Published online 29 February 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 11* ***e83***
doi:10.1093/nar/gks196

FANSe: an accurate algorithm for quantitative mapping of large scale sequencing reads

Gong Zhang[1,2,*], Ivan Fedyunin[1,3], Sebastian Kirchner[1], Chuanle Xiao[2], Angelo Valleriani[3] and Zoya Ignatova[1,*]

[1]Biochemistry, Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14467 Potsdam, Germany, [2]Institute of Life and Health Engineering, Jinan University, Huang-Pu Avenue West 601, 510632 Guangzhou, China and [3]Theory and Bio-Systems, Max Planck Institute of Colloids and Interfaces, 14424 Potsdam, Germany

Received January 11, 2012; Revised February 5, 2012; Accepted February 12, 2012

ABSTRACT

The most crucial step in data processing from high-throughput sequencing applications is the accurate and sensitive alignment of the sequencing reads to reference genomes or transcriptomes. The accurate detection of insertions and deletions (indels) and errors introduced by the sequencing platform or by misreading of modified nucleotides is essential for the quantitative processing of the RNA-based sequencing (RNA-Seq) datasets and for the identification of genetic variations and modification patterns. We developed a new, fast and accurate algorithm for nucleic acid sequence analysis, FANSe, with adjustable mismatch allowance settings and ability to handle indels to accurately and quantitatively map millions of reads to small or large reference genomes. It is a seed-based algorithm which uses the whole read information for mapping and high sensitivity and low ambiguity are achieved by using short and non-overlapping reads. Furthermore, FANSe uses hotspot score to prioritize the processing of highly possible matches and implements modified Smith–Watermann refinement with reduced scoring matrix to accelerate the calculation without compromising its sensitivity. The FANSe algorithm stably processes datasets from various sequencing platforms, masked or unmasked and small or large genomes. It shows a remarkable coverage of low-abundance mRNAs which is important for quantitative processing of RNA-Seq datasets.

INTRODUCTION

Rapid technological advances in massively parallel, high-throughput sequencing technologies (aka deep sequencing) can deliver datasets of gigabases (1), which are expressed in millions of 'reads' (i.e. short nucleotide sequences, usually 17–400 nt long, depending on the platform, the protocol and the sample). In the studies of *de novo* assemblies of whole genomes, millions of reads are assembled together to build up an unknown genome (2,3). In the analyses of genomic variations or epigenomic, transcriptomic and translatomic studies, millions of reads originating from DNA or RNA fragments are mapped to an already known reference genome (4–7). Read mapping approaches have to adequately respond to the specifications and errors of each type of technology. In RNA-based sequencing (RNA-seq) errors can be introduced by stochastic misincorporation or misreading of modified nucleotides by the reverse transcriptase (8), or by default sequencing errors by the sequencing platforms (9,10). Accurate dynamic programing algorithms, including Smith–Waterman algorithm (11) or conventional heuristic algorithms [FASTA (12) and BLAST (13)] are suitable for the detection of misincorporations since mismatches can be implemented within them. However, high accuracy of aligning vast amount of reads to a genome compromises the performance and speed of the algorithms (14).

In general, the algorithms that are currently used to map deep-sequencing datasets can be classified into two major groups: seed-based (or 'hash table-based') and Burrows–Wheeler Transform (BWT)-based algorithms (14,15). Conceptually, the seed-based algorithms, including BLAST, BLAT (16), SOAP (17), Genomemapper (18), MAQ (19), Stampy (20) and SHRiMP (21) extract short

*To whom correspondence should be addressed. Tel: +49 331 977 5130; Fax: +49 331 977 5128; Email: ignatova@uni-potsdam.de
Correspondence may also be addressed to Gong Zhang. Tel: +86 20 85224031; Fax: +86 20 85227039; Email: zhanggong@jnu.edu.cn

subsequences called ‘seeds’ from the query sequence and search for exact matches (in at least one of the seeds) to the reference genome sequence (15). For each exact match, the algorithms refine the alignment with more sensitive methods (e.g. Smith–Waterman algorithm) and thereafter, the best alignment is selected. BWT-based algorithms, e.g. Bowtie (22), BWA (23) and SOAP2 (24), compact the reference genome into a data structure and search ‘suffixes’ of a read through the index to find a match (15). BWT-approaches are faster than seed-based methods when the exact reference genome is available; however, if only transcriptomes of distant species are available, the seed-based algorithms show a much greater sensitivity (25).

The DNA modifications, single-nucleotide polymorphisms (SNPs) or misincorporations at modified ribonucleotides introduced by the reverse transcriptase can alter the query sequence: RNA-seq datasets bear higher error rates than DNA sequencing. Furthermore, the sequencing platforms also contribute some alterations to the read sequences, e.g. GS FLX sequencing is rather more likely to include insertions and deletions (indels) while the reads from Illumina-sequencing machines contain mismatches (26,27). The sensitivity of the BWT-based algorithms decreases exponentially with the number of mismatches. Comparative analysis revealed that Bowtie and BWA only map half of the reads compared with seed-based algorithms (28); thereby, reads with moderate- to low-abundance are markedly affected which will bias the quantitative processing of the RNA-seq data. Considering an error rate per nucleotide of ~1.5% in RNA-seq applications (Illumina platform) (10), a conventional 11-nt BLAST seed has >15% probability of containing at least one mismatch. Reducing the length of the seeds and/or applying larger numbers of seeds from the query read sequence increases the sensitivity, but it dramatically reduces the speed, specifically for large genomes. For example, BLAT needs 78 days to map 3.5 million reads to the human genome (28). Three or more mismatches are likely to occur in one read, particularly when using SOLiD and Helicos sequencing platforms (2–7% average error rate) (9,10). By allowing more mismatches, the accuracy is compromised while still maintaining a high speed. For example, BFAST has a sensitivity of only 80% when allowing five mismatches in 50-nt long reads (28).

Most of the currently available mapping algorithms offer a limited ability, if any at all, to map reads with indels (14), even though some deep sequencing platforms deliver a relatively high indel rate (indels account for more than two-thirds of the errors of GS FLX 454 pyrosequencing) (27). Furthermore, MAQ (19), SOAP (17) and Bowtie (22) handle mismatches but cannot detect indels, while PatMaN (29), SHRiMP (21) and BWA (23) can detect limited number of indels. Even a very low indel frequency (0.5/kb) can cause a mismapping rate of 4–13% (30). The sensitive read mapping and accurate detection of mismatches and indels to reference genomes is crucial in studies aimed at identifying genetic variations (e.g. SNPs) (5,31) or DNA methylation patterns (32,33), or studies quantitatively analyzing RNA-seq data (6,34–37). Thus, there is a demand for a versatile algorithm to quantitatively map sequencing datasets with various lengths of reads with a higher error rate. We developed a new mapping algorithm, FANSe, which accurately and quickly maps millions of reads with a scalable read length to reference genomes in various sequencing applications. We validated the performance of FANSe with short (24 nt) and long reads (>140 nt). Long reads were generated using a prokaryotic *Escherichia coli* DNA library sequenced with a 454 GS FLX pyrosequencing platform and short reads were obtained with an Illumina RNA-seq of randomly-digested prokaryotic *E. coli* mRNA and eukaryotic HeLa mRNA. We also verified FANse with an *in silico* simulated random sequencing reads of different lengths (24 nt and 50 nt).

