Introduction

Mapping (aligning) millions of next-generation sequencing (NGS) reads accurately to reference sequences is the basis of all deep sequencing applications that utilize reference genomes or transcriptomes, including variant analysis, gene expression and isoform analysis. Traditional alignment algorithms such as BLAST and BLAT could not process the massive amount of sequencing data in hours [reviewed in [1]]. A series of early mapping algorithms such as SSASHA, MAQ and SOAP started to tackle this speed hindrance. These algorithms extended the basic idea of “seeding” (hash table indexing) from BLAST, which is simple in design and easy to implement, bringing the NGS technology into quantitative era [reviewed in [2,3]]. The computational time of this type of algorithms is less error-tolerant and thus usually less sensitive than seed-based algorithms at higher error rate (reviewed in [12]). Due to the principle of BWT, the blowout of NGS applications. According to a statistics till the end of 2012, two among the top three cited mapping algorithms are of this type (Bowtie and BWA) [reviewed in [6]]. In real-world benchmarks, although the sensitivity of earlier BWT-based algorithms like Bowtie and SOAP2 (<80%) is still to be improved when mapping DNA sequencing reads, the sensitivity of the upgraded Bowtie2 is almost the same as the traditional seed-based algorithms while being more than 20x faster [6].

However, deviations between reads and reference sequences set a great challenge of the sensitivity and speed to the mapping algorithms. Origins of the deviation include single nucleotide polymorphisms [reviewed in [8]], PCR amplification [9], base calling [reviewed in [10]] and sequencer errors [11]. When the mismatch rate exceeds 2% or the indel rate exceeds 0.5%, most algorithms lose their accuracy [12]. Due to the principle of BWT, this type of algorithms is less error-tolerant and thus usually less sensitive than seed-based algorithms at higher error rate [reviewed in [1,7]]. For RNA-seq, the error rate is higher due to RNA editing [13], modifications [14] and nucleotide misincorporation in reverse transcription [15]. Indeed, in simulated tests, Bowtie and BWA remained 55%~75% accuracy at 4% error rate, while the seed-based algorithms SOAP and Novoalign maintain 80%~90%
accuracy [12]. This result coincides with the real-world test: even when adding the splice-mapped reads, BWT-based algorithms TopHat and SOAPsplice mapped 12–19% of reads less than seed-based algorithms [6]. These algorithms tend to unproportionally lose mappable reads of the medium to low abundance RNA, generating a significant bias in quantification [5]. Moreover, the accuracy of BWT-based algorithms was shown to be highly dependent on the dataset in various comparative tests, from very high [7] to moderate [6,12] to very low [16], in contrast to the seed-based algorithms. The inconsistency of quantitative results given by RNA-seq and microarray may reflect this bias and unrobustness [17–19].

It would be ideal to combine the speed of BWT and the robust accuracy of seed-based algorithms, especially for the cases with higher error rates like RNA-seq. To improve the robustness and indel detection of the BWT-based algorithm Bowtie, the upgraded Bowtie2 partially took the advantage of the seeding principle, and it truly exceeded Bowtie, BWA and SOAP2 [20]. However its accuracy and robustness are difficult to be theoretically estimated. To overcome this problem, we took the advantage of our previously developed FANSe algorithm, which is a seed-based algorithm with theoretical estimation of high accuracy and robustness (miss rate can be as low as 10⁻⁶) [5], and further developed FANSe2 algorithm. FANSe2 can map a billion reads to human genome in hours using normal office computers without compromising the high and robust accuracy. We also tested this algorithm using real-world RNA-seq datasets and experimentally validated its results by RT-PCR and microarray.

**Materials and Methods**

**Design of FANSe2**

FANSe2 is an iterative and parallel seed-based read mapping algorithm with a simple design to ensure all advantages of FANSe and largely improve the speed and parallelization. The following major steps were implemented: (Figure 1).

**Step 1.** Segmentation of reference sequences. To reduce the memory consumption, large reference sequences like human genome are split to segments. Two adjacent segments are overlapped with maximum read-length. Each segment will be processed as a task package and assigned to a processor core.

**Step 2.** Initialize parallel computing environment. To avoid resource competition, FANSe2 parallelizes multiple processes via the industrial standard MPICH2 environment instead of multi-threading. Unlike FANSe that uses the 6- or 8-nt seeds, FANSe2 initially set the seed length as 14-nt.

**Step 3.** Each CPU core starts to process the assigned task package, mapping all reads to the reference sequence segment using the seed length based on the principle of FANSe. The final refinement of hotspots is performed by calculating Hamming distance (indel detection off) or by using accelerated Smith-Waterman method (indel detection on) [5].

**Step 4.** After all the task packages are processed, the mapping results are combined and the best mapping location of a read is written to the final result file.

**Step 5.** FANSe2 decreases the seed length by 2-nt and tries to map the unmapped reads using the shorter seed length (back to step 3). Iterative mapping process stops when the seed length reaches the minimum seed length or all the reads are mapped.

