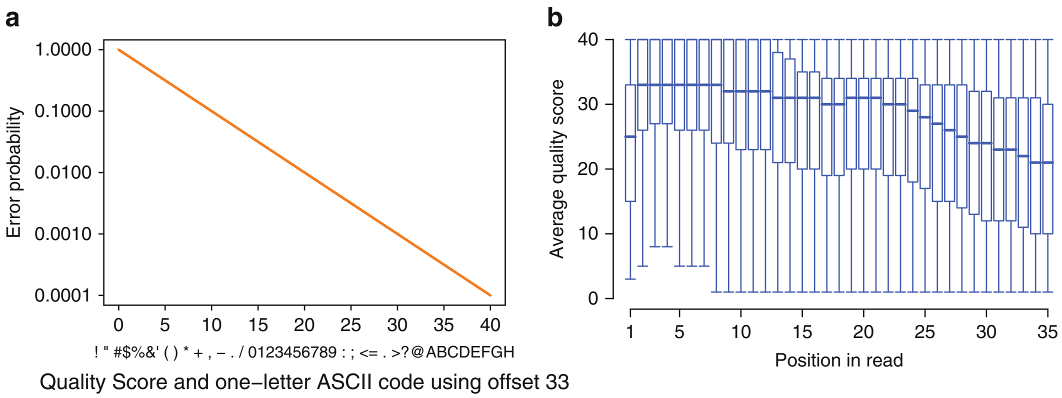


Fig. 5 (**a**) Quality scores: Conversion between error probabilities and Phred-type quality scores. The quality score Q is calculated from the error probability e by $Q = -10 \times \log e$ and can be represented as an ASCII character, here using an offset 33. (**b**) Quality profile: Example of a quality profile for the 35 first nucleotides of the reads from the Illumina CLIP-Seq data set used in the Introduction. The average quality scores decrease towards the 3′-end of the reads

which can be represented as characters from the ASCII code using a certain offset (e.g., 33, which corresponds to "!") to skip non-printable characters at the beginning of the ASCII code. Figure 5a shows the relation between error probabilities and quality scores.

The reads and their corresponding quality scores are usually stored in the FASTQ format [15]. Note that there are several possible ways to calculate the quality scores, depending on the sequencing technology and instrument version. For example, quality scores in Roche/454 reads indicate the probability that the homopolymer length at a given position is correct [16, 17], while Illumina quality scores denote the probability of an incorrect base call at this position. Also different offsets for the conversion to ASCII characters are used, e.g., by different Illumina instruments, so that programs for downstream analysis usually require an explicit setting of the used offset or try to guess it from the quality score characters.

The quality scores contain valuable information about the precision of base calls that can be used in the mapping process. As a simple example, if we have a mismatch between read and genome in the first few bases, we can trust it more to be a real mismatch compared to a mismatch in the low-quality end of a read where the base in the read has higher probability to be a sequencing error. Figure 6 illustrates the possible mapping of a read to two different locations in the genome, where the mapping with two mismatches has a higher probability of being correct than the mapping with only one mismatch.

Most current mappers use the quality scores in one way or another. Typically, programs provide an option to truncate a read at a specified