MATERIALS AND METHODS

Design of FANSe

FANSe is a seed-based algorithm with a simple design to ensure accuracy, which comprises the following steps:

Step 1: A read is split into several non-overlapping seeds. Each seed is *n*-base long with a typical seed size of 6–8 nt (Figure 1A). For reads that are not completely covered by the non-overlapping 6- or 8-nt seeds, an extra seed is taken at the end of the read that overlaps with the penultimate one (Figure 1B).

Step 2: All seeds are aligned to the reference genome sequence. Seeds with no mismatch and no indel ensure the correct mapping of the whole read (Figure 1C). A read can only be missed when all seeds contain at least one error (mismatch or indel).

Step 3: Adjacent seeds are combined if they are likely to be within one read, based on their offsets, and independent potential locations (‘hotspots’) are defined. The number of combined matched seeds defines the score of the hotspot. Hotspots with high scores are refined with priority (Figure 1C).

Step 4: The alignment for each hotspot is refined in the order of decreasing hotspot scores and the best alignment, i.e. the hotspot that contains the least number of mismatches, is chosen. Two methods can be used here: (i) a simple alignment that is based on a nucleotide-by-nucleotide comparison that does not detect indel in order to achieve a faster speed or (ii) accelerated Smith–Waterman alignment, providing 100% sensitivity for indels (Figure 1D).

Acceleration of the FANSe

The relatively small seeds (6- or 8-nt) may be matched several times, even within small reference genomes (e.g. bacteria, yeasts). Thus, Step 2 is the most time-consuming step. Acceleration is achieved by using a seed lookup table comprising either $4^6 = 4096$ different 6-nt seeds or $4^8 = 65536$ various 8-nt seeds. Prior to mapping, a search of all possible seeds is performed through the reference genome and the locations of exact matches are recorded in the seed lookup table. When designating a seed from a real read, the locations of the exact matches are obtained

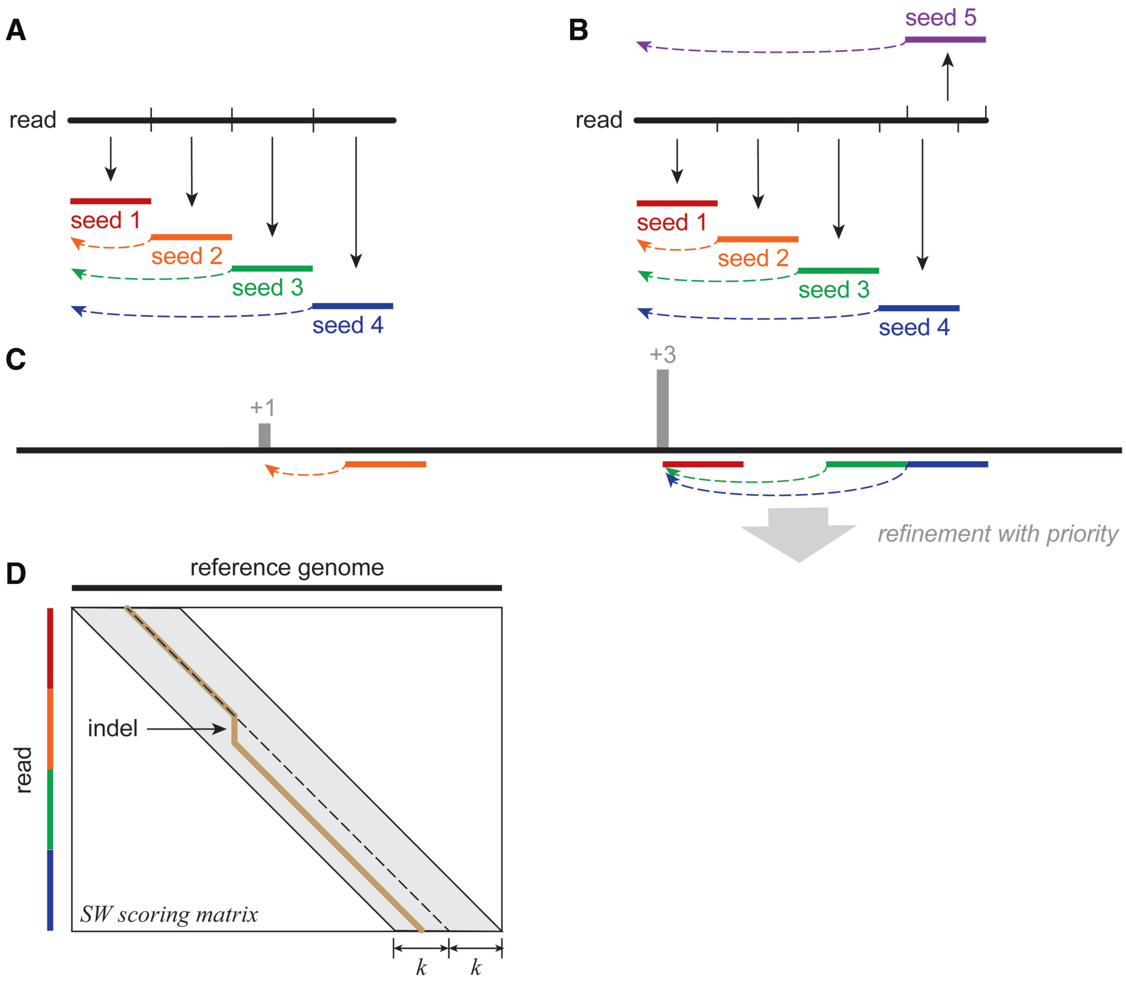

Figure 1. Principle of the FANSe algorithm. Scheme of a read covered by non-overlapping seeds (**A**) or an additional overlapping seed (**B**). The dashed lines mark the offset, i.e. the distance between the seed start position and the read start position. (**C**) Alignment of seeds to a reference genome (black line). Hotspots are represented as gray bars and the number represents the hotspot score. (**D**) Accelerated Smith–Waterman scoring to detect indels. Only the scoring area near the diagonal (gray shadow) is calculated. The dashed line represents the backtracking path without indel; the brown line depicts the backtracking path with indels; k is the number of allowed errors.

directly from the table that can be accomplished very fast (within microseconds). The number of entries in the seed lookup table equals the length of the genome, and is not expensive regarding the consumption of memory.