**Datasets and reference sequences**

To analyze the nucleotide error distribution in the sequencing datasets, we downloaded six datasets from DDBJ Sequence Read Archive [http://trace.ddbj.nig.ac.jp/dra/index_e.shtml], as listed in Table S1 in File S1. Each read was truncated at the nucleotide, whose sequencing quality is lower than 20 in Phred scale. Reads shorter than 18 nt were discarded. The *E. coli* datasets were mapped to *E. coli* K-12 substrain MG1655 genome sequence (NCBI Reference Sequence: NC_000913.2). The yeast datasets were mapped to *S. cerevisiae* genome sequence sacCer3 (downloaded from UCSC genome browser, http://hgdownload.cse.ucsc.edu). FANSe was used to perform these mappings with the errors allowed as listed in Table S1 in File S1 and indel detection on.

The *E. coli* mRNA dataset reported previously was used to test the sensitivity and speed of FANSe2 [5]. The datasets of the whole Flowcell A (FCA) of Human Body Map 2.0 project, containing altogether 608 million 75-nt reads of human polyA⁺ mRNA sequenced on an Illumina HiSeq-2000 sequencer, were used to test the parallel computing capacity of FANSe2. The human datasets were mapped to human genome sequence hg19/GRCh37 (downloaded from UCSC genome browser).

Simulated datasets with 2% and 4% error rate were generated from human chromosome 1 non-masked and masked genome sequence (hg19/GRCh37). Each datasets contained 500,000 reads, 75-nt long. These reads were generated from the non-masked regions. These datasets were mapped to human chromosome 1 non-masked and masked genome sequence, respectively. Reads with homopolymeric stretch or dinucleotide repeats longer than half of the read length were filtered out to avoid unnecessary and ambiguous alignment [21].

---

**Figure 1. Flowchart of FANSe2.** For details please refer to the Materials and Methods section. SL = seed length. doi:10.1371/journal.pone.0094250.g001
Comparison of mapping programs

We compared FANSe2 with FANSe, Bowtie, Bowtie2, BWA, SHRIMP2 and Novoalign (for details please refer to Table S3 in File S1). The performance tests were carried out on quad-core Intel i5-3570K computers with 16 GB RAM. We used -n 3 — tryhard—best for Bowtie and -n 7 -o 1 for BWA. Unless specified, —very-sensitive option was used for all Bowtie2 tests. The memory consumption of these programs was recorded using either Task Manager (Windows) or System Monitor (Linux).

Sensitivity and correctness were defined previously [5]. In brief, a read that is truly originated from the reference sequence can have one of the following three outcomes after being processed by an algorithm: (i) mapped to its correct position (C); (ii) mapped to a wrong position (J); (iii) failed to be mapped to the reference genome (U). Sensitivity is defined as $\frac{C}{C+U}$, and the correctness is defined as $\frac{C}{C+J}$. Sensitivity can be calculated from a deep-sequencing dataset, which is proportional to the number of mapped reads. Correctness can be only evaluated using simulated random datasets.

Results

Iterative step-down acceleration strategy based on the real-world alignment error distribution

When mapping a read, FANSe takes seeds (6- or 8-nt long) from the read and searches for exact matches in the reference genome with a pre-built look-up table [5]. These exact matches are then merged into hotspots and then refined to determine the best mapping location. An n-nt long seed has in average $N/n^4$ exact matches in the genome (where $N$ is the genome size), a large number for large genomes and $n=8$, thus creating a heavy workload for the hotspot merging and refinement, especially when indel detection is enabled. Longer seeds decrease the number of exact matches exponentially and thus largely accelerate the mapping: 14-nt seed decreases the number of exact matches by $\frac{1}{4^{14-8}} = 4096$ folds than 8-nt seeds. However, longer seeds are more likely to contain error and may lose the reads with higher number of mismatches, thus impairing the sensitivity. A read containing maximum $f$ errors with a minimal read length of $n(f+1)$ can be reliably mapped to a genome when using $n$-nt seeds, indicating that a long read with a few errors may be still stably mapped with longer seeds (Figure 2A). For example, up to 5 errors are guaranteed to be detected in 75-nt reads using 14-nt seeds. To achieve theoretical miss rate less than 1%, 12-nt seeds are sufficient for 50-nt reads, whereas 14-nt seeds are more likely to map reads in much higher speed. Therefore, we implemented an iterative step-down strategy: long seeds (e.g. 14-nt or 12-nt) are used to map most reads with high speed, and the unmapped reads (a small fraction) are mapped in the next iteration with shorter seed. This iterative process terminates when the seed length reaches the limit set by the user (Figure 1).

We tested this strategy with the E. coli mRNA-seq dataset that was previously used in FANSe test [5]. Stepping down to 8-nt seed length, FANSe2 exported the same mapping result as FANSe at much faster speed using single CPU core when allowing 3 mismatches (Figure 2D and 2E). Indeed, most of the mappable reads were mapped in the initial iteration using 14-nt seeds. When stepped down to 12-nt seeds, FANSe2 mapped 8.26 M reads using 0.28 minutes in total. At this stage, the sensitivity is already higher than the widely-used Bowtie and Bowtie2 (7.93 M and 8.12 M reads, respectively, Figure 2D), while faster than Bowtie2 (1.13 minutes). Stepping further down to 8-nt stage may not be practically necessary, since this significantly increased the running time for three times, however only mapped 0.21 M more reads. Even down to 8-nt stage, the speed of FANSe2 is 3~21x faster than FANSe, Bowtie and BWA, only slightly slower than Bowtie2 (Figure 2E).