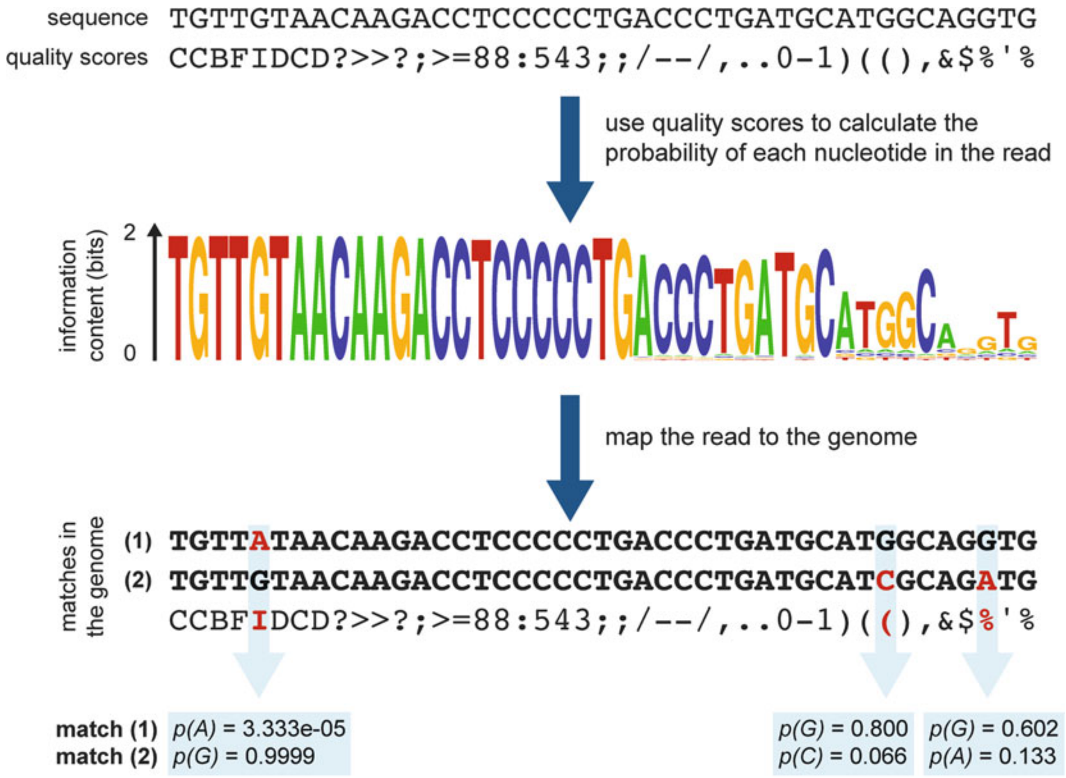

sequence TGTTGTAACAAGACCTCCCCCTGACCCTGATGCATGGCAGGTG
quality scores CCBFIDCD?>>?;>=88:543;;/--/,..0-1)((),&$%'%
use quality scores to calculate the probability of each nucleotide in the read
information content (bits)
2
0
map the read to the genome
matches in the genome
(1) TGTTATAACAAGACCTCCCCCTGACCCTGATGCATGGCAGGTG
(2) TGTTGTAACAAGACCTCCCCCTGACCCTGATGCATCGCAGATG
CCBFIDCD?>>?;>=88:543;;/--/,..0-1)((),&$%'%
match (1) p(A) = 3.333e-05
match (2) p(G) = 0.9999
p(G) = 0.800
p(C) = 0.066
p(G) = 0.602
p(A) = 0.133

Fig. 6 Most HTS platforms report a quality score for each nucleotide in a read. It denotes the probability *e* that the called nucleotide is incorrect, here the probability it is correct 1-*e* is shown as a sequence logo. Assuming a uniform background distribution for the genome, we can calculate the probability of each of the other three possible nucleotides to be *e*/3. When a read matches several locations in the genome, the mappers typically report the mapping with fewest mismatches as uniquely mapped. In our example, when mapping the read directly to the genome, the mapper would report match 1 as uniquely mapped and disregard match 2. However, when using the probabilities based on the quality scores, instead of considering mismatches we can calculate the probability of each mapping. In this case, we would obtain an approximately 550 times higher probability for match 2 compared to match 1, and we would choose match 2 as the best mapping

quality score cutoff in order to increase the chance of a correct mapping by removing the low-quality end of the read. In `Bowtie2`, the penalty of a mismatch can be set to depend on the quality score—by default the penalty is 6 for a quality of 40, 5 for values from 30 to 39, 4 from 20 to 29, 3 from 10 to 19, and 2 from 0 to 9.

3.2 PSSMs

From the quality score we know the probability of an error e and thus the probability of the called base is $1-e$. If we assume no nucleotide bias for wrong base calls by the sequencing machine, then the probabilities for having each of the other three bases are $e/3$. Given the four base probabilities for each position in the read, we can convert the read into a position-specific scoring matrix (PSSM), which is a construct often used in bioinformatics. If this probability of a base a at position i is called $p_i(a)$, the corresponding PSSM score is $s_i(a) = \log(p_i(a)/q(a))$, where $q(a)$ is the "background" probability, which could be the base frequencies in the genome, but often we just set it uniformly to $\frac{1}{4}$.