The hotspots with the highest scores will be processed with priority since they may contain a lower number of errors and are thus more likely to be mapped (Figure 1C). If a high-scoring hotspot is successfully mapped, other hotspots are not considered, which in turn saves comparison operations. If no high-scoring hotspots are mapped, lower-scoring hotspots will then be processed. For a read with x non-overlapping seeds, a hotspot with a score s contains minimum $x\!-\!s$ errors. Allowing k mismatches, this read needs a minimum hotspot score of $x\!-\!k$ to be mapped successfully. All of the hotspots with $x\!-\!s\!>\!k$ will be neglected which minimizes the number of hotspots examined.

Indel detection is a challenge for the mapping algorithms since the most accurate Smith–Waterman method is computationally very costly. When the indel detection of FANSe is on, it first maps a read in a first trial considering no indel. An indel-containing read, however, contains a large number of mismatches in a simple alignment check. Only reads that failed to be mapped in the first round will be further analyzed using the Smith–Waterman refinement. An accelerated refinement of the Smith–Waterman algorithm for complete indel sensitivity is implicated in FANSe. The most time-consuming step in this algorithm is the calculation of the scoring matrix with a time complexity of $O(n^2)$. If maximum k errors, including mismatches and indels are allowed, the final back-tracking route (11) would maximally deviate by k cells located away from the main diagonal in the scoring matrix (Figure 1D). Therefore, it is only necessary to calculate the cells near the main diagonal, reducing the time complexity to $O(n)$. The majority of the hotspots will contain too many errors and in those cases the Smith–Waterman scores in the first few rows will be very low. If this is detected, the Smith–Waterman refinement is aborted since the errors of the alignment at this hotspot would exceed the allowed limit of k. This acceleration of the Smith–Waterman refinement does not affect its accuracy; it markedly reduces the running time by ~90%, thus providing high sensitivity for indels at a minimal computational cost.

RNA sequencing experiments to validate FANSe

One RNA-sequencing dataset was generated by the high-throughput sequencing of randomly fragmented *E. coli* mRNA using the following protocol: *E. coli* MC4100 cells were grown in LB medium at 37°C until the mid-log phase (OD600 ~ 0.5), then rapidly cooled down by pouring through crushed ice and harvested by centrifugation for 5 min at 5000 *g* at 4°C. The cell pellet was dissolved in resuspension buffer (0.016 M Tris–HCl, pH = 8.1, containing 0.05 M KCl and 0.2% EDTA) and treated with 1 mg/ml lysozyme for 5 min on ice, followed by a total RNA extraction with TRIzol (Invitrogen). The mRNA fraction was enriched via the subtraction of small RNAs (5 S and tRNAs) with the GeneJET RNA Purification Kit (Fermentas) and depletion of 16 S and 23 S rRNA using the MICROBExpress Bacterial mRNA Enrichment Kit (Ambion) (38). The enriched mRNAs were heated up to 95°C for 40 min in alkaline fragmentation buffer (100 mM $NaCO_3$, pH 9.2, containing 1 mM EDTA), which cleaves the mRNA into short fragments in a random and unbiased manner (35). Chemically digested fragments were resolved on 15% denaturating polyacrylamide gel and fragments between 20 - and 35-nt long were eluted from the gel with 300 mM sodium acetate buffer, pH 5.5. The complementary DNA (cDNA) library was prepared via direct adapter ligation according to the method described already (39), followed by reverse transcription with RevertAid^{TM} H Minus Reverse Transcriptase (Fermentas) and PCR-based amplification with *Pfu* DNA Polymerase (Fermentas). The sequencing was performed on the Illumina GAIIx platform. After the sequencing, the adapter sequences were removed and the high-quality reads with Phred score >20 were further processed using different algorithms. The reads which were aligned with rRNA sequences with no mismatch or one mismatch were also removed. The remaining reads, which were enriched in mRNA reads, were used as the input of the mapping algorithm; in total 9 387 287 reads with a length ranging from 18 to 36 nt were used.

HeLa cells (ATCC CCL-2) were cultivated to 80% confluence in DMEM (PAN Biotech) supplemented with 10% FCS (PAN Biotech GmbH) and 2 mM L-glutamine (Gibco), at 37°C and 5% CO_2. Cells were washed in 1× DPBS (Gibco), harvested by trypsinization and total RNA was isolated using TRIzol (Invitrogen), according to manufacturer's instructions. The poly(A)$^+$ mRNA was isolated from total RNA using the Dynabeads mRNA Purification Kit (Invitrogen). Alkaline fragmentation of mRNA, size selection of the fragments, preparation of the sequencing library and the sequencing reaction was performed as described above. The set contained 19 347 370 reads with a length ranging from 16 to 34 nt. The first one-tenth of the reads (1 934 737 reads) was used as the input of the mapping algorithm. The reads were mapped to the human chromosome 21 reference sequence (hg19/GRCh37, downloaded from UCSC Genome browser, http://hgdownload.cse.ucsc.edu).

Furthermore, we downloaded one dataset of *E. coli* genomic DNA sequenced with the 454 GS FLX pyrosequencing platform as a typical dataset for long reads (Human Microbiome Project, data accession number SRR057661, downloaded from DDBJ (DNA Data Bank of Japan) Sequence Read Archive (DRA, https://trace.ddbj.nig.ac.jp/dra/index.shtml) (40). In total 168 890 high-quality reads (with a Phred score >5) longer than 140 nt were selected and used to feed the algorithms.

Comparison of various mapping programs

We compared eight widely used, non-commercial read mapping algorithms, including BLAT (16), SOAP (17), Genomemapper (18), mrsFAST (41), SHRiMP (21), SOAP2 (24), Bowtie (22) and BWA (23) (Supplementary Table S1). The BLAT, SOAP, Genomemapper, mrsFAST and SHRiMP programs are seed-based algorithms, whereas SOAP2, Bowtie and BWA belong to the BWT-based algorithms. The performance test was performed on a quad-core Intel i5-2300 machine with 8GB RAM. Windows 7 64-bit and Ubuntu 10.10 64-bit (Linux) were installed to run the programs accordingly.