If the base is of very high quality, the score of the correct base is high: almost $\log\left(1/\frac{1}{4}\right) = \log(4)$, which is 2 if we are using the logarithm base 2 as is customary. On the other hand, the score for the other bases at a high-quality position will be large negative; e.g., if the error probability is $1/1{,}000$, it would be $\log\left(\frac{1/3{,}000}{1/4}\right)$, which is almost -10. For low-quality positions, the score is small, and if the probabilities of the four bases are close to $\frac{1}{4}$, the scores become close to 0 for all of them. To score a sequence in the genome, we just add up these scores—if we want to score the PSSM for a read against the sequence CTAAG$\cdots$, we would calculate $s_1(\mathrm{C}) + s_2(\mathrm{T}) + s_3(\mathrm{A}) + s_4(\mathrm{A}) + s_5(\mathrm{G}) + \cdots$.

In principle one can model sequence error biases and dinucleotide biases and so forth, but this is probably rarely done. One sophistication should be considered, namely, the possibility of having differences between the sample and the reference genome, such as SNPs, i.e., differences that are not due to sequencing errors.

In its simplest form, we can assume that the probability of a base difference between the sample and the reference genome is p_0. Then the probability of seeing base b in the genome, given base a in the sample, would be $p(b|a) = 1 - p_0$ for $a = b$ and $p(b|a) = p_0/3$ for the rest. Now we can simply replace the above probability $p_i(a)$ with $\tilde{p}_i(b) = \sum_a p(b|a) \times p_i(a)$ in the calculation of the PSSM. For the human genome, the expected frequency of SNPs is around 1 in 1,000, so one could use $p_0 = 0.001$. It is also possible to incorporate more sophisticated evolutionary models than the above and, for instance, use different probabilities for transitions and transversions.

3.3 The Probability of a Match

Given the quality scores, we can calculate the probability that a mapping is correct. Let us assume that we have calculated the score of a read for every possible location in the genome using the PSSM approach described above. Then the probability of a match at a given location l is $P(l) = 2^{S_l}/\sum_k 2^{S_k}$, where S_l is the score for a match at position l and the denominator is a sum over all positions in the genome [3, 18, 19]. If there is contamination in the sample, which is almost always the case, as discussed above, the correct posterior match probability is [19]

$$P(l) = \frac{2^{S_l}}{\sum_k 2^{S_k} + L(1 - P_{\mathrm{M}})/P_{\mathrm{M}}}. \tag{4}$$

Here P_{M} is the prior probability of a match in the genome and L is the number of possible match positions in the genome (~6 billion for the human genome).

Instead of considering whether a match is correct or not, we can now quantify uniqueness. If a long high-quality read matches very well to the genome, and there are no competing high-scoring matches, the above probability $P(l)$ will be very close to one. On the other hand, if the read is short and/or of low quality, so that 2^{S_l} is similar in size or less than $L(1 - P_{\mathrm{M}})/P_{\mathrm{M}}$, the probability will be low. The probability also gets lower if there are competing matches with scores similar to the best match. If there are two matches with identical score, for instance, the probability will be at most 0.5.

The PSSM mapping has been implemented in a version of BWA called BWA-PSSM [19]. The whole sum over all possible sites in the genome would take way too long to calculate, but it is well approximated by a sum over the high-scoring positions. This means that the PSSM search and the calculation of the match probability do not increase the run-time dramatically (depending on parameter settings).

Mappers like BWA and Bowtie2 calculate a mapping quality (MapQ). In BWA, it is derived as an approximation to the above $P(l)$ and log-transformed like the base quality reads [3]. A MapQ value of 37 means a single match with less mismatches than the maximum allowed, 25 means a match with exactly the maximum number of mismatches allowed, whereas matches having competing matches (with more mismatches than the best match) are scored from 23 to 0 as the number of competing matches increases. The MapQ calculated in Bowtie2 is based on the difference between the score of the best match and the score of the second best match divided by the maximum possible score and the minimum accepted score. If there are no competing matches, the difference is replaced by the difference between the score of the read and the smallest score allowed. The maximum MapQ is 42 for a fraction above 0.8 and no competing matches and goes towards zero for matches with lower fractions and/or competing matches.

To see how useful these mapping qualities are to avoid undesired mappings, we simulated Illumina reads with lengths from 20 to 30 nt without insertions/deletions from the *E. coli* genome (Genbank acc. FM180568, 1× coverage) using ART [20] and mapped them to the human genome (hg19) with BWA, Bowtie2, and BWA-PSSM. We used the mapper's default settings and chose a prior match probability $P_{\mathrm{M}} = 0.8$ in BWA-PSSM. Figure 7 shows the fraction of uniquely mapped reads for the three mappers. We see that the number of uniquely mapped reads is lower for BWA-PSSM, which can make use of the prior match probability. By using a probability cutoff with BWA-PSSM, we can influence the amount of matches from very strict to more permissive. When requiring a posterior probability of 0.99 (dotted green line), only ~1 of all reads are mapped uniquely. Bowtie2 maps only around half as many reads uniquely as BWA and for both of them, even so random matches are not modelled directly, having a minimum cutoff on the

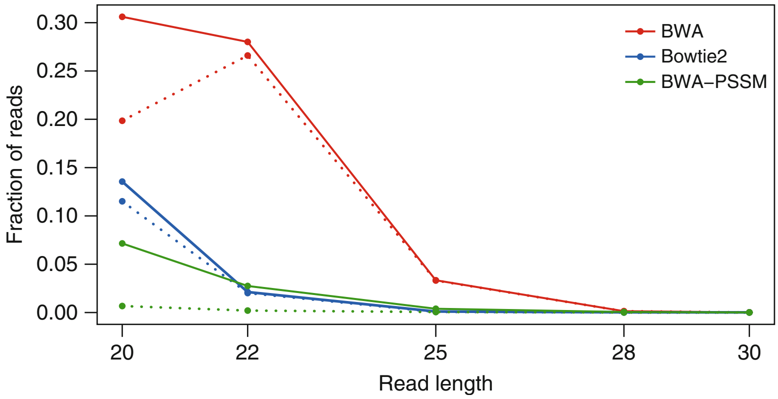


Fig. 7 Mapping of *E. coli* reads to the human genome. The figure shows the fraction of all reads that are mapped uniquely for BWA, Bowtie2, and BWA-PSSM using their default settings. The *full lines* denote all uniquely mapped reads, while the *dotted lines* denote only those uniquely mapped reads that have a MapQ value greater than 20 in BWA and Bowtie2, or equivalently a posterior probability greater than 0.99 in BWA-PSSM. For BWA-PSSM we define uniquely mapped reads as reads with a posterior probability greater than 0.5. *See* also [19]

score helps to reduce undesired mappings in this case (dotted lines). When using a cutoff, BWA still maps nearly twice as many reads than any of the other mappers for read lengths 20 and 22 nt. In practice, when using BWA, we would recommend using the *p*- or *E*-values given in this chapter together with the MapQ score (or avoid short and/or low-quality reads).

Posterior probabilities are also explicitly used in other recent mappers. For example, Stampy [21] uses a Bayesian model to estimate the probability of an incorrect mapping based on errors and variation in the read and repetitiveness of the genome or if the read is contamination. For 454 pyrosequencing, the FLAT mapper [22] also employs a probabilistic framework that considers sequencing errors for short reads. LAST [18] also converts reads into scoring matrices based on the quality values and employs a probabilistic model for the mapping.

The FASTQ quality scores may not be accurate and may vary between runs of the sequencing machine. Therefore it is worth to consider a recalibration of the quality scores; *see*, e.g., DePristo et al. [23].

4 Conclusion

The success of many experiments is dependent on high-quality mapping of short reads. From our introductory example with six read data sets from different types of experiments, we saw that different mapping programs yield quite differing mappings depending on the type of data. Even this simple experiment shows that, although mapping is a routine task nowadays, it is important to think about the choice of the mapper and the mapping parameters, which are crucial for a high-quality mapping.

For short reads, the most severe problem in complex genomes, like human and other eukaryotic genomes, is wrongly mapped reads. In our statistical analysis (Fig. 2), we show that when considering unique matches in the human genome, the highest rate of wrongly mapped reads is observed at a read length around 18 nt. When the reads have a high error rate, the problem persists with more than 1 % wrongly mapped reads up to a length around 50 nt. Because of the repetitive nature of genomes, the risk of wrongly mapped reads will always be there. It may actually be more severe than it seems, because repeat regions usually also involve assembly problems. If a region is missing in the reference genome or several regions are mixed in the assembly, the chance of wrong mappings increases.

Another problem is the unintended mapping of reads to a genome, which do not originate from that genome, such as contamination. From our theoretical considerations, we can see that this problem is most severe for very short reads. For example, when mapping random sequences with length below 20 nt to the human genome, more than 1 % of them will map perfectly, which cannot be influenced by any mapping parameters. We confirmed this estimate by mapping randomly sampled reads from *E. coli* to the human genome. For more permissive criteria on the mapping, i.e., allowing for mismatches, this error rate shifts to around a 25–30 nt read length, and for longer reads it should be below 1/1,000, and decreasing exponentially with length.