Two terms are defined here that were used to evaluate the accuracy of read mapping algorithms: sensitivity and correctness [adapted from Reference (42)]. For the algorithms that report mapping quality values we considered a read with a minimum mapping quality of 20 in the Phred score scale (i.e. <1% possibility of false-positive mapping) as a 'mapped read'. A read that is processed by an algorithm can result in one of the following three categories: (i) correctly mapped (*C*), if the read is mapped to the genome at the correct place; (ii) incorrectly mapped (*I*), if the read is mapped to the genome but at an incorrect place or (iii) unmapped (*U*), when a read fails to be mapped to the genome and is then discarded. Sensitivity is defined as a fraction of the total mapped reads out of all reads, $\frac{C+I}{C+I+U}$, and the correctness means a fraction of the correctly mapped reads from the total mapped reads, $\frac{C}{C+I}$. Only the sensitivity can be calculated from a deep-sequencing dataset, which is proportional to the number of mapped reads. Correctness can be evaluated using simulated random datasets. Random datasets were simulated from the *E. coli* genome and human chromosome 21 masked reference sequences. Each dataset contained 500 000 reads with an identical read length (24 nt or 50 nt). We also simulated a series of indel-free datasets with a substitution rate ranging from 0.5% to 8% and a series of indel-containing datasets fixed at a substitution rate of 1% and variable indel rate from 0.5% to 4%, in which the indel length (the number of consecutive nucleotide insertions or deletions) was set as 1.

The FANSe algorithm is accessible at http://bioinformatics.jnu.edu.cn/software/fanse/. The web site contains a detailed tutorial and the source code for download.

RESULTS

Concept of FANSe

The large amount of reads generated by high-throughput sequencing has triggered the development of many mapping algorithms towards greater speed but with compromises regarding completeness (15). Quantitative processing of the sequencing data, which may contain mismatches and indels, rather sets the demands for a higher accuracy. Here, we aimed to develop an algorithm that will accurately and quantitatively map sequencing reads while still maintaining a reasonable speed. FANSe uses the core of a seed-based algorithm, but unlike most seed-based mapping algorithms that usually use large seeds (10–14 nt), the typical seed size here is 6–8 nt. In addition, it uses the entire information from a sequencing read which added to the small seed size increases the sensitivity. Importantly, the reads are also designed in a non-overlapping manner which minimizes their number and achieves their independency (Figure 1A and B). A read can be mapped if at least one of the seeds aligns without a mismatch. Offsets are further used to combine the seeds within one read (Figure 1A and B); this operation reduces the number of hotspots, which are the putative alignment locations in the genome. A 24-nt-long read usually generates 1000–4000 hotspots in the *E. coli* reference genome when 6-nt seeds are used and one order of magnitude fewer hotspots when 8-nt seeds are used. The hotspot scoring approach prioritizes the processing of hotspots with the highest number of exact matches, thus reducing any further efforts to find the best hotspot (Figure 1C). Scoring of the hotspots is a novel feature of FANSe which decreases significantly the number of hotspots to be refined and consequently accelerates the mapping. The alignments need to be further refined, which computationally is an inexpensive operation (Figure 1D).

When detecting indels, FANSe implements a reduced Smith–Waterman refinement that significantly accelerates the calculation and unlike other Smith–Waterman algorithms is hardware unspecific. Furthermore, instead of using '2-bits-per-base encoding' to process masked genomes and/or genomes with undefined nucleotides, FANSe implies 8-bits-per-base which is not restricted to only four characters (A T G C) and can also identify masked or undefined nucleotides (N).

Sensitivity, correctness and scalability of FANSe

A read split into x non-overlapping seeds can be reliably mapped to a genome when $\leq x-1$ mismatches are allowed, so that at least one seed contains no errors. Alternatively, reads that contain f mismatches and $(f+1)$ seeds can always be successfully mapped to a genome. This corresponds to a minimal length of $n(f+1)$, where n is the seed length (Table 1). Commonly used sequencing platforms typically provide read length of 18–24 nt (microRNA) or longer 36–125 nt (mRNA, DNA, etc.) that can be fully mapped, typically allowing two to three mismatches. When allowing more mismatches the performance of some algorithms decreases, e.g. the sensitivity of BFAST drops to 80% for 50-nt reads allowing five mismatches (28). The simple design of FANSe allows a theoretical estimation of the mapping error rate when more errors are allowed (see Supplementary Data). Within the range of error rate in the current next-generation sequencing platforms, the rate of losing a mappable read is very low: 10^{-3}–10^{-5} (Supplementary Figure S1).

Table 1. Minimal read length and errors allowed per read to achieve complete mapping with FANSe

Error(s) allowed	Minimal read length for complete mapping	
	6-nt seeds	8-nt seeds
1	12	16
2	18	24
3	24	32
4	30	40
5	36	48
10	66	88

Next, we compared the ability of FANSe to map short reads generated from RNA-seq with other algorithms. We extracted the total mRNA from exponentially growing *E. coli* or eukaryotic HeLa cells, randomly fragmented them into short fragments and sequenced them on the Illumina GAIIx platform. Compared with all of the tested programs, FANSe showed the highest sensitivity in read mapping with disabled indel detection (Figure 2A). When indels were considered, FANSe also achieved the best sensitivity among the algorithms that are capable of handling indels (e.g. BLAT, BWA, mrsFAST and SHRiMP) (Figure 2B). Even though the Illumina GAIIx platform operates at a very low indel rate [estimated to be 0.0032% per nucleotide (43)], the indel search with the Smith–Waterman refinement in FANSe was enabled that increased the mapped reads by 6.5%. With a minimum read length of 18 nt in this dataset, FANSe achieved a complete mapping of all reads when 6-nt seeds were used and one or two mismatches were allowed (Table 1). Note that SOAP2 did not work for this dataset because of an internal error, and Genomemapper only mapped a very small fraction of the reads. When using 8-nt seeds, only 0.27% fewer reads were mapped with FANSe than when 6-nt seeds were used; however the mapping speed was accelerated by >12-fold.

We next compared the read hits for each gene mapped by FANSe, BWA and BLAT (Figure 2C). Similar to FANSe, BWA showed a high ability to map the reads of highly-abundant mRNAs, whereas BLAT mapped significantly fewer reads, thereby proportionally losing also reads of high-abundance mRNAs. Both BWA and BLAT algorithms, however missed reads of low-abundance mRNAs (Figure 2C): when the read hits of a gene dropped below 200 (BWA) or 1000 hits (BLAT), these algorithms disproportionally lost mappable reads that could create a bias if the RNA-seq set is used for further quantitative analysis.