Both of the above problems can in principle be dealt with by calculating a probability that the read is correctly mapped as shown in the last section. Several mappers do this in one way or the other, and assign either a probability or a mapping quality score, which can represent a log-transformed mapping (or error) probability. The calculation of these quantities varies a great deal between mappers and is often based on approximations in order to maintain the speed of mapping. However, we saw that using a cutoff on this posterior probability can limit the number of matches of *E. coli* reads to the human genome (Fig. 7).

Another way to limit the problem of wrongly mapped reads, which is used in many applications, is to require multiple matches to the same region. For instance, one would rarely accept a binding site unless it is covered by many reads in a ChIP-Seq experiment, and one would not call a polymorphic site based on a single read in a re-sequencing experiment. This is an excellent strategy if one is careful with removing PCR artifacts (e.g., reads that map with exactly the same start and end position). However, one should keep in mind that (1) a few reads might happen to map incorrectly to the wrong repeat, and (2) the higher the quality of mapping, the smaller the number of reads required.

In this chapter we have not considered the problem of sensitivity: reads that should have been mapped but are missed by the mapper. This problem may cause higher expenses, because a higher sequencing depth is needed if too few reads are mapped, and does therefore receive more attention. In our opinion it is less of a problem than wrongly mapped reads.

Exactly how to map reads depends heavily on the application. We hope that this chapter has sharpened your intuition on the subject and given you some pointers on how to attack the problems.

5 Notes

5.1 Probability of Correct Exact Matches

In the text we consider a simple example, where a read matches exactly and uniquely, but there are n other matches with one mismatch. We want to calculate the probability that an exact match is also the correct match $P(\text{correct}|\text{exact})$. Using Bayes' theorem, it can be written as

$$P(\text{correct}|\text{exact}) = \frac{P(\text{exact}|\text{correct}) \times P(\text{correct})}{P(\text{exact})}.$$

The first two terms are easy (see below), but the denominator needs to be rewritten as

$$P(\text{exact}) = P(\text{exact}|\text{correct}) \times P(\text{correct}) + P(\text{exact}|\text{incorrect}) \\ \times P(\text{incorrect}).$$

The individual terms are

$$P(\text{exact}|\text{correct}) = (1 - p)^l,$$

where l is the read length and p is the error rate for each base, i.e., the probability that a base is incorrect. Similarly, for precisely one mismatch:

$$P(\text{exact}|\text{incorrect}) = \frac{p}{3}(1 - p)^{l-1}.$$

The prior probability that a match is correct (before even comparing the sequences) is just $P(\text{correct}) = 1/(n + 1)$ and $P(\text{incorrect}) = 1 - P(\text{correct})$. Now inserting all the terms. we get after rearrangements

$$P(\text{correct}|\text{exact}) = \frac{1}{1 + \frac{np}{3(1-p)}} \approx \frac{1}{1 + np/3},$$

where the last approximation holds only if p is small.

This calculation ignores the possibility that there are matches in the genome with more than one mismatch. These will further lower the probability that the exact match is the correct one, so the above is an upper bound for this probability.

5.2 Approximation of p-Value for Random Matches

The probability of not having a match anywhere in the genome is $\left(1 - F\left(m; l, \frac{1}{4}\right)\right)^L = \left(1 - E(m, l, L)/L\right)^L$. Assuming that $F\left(m; l, \frac{1}{4}\right)$ is small, this is very well approximated by $1 - e^{-E(m,l,L)}$.

5.3 Read Preprocessing and Mapping for Fig. 1

Illumina reads of length 30 nt (without indels) were simulated with ART [20] from the human reference genome hg19 with 0.01× coverage, which results in 953147 reads. Experimental Data sets were downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) and SRA (http://www.ncbi.nlm.nih.gov/sra, [24]) databases. The accession numbers of the different data sets used are listed below in Table 1. To identify barcodes and adapters we analyzed the data sets with FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to find overrepresented sequences in the reads and identify adapters used in Illumina sequencing protocols. We used AdapterRemoval [25] to remove barcodes, adapters, and trimming low-quality bases and trailing Ns. The primer sequences and barcodes are listed in Table 1. After preprocessing, all reads less than 20 nt long were discarded.

500,000 reads were randomly sampled from each of the above data sets. The indices for Bowtie2 and BWA were created from the human reference genome hg19 and reads were mapped to the

Table 1
Accession numbers and barcode/adapter sequences

Data set	Accession numbers	Barcode	Adapter
Ancient DNA	SRX013912	–	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT
CAGE	GSM849338	–	AGACAGCAGG (5′-adapter)
Small RNA-Seq	GSM893567 GSM893569 GSM893584 GSM893585 GSM893571	ACTT GTTA GGGT GTTA TCGC	CTGTAGGCACCATCAAT
CLIP-Seq	GSM859978 GSM859979 GSM859980 GSM859981	–	TCGTATGCCGTCTTCTGCTTG
ChIP-Seq	GSM727557	–	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG

genome with sensitive settings for both mappers. The commands used for preprocessing and mapping are as follows:

Read preprocessing:

```
AdapterRemoval --trimns -trimqualities --5prime
[5'-adapter/barcode sequence] --pcr1 [3'adapter
sequence] < file.fastq
```

Mapping with `Bowtie2`
Build index:

```
bowtie2 -build ref_genome.fa bowtie_index
```

Read mapping:

```
bowtie2 -i L,4,0 -L 18 -M 2000000 -N 1 -x bowtie_index -U
infile.fastq -S outfile.sam
```

Options:

- **`M n`** number of distinct alignments to consider for a read is $n + 1$.
- **`i f,b,a`** interval length as function of read length x, $L(x) = b + a \times f(x)$. It defines the number of seeds to extract from the read and thereby the number of allowed mismatches across the reads.
- **`L n`** n is the seed length.
- **`N`** number of allowed mismatches in each seed.

Mapping with `BWA` **(0.6.1-r104)**
Build index:

```
bwa index -a bwtsw ref_refgenome.fa
```

Read mapping:

```
bwa aln -n 0.01 -l 1024 -m 2000000 -t 28 ref_genome.fa
data.fastq > data.sai
bwa samse -f data.sam ref_genome.fa data.sai data.
fastq
```

Options:

- **`n p`** ratio of the reads that would not be mapped at an error rate of 2 % (for more details *see* [26]).
- **`l n`** seed length. If the length of a read is shorter than `n` then seeding is disabled for the given read. By setting the value of n to 1,024, seeding is disabled for all reads in the datasets.
- **`m n`** number of alignments to consider when looking for the optimal alignment.

Acknowledgments

This work was supported by grants from the Danish Strategic Research Council (COAT), the Novo Nordisk Foundation, and European Union (Project #265933 *Hotzyme*).

References

1. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
2. Li L, McCorkle S, Monchy S, Taghavi S, van der Lelie D (2009) Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. Biotechnol Biofuels 2:10. doi:10.1186/1754-6834-2-10
3. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858. doi:10.1101/gr.078212.108
4. Li R, Yu C, Li Y, Lam T, Yiu S, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967. doi:10.1093/bioinformatics/btp336
5. Langmead B, Salzberg S (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923
6. Ruffalo M, LaFramboise T, Koyutürk M (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 27:2790–2796. doi:10.1093/bioinformatics/btr477
7. Stiller M, Green R, Ronan M, Simons J, Du L, He W, Egholm M, Rothberg J, Keates S, Keats S, Ovodov N, Antipina E, Baryshnikov G, Kuzmin Y, Vasilevski A, Wuenschell G, Termini J, Hofreiter M, Jaenicke-Després V, Pääbo S (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. Proc Natl Acad Sci U S A 103(13):578–584. doi:10.1073/pnas. 0605327103
8. Kircher M (2012) Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol 840:197–228. doi:10.1007/978-1-61779-516-9\textunderscore23
9. Rasmussen M, Li Y, Lindgreen S, Pedersen J, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert M, Wang Y, Raghavan M, Campos P, Kamp H, Wilson A, Gledhill A, Tridico S, Bunce M, Lorenzen E, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre T, Grønnow B, Meldgaard M, Andreasen C, Fedorova S, Osipova L, Higham T, Ramsey C, Hansen T, Nielsen F, Crawford M, Brunak S, Sicheritz-Pontén T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463:757–762. doi:10.1038/nature08835
10. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100 (15):776–781. doi:10.1073/pnas.2136655100
11. Morin R, O'Connor M, Griffith M, Kuchenbauer F, Delaney A, Prabhu A, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves C, Marra M (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18:610–621. doi:10.1101/gr.7179508
12. Zhang C, Darnell R (2011) Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotechnol 29:607–614. doi:10.1038/nbt.1873
13. Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov J, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin J, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston R, Wilson R, Hillier L, McPherson J, Marra M, Mardis E, Fulton L, Chinwalla A, Pepin K, Gish W, Chissoe S, Wendl M, Delehaunty K, Miner T, Delehaunty A, Kramer J, Cook L, Fulton R, Johnson D, Minx P, Clifton S, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs R, Muzny D, Scherer S, Bouck J, Sodergren E, Worley K, Rives C, Gorrell J, Metzker M, Naylor S, Kucherlapati R, Nelson D, Weinstock G, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith D, Doucette-Stamm L, Rubenfield M,