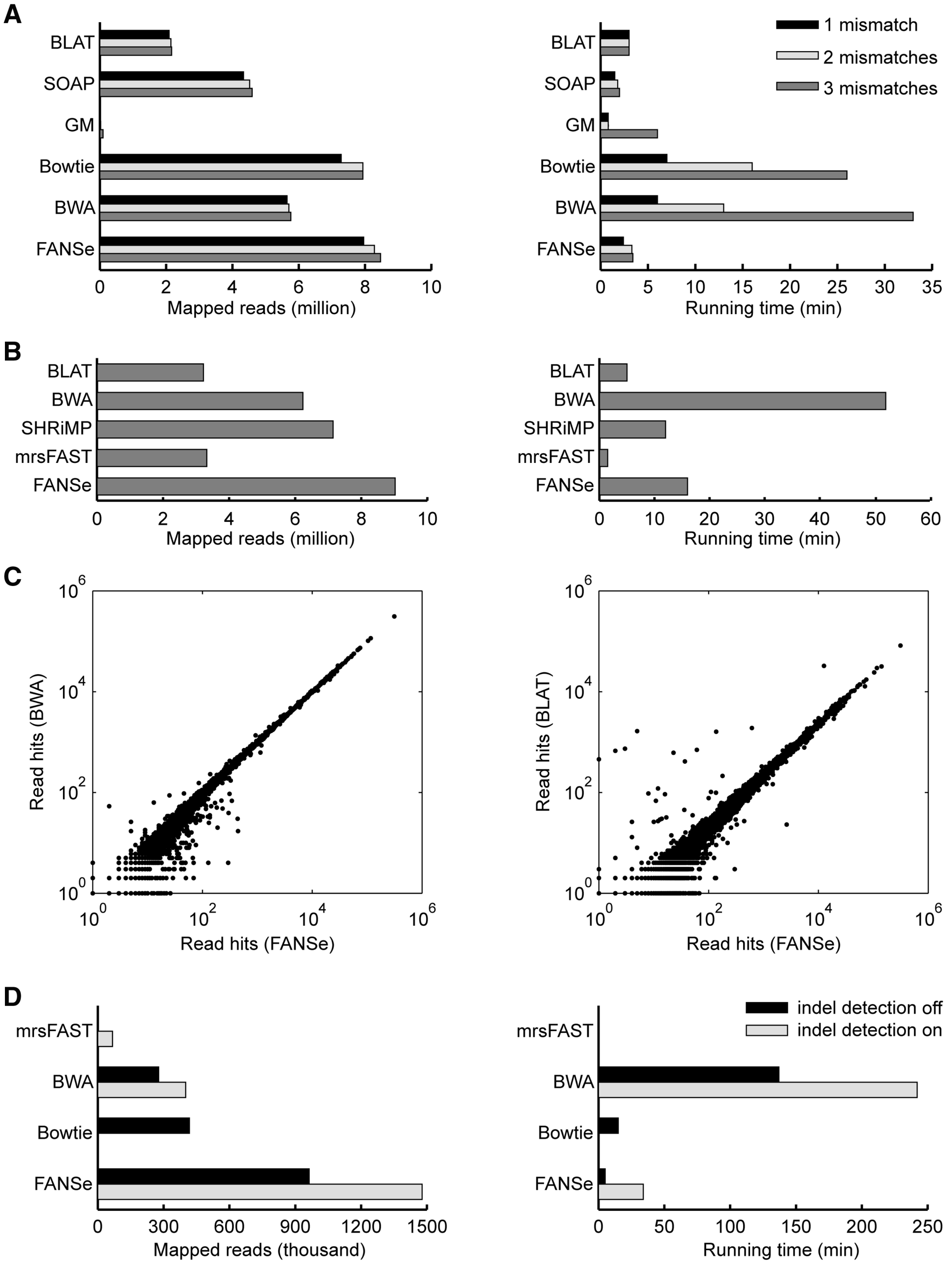

Figure 2. Sensitivity and speed of FANSe compared with other mapping algorithms. Mapped reads (left panels) and running time (right panels) for the mapping of *E. coli* mRNA random fragments to the reference genome with deactivated (**A**) or activated (**B**) indel detection using 8-nt seeds. One, two or three mismatches were allowed when indel detection was switched off. (**C**) Comparison of the read hits for the mRNA random fragments of each *E. coli* gene mapped by FANSe and BWA (left panel) or BLAT (right panel). (**D**) Mapped reads (left panel) and running time (right panel) by mapping the HeLa mRNA random fragments to the masked human chromosome 21 allowing three mismatches. Note that some algorithms were only run in indel-enabled (mrsFAST, BLAT) or indel-disabled mode (Bowtie). BLAT mapped 1459 reads within 1 min and is not included in the plots as it is out of scale compared with the other algorithms. GM, Genomemapper.

The FANSe algorithm showed a high level of sensitivity not only for mapping reads to small reference genomes, e.g. bacteria, but also to large eukaryotic genomes. We compared the performance of FANSe and the other algorithms in mapping reads generated by sequencing randomly fragmented mRNA from HeLa cells to the human chromosome 21 reference sequence. We used a masked genome sequence, in which the highly repetitive regions are already masked to avoid ambiguous multiple mapping of one read to the repetitive regions. FANSe mapped double amount of reads compared with the other algorithms (Figure 2D). When indel detection was enabled, the number of the mappable reads increased by 53% compared with 44% when using BWA. The SOAP, Genomemapper and SHRiMP algorithms do not support masked genomes; SOAP2 failed to run because of an

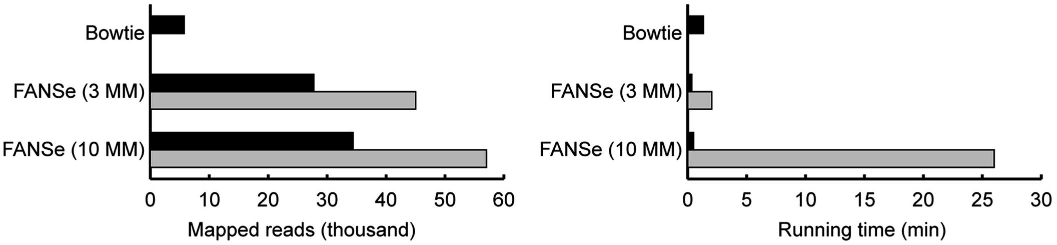

Figure 3. Comparison of the sensitivity and running time of FANSe and Bowtie on mapping long reads (>140 nt). *Escherichia coli* genomic DNA sequenced with a 454 GS FLX pyrosequencing platform (40) was used as a dataset. FANSe was set to allow 3 or 10 mismatches; three mismatches were allowed for Bowtie. Mapping was performed with the indel detection switched on (gray bars) or off (black bars). MM, mismatches.

internal error. Although some of these reads may be mapped to other human chromosomes with the same or even fewer mismatches, FANSe has the potential to report more locations of the alignments compared with the other algorithms if all mapping locations need to be reported instead of the just the best one. Such a requirement was recently demanded by some RNA quantification applications for eukaryotes (44). The high sensitivity of FANSe is a tradeoff with its speed: the running time of FANSe was slightly slower than the other seed-based algorithms (e.g. BLAT, SOAP and SHRiMP), particularly when indel detection was enabled (Figure 2A, B and D).