Weinstock K, Lee H, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis R, Federspiel N, Abola A, Proctor M, Myers R, Schmutz J, Dickson M, Grimwood J, Cox D, Olson M, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans G, Athanasiou M, Schultz R, Roe B, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie W, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey J, Bateman A, Batzoglou S, Birney E, Bork P, Brown D, Burge C, Cerutti L, Chen H, Church D, Clamp M, Copley R, Doerks T, Eddy S, Eichler E, Furey T, Galagan J, Gilbert J, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson L, Jones T, Kasif S, Kaspryzk A, Kennedy S, Kent W, Kitts P, Koonin E, Korf I, Kulp D, Lancet D, Lowe T, McLysaght A, Mikkelsen T, Moran J, Mulder N, Pollara V, Ponting C, Schuler G, Schultz J, Slater G, Smit A, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf Y, Wolfe K, Yang S, Yeh R, Collins F, Guyer M, Peterson J, Felsenfeld A, Wetterstrand K, Patrinos A, Morgan M, de Jong P, Catanese J, Osoegawa K, Shizuya H, Choi S, Chen Y, Szustakowki J, International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. doi:10.1038/35057062

14. Longo M, O'Neill M, O'Neill R (2011) Abundant human DNA contamination identified in non-primate genome databases. PLoS One 6:e16,410. doi:10.1371/journal.pone.0016410

15. Cock P, Fields C, Goto N, Heuer M, Rice P (2010) The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771. doi:10.1093/nar/gkp1137

16. Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, Berka J, Braverman M, Chen Y, Chen Z, Dewell S, Du L, Fierro J, Gomes X, Godwin B, He W, Helgesen S, Ho C, Ho C, Irzyk G, Jando S, Alenquer M, Jarvie T, Jirage K, Kim J, Knight J, Lanza J, Leamon J, Lefkowitz S, Lei M, Li J, Lohman K, Lu H, Makhijani V, McDade K, McKenna M, Myers E, Nickerson E, Nobile J, Plant R, Puc B, Ronan M, Roth G, Sarkis G, Simons J, Simpson J, Srinivasan M, Tartaro K, Tomasz A, Vogt K, Volkmer G, Wang S, Wang Y, Weiner M, Yu P, Begley R, Rothberg J (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380. doi:10.1038/nature03959

17. Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin J (2011) Accuracy and quality assessment of 454 GS-FLX titanium pyrosequencing. BMC Genomics 12:245. doi:10.1186/1471-2164-12-245

18. Hamada M, Wijaya E, Frith M, Asai K (2011) Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. Bioinformatics 27:3085–3092. doi:10.1093/bioinformatics/btr537

19. Kerpedjiev P, Lindgreen S, Frellsen J, Krogh A (2013) Adaptable probabilistic mapping of short reads using position specific scoring matrices. Unpublished

20. Huang W, Li L, Myers J, Marth G (2012) ART: a next-generation sequencing read simulator. Bioinformatics 28:593–594. doi:10.1093/bioinformatics/btr708

21. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. Genome Res 21:936–939. doi:10.1101/gr.111120.110

22. Vacic V, Jin H, Zhu J, Lonardi S (2008) A probabilistic method for small RNA flowgram matching. Pac Symp Biocomput 75–86

23. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas M, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. doi:10.1038/ng.806

24. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration (2012) The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 40:D54–D56. doi:10.1093/nar/gkr854

25. Lindgreen S (2012) AdapterRemoval: easy cleaning of next generation sequencing reads. BMC Res Notes 5:337. doi:10.1186/1756-0500-5-337

26. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595. doi:10.1093/bioinformatics/btp698