The majority of the read mapping algorithms are designed to map short reads with a maximum length of 60–127 nt. Tools to map long reads generated by sequencing platforms like 454 GS FLX are limited. Next, we tested the scalability of FANSe to map long reads. We mapped a dataset generated on the 454 GS FLX sequencing platform to the *E. coli* reference genome. Although DNA-seq methods have a lower error rate than the RNA-seq, it is still very likely that long reads (140–300 nt) contain more than three mismatches and indels. The number of mismatches in FANSe is flexible and we compared its mapping performance using 3 or 10 mismatch settings with a maximal allowance of three mismatches for Bowtie (Figure 3). FANSe mapped a higher number of reads compared with Bowtie, which only identified a small fraction of the mappable reads (Figure 3). To validate the mapping result of FANSe, we randomly chose 20 mapped reads (indel-free and indel-containing reads) and manually verified the unique and correct mapping of all these reads using the NCBI nucleotide BLAST tool. Clearly, by allowing a higher mismatch number, 24% more reads were mapped with FANSe without losing much speed (Figure 3). Furthermore, by enabling the Smith–Waterman refinement 66% more reads were mapped, albeit at a slower speed (Figure 3). The BWA-Smith Waterman Alignment algorithm, a variant of BWA that is designed to map long reads, failed to function, most likely due to its limitations on mapping small (i.e. bacterial) genomes (45). BLAT gave a large amount of mapped reads; however were mostly local alignments, only aligning part of the read instead of the whole read with the reference genome.

To avoid bias as a result of choosing the dataset and application, we used simulated, random sequencing datasets and compared the accuracy of FANSe and Bowtie. For both simulated *E. coli* reads and human chromosome 21 reads, FANSe and Bowtie achieved a comparably high level of sensitivity, ~100 %, when the substitution rate was varied from 0.5% to 1% (Figure 4A). Further increase in the substitution rate of up to 8% caused a decrease in the sensitivity of Bowtie by 30–80% depending on the read length, whereas the sensitivity of FANSe only decreased by 10% (Figure 4A). In almost all indel-free cases (Figure 4A) the correctness of FANSe ranged from 97.2% to 99.7% (average 98.8%) which is similar to the correctness of Bowtie (98.0–99.7%, average 98.8%). Increasing the mismatches from three to four decreased the number of unmapped reads by half (Figure 4A); however only a marginal decrease in the correctness, from 98.2% to 99.6% to 97.2% to 99.5%, was detected. The high sensitivity and correctness found when mapping *in silico*- generated datasets confirms the theoretically estimated accuracy (Supplementary Figure S1).

Furthermore, the sensitivity and correctness were almost identical for both the bacterial and eukaryotic reference sequences under the same settings, illustrating the high robustness of FANSe. For the more difficult datasets with a 1% substitution rate mixed with a 0.5–4% indel rate, FANSe provided a sensitivity higher than 99.7% and correctness between 95.9% and 98.6% (average of 98.0%) when the indel search was enabled, whereas Bowtie discarded a large fraction of mappable reads most likely due to its limited ability to handle indels (Figure 4B). Together these data underpin the advantage of using FANSe, particularly when mapping datasets with relatively high error rates.

Mapping speed and memory consumption of FANSe

Next, we tested the running time of FANSe and compared it with the other algorithms by recording it on the same computer with one CPU core. Construction of the index file is a separate step in some algorithms (FANSe, BWA, Bowtie, SOAP2 and Genomemapper) and the time consumed during this step was not included in our comparison because the file can be reused and therefore is not included in these comparisons. The time required to create an index file varies between the algorithms; e.g. construction of the lookup table for human chromosome 21 for FANSe only took several seconds. This step is integrated

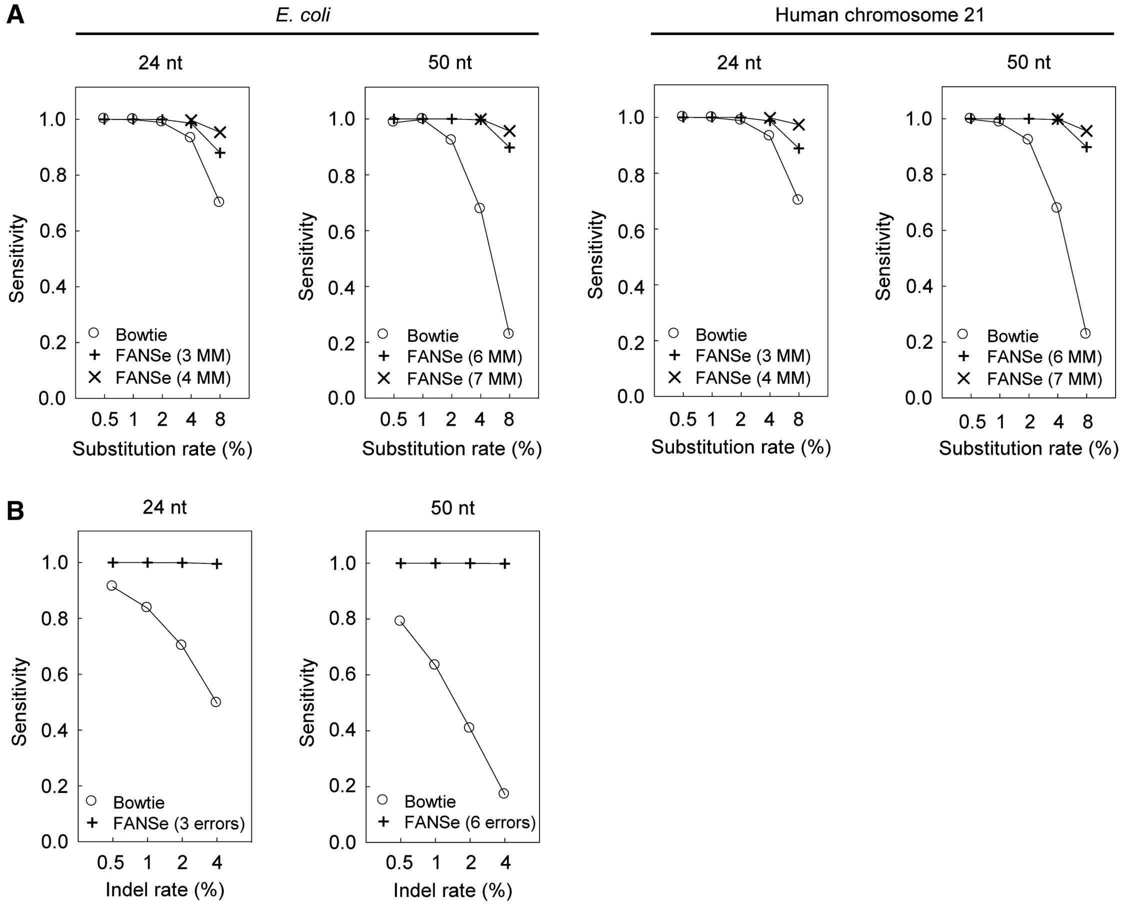

Figure 4. Comparison of the sensitivity between FANSe and Bowtie by mapping of *in silico* simulated datasets. (**A**) Sensitivity of mapping indel-free reads from the *E. coli* genome and masked human chromosome 21 reference sequence. FANSe was run with 6-nt seeds. (**B**) Sensitivity of mapping reads from the *E. coli* genome with a 1% substitution rate and an indel rate ranging from 0.5% to 4%. Indel search is enabled. All tests with Bowtie were run with three mismatch allowance. MM, mismatches.

within FANSe as it simplifies usage and saves disk space. For short reads, FANSe performed slower than the other algorithms when using 6-nt seeds due to its high accuracy (Supplementary Figure S2). When using 8-nt seeds for reads longer than 24 nt, the speed of FANSe increased by 2-fold, whereas a marginal decrease in sensitivity was detected and 0.07% of the reads were missed. When using 8-nt seeds for all reads, the speed was significantly faster than the BWT-based algorithms (Bowtie and BWA), whereas missing only 0.27% of the mappable reads when three mismatches were allowed (Figure 2A and B and Supplementary Figure S2). Ten million reads can be mapped to large genomes, e.g. mouse or human, on one quad-core computer within 1 day. When a few errors were allowed, enabling the Smith–Waterman refinement in order to detect all indels increased the running time two to five times for short reads and more than six times for long reads (Figures 2 and 3 and Supplementary Figure S2).

It should be noted that allowing a lot of mismatches while enabling indel detection significantly increased the running time due to the much larger area of the Smith–Waterman scoring matrix to be calculated, especially for long reads (Figure 3). For platforms that provide short reads and intrinsically have very low indel rates (e.g. Illumina platforms), enabling the indel detection only gained 6.5% more reads. For practical reasons, the indel search might be disabled when mapping short reads from these platforms. In the platforms with high indel rates (e.g. Helicos) or when long reads are generated (e.g. 454 GS FLX), where indels are more likely to occur, it is recommendable to enable the indel search. As the current sequencing platforms that generate very long reads (e.g. 454 pyrosequencing sequencers) do not provide multi-million reads in a single run, the current mapping speed is still acceptable when using multi-core processors.

The FANSe algorithm requires a memory approximately six times the size of the reference genome and is almost independent of the length of the reads and the errors allowed. This keeps the memory consumption within a reasonable range. In comparison, some algorithms require gigabytes of RAM even when mapping reads to the *E. coli* genome (Genomemapper, mrsFAST and SHRiMP) that is prohibitive, especially in the case of parallel computing of multiple datasets (Figure 5).

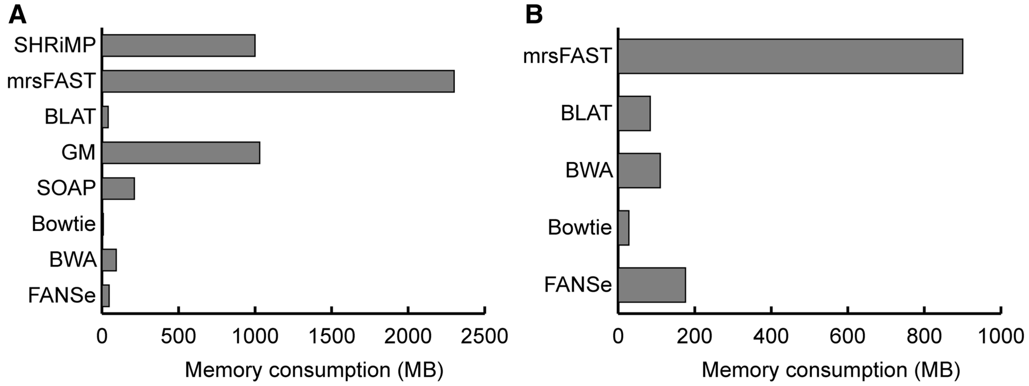

Figure 5. Memory consumption of different algorithms when mapping randomly fragmented *E. coli* (**A**) and HeLa (**B**) mRNA short reads. Panel A represents the memory consumption when running the mapping with the same parameter as in Figure 2A; panel B corresponds to the mapping shown in Figure 2D. GM, Genomemapper.

DISCUSSION

Massively parallel deep-sequencing technology provides a powerful tool for unraveling new biological information and brings new challenges to data processing. The alignment or mapping of the reads to a reference genome is a fundamental step of the data processing from RNA-seq and all subsequent analyses are based on it. Therefore, the accuracy of the mapping algorithm (including sensitivity and correctness) is crucial, whereas the speed might be considered as a subordinate feature. We developed a seed-based mapping algorithm that performs with a high sensitivity and accuracy when aligning the reads to small (*E. coli* genome) and large (e.g. human masked genome) reference genomes. Even at high error rates of the sequencing datasets, FANSe maintains a high sensitivity. The flexibility in the allowance of mismatches increases the accuracy and coverage of data processing. This is particularly crucial in transcriptome analysis, since prior to sequencing, RNA is converted into cDNA by reverse transcriptase whose fidelity is imperfect and may introduce multiple mis-incorporations at modified nucleotides, thus adding an extra error to the error rate generated by the sequencing machine itself.

Furthermore, the flexible mismatch settings within a read and the ability to completely detect indels provide advantages in the analysis of genetic variations (e.g. SNPs) and methylation patterns. The high sensitivity of FANSe is traded off with mapping speed which compared with the other seed-based algorithms is slightly slower. The implementation of an accelerated Smith–Waterman refinement increases the speed without compromising the accuracy and reduces the computational cost compared with the traditional Smith–Waterman algorithm (9). FANSe is the first mapping algorithm that provides theoretical estimation of the sensitivity thus allowing for chosing the best parameter sets to achieve the desired sensitivity.

The mRNA abundance of different genes differs by more than three orders of magnitude (46) and low-abundance mRNAs are a major technical challenge for transcriptome sequencing. Incomplete mapping can lose critical information and create a significant bias in quantification and downstream analyses. Notably, FANSe shows a remarkable coverage of low-abundance mRNAs: the mapping increased between 7% and 131% for short reads and by six times for long reads compared with other mapping algorithms when run at comparable settings (Figures 2 and 3). Currently, FANSe does not detect reads across the splicing junctions like some other mapping programs (e.g. BLAT). However, FANSe applications can be extended towards detecting splicing junctions. Alternatively, an algorithm that is designed to specifically detect splicing junctions [e.g. TopHat (14) and MapSplice (4)] can be used to process any reads that fail to be directly mapped to the reference genome.

The deep understanding of the diversity in biology and human disease biology is dependent on accurate genome sequencing. The increasing variety of deep sequencing techniques and applications requires versatility of the data processing tools. There is a great variability in the maturity of the available computational tools (25). We believe that FANSe will find broad applications as it can accurately and sensitively map millions of short reads with different lengths from a large variety of platforms, containing different mismatch, error and indel rates. Importantly, FANSe can stably map to both short and large reference genomes and even to masked genomes or reference sequences containing unspecific nucleotides (‘N’-s). Finally, FANSe can be compiled in various operating systems (Windows, Linux, MacOS, etc.) thus it is suitable for users who might only be familiar with one operating system.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1 and 2, Supplementary Methods and Supplementary References [47,48].

ACKNOWLEDGEMENTS

We are grateful to Wei Chen and Na Li (MDC Berlin) for their help with the Illumina GAIIx sequencing instrument.

We thank Samuel Arvidsson (Max Planck Institute of Molecular Plant Physiology) and Detlef Groth (Bioinformatics, University of Potsdam) for a constructive discussion on comparing algorithms and Shuai Liu (Jinan University) for technical hints on performance optimization.

FUNDING

Funding for the open access charge: The Deutsche Forschungsgemeinschaft (Grants IG 73/10-1 and TP A5 within SFB 740 to Z.I.).

Conflict of interest statement. None declared.

REFERENCES

1. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
2. Li,R., Fan,W., Tian,G., Zhu,H., He,L., Cai,J., Huang,Q., Cai,Q., Li,B., Bai,Y. *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.
3. Paszkiewicz,K. and Studholme,D.J. (2010) De novo assembly of short sequence reads. *Brief. Bioinform.*, **11**, 457–472.
4. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
5. Ossowski,S., Schneeberger,K., Clark,R.M., Lanz,C., Warthmann,N. and Weigel,D. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.*, **18**, 2024–2033.
6. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
7. Park,Y.J., Claus,R., Weichenhan,D. and Plass,C. (2011) Genome-wide epigenetic modifications in cancer. *Prog. Drug Res.*, **67**, 25–49.
8. Iida,K., Jin,H. and Zhu,J.K. (2009) Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in Arabidopsis thaliana. *BMC Genomics*, **10**, 155.
9. Lipson,D., Raz,T., Kieu,A., Jones,D.R., Giladi,E., Thayer,E., Thompson,J.F., Letovsky,S., Milos,P. and Causey,M. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.*, **27**, 652–658.
10. Magi,A., Benelli,M., Gozzini,A., Girolami,F., Torriceli,F. and Brandi,M.L. (2010) Bioinformatics for next generation sequencing data. *Genes*, **1**, 294–307.
11. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
12. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Trapnell,C. and Salzberg,S.L. (2009) How to map billions of short reads onto genomes. *Nat. Biotechnol.*, **27**, 455–457.
15. Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.*, **11**, 473–483.
16. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
17. Li,R., Li,Y., Kristiansen,K. and Wang,J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.
18. Schneeberger,K., Hagmann,J., Ossowski,S., Warthmann,N., Gesing,S., Kohlbacher,O. and Weigel,D. (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol.*, **10**, R98.
19. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
20. Lunter,G. and Goodson,M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.*, **21**, 936–939.
21. Rumble,S.M., Lacroute,P., Dalca,A.V., Fiume,M., Sidow,A. and Brudno,M. (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.*, **5**, e1000386.
22. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
23. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
24. Li,R., Yu,C., Li,Y., Lam,T.W., Yiu,S.M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.
25. Garber,M., Grabherr,M.G., Guttman,M. and Trapnell,C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.
26. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
27. Huse,S.M., Huber,J.A., Morrison,H.G., Sogin,M.L. and Welch,D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, R143.
28. Homer,N., Merriman,B. and Nelson,S.F. (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, **4**, e7767.
29. Prufer,K., Stenzel,U., Dannemann,M., Green,R.E., Lachmann,M. and Kelso,J. (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, **24**, 1530–1531.
30. Krawitz,P., Rodelsperger,C., Jager,M., Jostins,L., Bauer,S. and Robinson,P.N. (2010) Microindel detection in short-read sequence data. *Bioinformatics*, **26**, 722–729.
31. Cooper,G.M. and Mefford,H.C. (2011) Detection of copy number variation using SNP genotyping. *Methods Mol. Biol.*, **767**, 243–252.
32. Bibikova,M. and Fan,J.B. (2010) Genome-wide DNA methylation profiling. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **2**, 210–223.
33. Zilberman,D. and Henikoff,S. (2007) Genome-wide analysis of DNA methylation patterns. *Development*, **134**, 3959–3965.
34. Hebenstreit,D., Fang,M., Gu,M., Charoensawan,V., van Oudenaarden,A. and Teichmann,S.A. (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.*, **7**, 497.
35. Ingolia,N.T. (2010) Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol.*, **470**, 119–142.
36. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of Mammalian proteomes. *Cell*, **147**, 789–802.
37. Oh,E., Becker,A.H., Sandikci,A., Huber,D., Chaba,R., Gloge,F., Nichols,R.J., Typas,A., Gross,C.A., Kramer,G. *et al.* (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*, **147**, 1295–1308.
38. He,S., Wurtzel,O., Singh,K., Froula,J.L., Yilmaz,S., Tringe,S.G., Wang,Z., Chen,F., Lindquist,E.A., Sorek,R. *et al.* (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods*, **7**, 807–812.
39. Guo,H., Ingolia,N.T., Weissman,J.S. and Bartel,D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.
40. Turnbaugh,P.J., Ley,R.E., Hamady,M., Fraser-Liggett,C.M., Knight,R. and Gordon,J.I. (2007) The human microbiome project. *Nature*, **449**, 804–810.
41. Hach,F., Hormozdiari,F., Alkan,C., Birol,I., Eichler,E.E. and Sahinalp,S.C. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods*, **7**, 576–577.
42. Ruffalo,M., LaFramboise,T. and Koyuturk,M. (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, **27**, 2790–2796.

43. Albers,C.A., Lunter,G., MacArthur,D.G., McVean,G., Ouwehand,W.H. and Durbin,R. (2010) Dindel: accurate indel calls from short-read data. *Genome Res.*, **21**, 961–973.
44. Nagaraj,N., Wisniewski,J.R., Geiger,T., Cox,J., Kircher,M., Kelso,J., Paabo,S. and Mann,M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.*, **7**, 548.
45. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
46. Bernstein,J.A., Khodursky,A.B., Lin,P.H., Lin-Chao,S. and Cohen,S.N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl Acad. Sci. USA*, **99**, 9697–9702.
47. Creighton,C.J., Reid,J.G. and Gunaratne,P.H. (2009) Expression profiling of microRNAs by deep sequencing. *Brief. Bioinform.*, **10**, 490–497.
48. Hafner,M., Landgraf,P., Ludwig,J., Rice,A., Ojo,T., Lin,C., Holoch,D., Lim,C. and Tuschl,T. (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, **44**, 3–12.