Memory consumption, speed and scalability when handling huge datasets

The memory consumption of FANSe2 is tunable by the user, because it is only relevant to the genome segment size: FANSe2 uses 1.2~1.7 GB memory for each activated CPU core when the reference sequences are split to 50~200 MB segments (Figure 3A).
FANSe2: A Robust and Fast Mapping Tool for Next Generation Sequencing

A

B

C

D

E
Therefore, FANSe2 can accelerate mapping large datasets by using 2~4 CPU cores on one computer using 4~8 GB RAM. This means that even laptop computers can perform mapping with ease. When mapping reads to human genome, Bowtie2 and Bowtie needs more than 5 GB RAM. BWA and novoalign needs 7.3~7.8 GB RAM, which hardly fits a computer with 8 GB RAM because the operating system usually requires additional 0.7~1.2 GB RAM (Figure 3A). According to the manual, SHRiMP2 needs 48 GB RAM to map reads to human genome, which is already far beyond the capacity of high-end workstations, including our computers (Figure 3A and 3B).

In addition, FANSe can parallelize across multiple normal computers with simple LAN connection, providing an economic and scalable solution for biology labs. This feature is not offered by any other current mainstream mapping tools. We tested the scalability of FANSe2 in our real office environment with three heterologous computers: two Intel i5-3570s and one Intel i5-2500 with 8 GB~16 GB RAM installed, connected with gigabit LAN. One such inexpensive office computer (~$600) mapped an mRNA-seq dataset of Human Body Map 2 generated from an entire Illumina HiSeq-2000 flowcell (8 lanes, 608 M reads, 75-nt) to human reference genome within 10 hours, 3.6x faster than

Figure 2. Rational and validation of the iterative strategy of FANSe2. (A) Errors in a read which can be perfectly detected by FANSe algorithm versus the seed length for 50-, 75- and 100-nt reads, respectively. (B) Theoretical miss rate of FANSe algorithm with different seed length for various read length. (C) Error distribution of six sequencing datasets (listed in Table S1 in File S1) sequenced on various types of sequencers, respectively. Reads were mapped with FANSe. ARL = average read length. (D, E) Mapping the E. coli mRNA dataset reported in [5]. For FANSe2 algorithm, the reads mapped (D) and the calculation time (E) used using different read length stages were shown in colors. The test was performed in a quad-core Intel i5-3570K computer using one CPU core.

doi:10.1371/journal.pone.0094250.g002

Figure 3. Scalability, sensitivity and speed of FANSe2 compared to other algorithms. (A) Memory consumption of the tested algorithms when mapping 75-nt reads to human genome. The memory consumption of FANSe2 using 1 CPU core and 4 CPU cores are indicated using circles and diamonds, respectively. SHRiMP2 failed to run this test in our 16 GB memory system; thus its memory consumption was taken from its manual. (B, C) Mapping data from an entire Illumina HiSeq-2000 flowcell (8 lanes, 608 M 75-nt reads) to masked human genome. (B) The number of reads mapped by the tested algorithms using one computer (4 CPU cores). FANSe2 was tested with indel detection on and off, respectively. SHRiMP2 failed to run in our system due to its high memory consumption. Novoalign failed to finish the task within 96 hours. (C) The number of CPU cores and computers. Plus sign: FANSe2 without indel detection; cross: FANSe2 with indel detection. Bowtie2 (circle) and BWA (rectangle) do not support automatic parallelization across multiple computers.

doi:10.1371/journal.pone.0094250.g003
Figure 4. Comparison of FANSe2 and other algorithms on their sensitivity, speed and correctness using simulated datasets from non-masked and masked human chromosome 1 reference sequence (hg19) with 2% and 4% sequencing error rate, respectively. Each dataset contains 500,000 reads with the length of 75-nt. The test parameters are listed in Table S4 in File S1. (A) Comparative test on sensitivity and speed. Reads mapped and the time used at different stages of seed lengths in FANSe2 are shown in colors. The BWT-based algorithms are shown in light blue bars, and the other seed-based algorithms are shown in gray bars. (B) Comparative test on correctness. For Bowtie2, BWA and novoalign, mapped reads were filtered using various mapping quality threshold (Q0 → Q20) represented in Phred score scale (black circle). The correctness of FANSe2 results were marked on the same plot when considering all mapping results (red triangle, 5–7 errors allowed) or considering only the reads.
FANSe2: A Robust and Fast Mapping Tool for Next Generation Sequencing

that were uniquely mapped (red cross, 7 errors allowed). The results of Bowtie and SHRiMP2 were not filtered according to the mapping quality due to their low mapping sensitivity.

doi:10.1371/journal.pone.0094250.g004

Bowtie2 in very sensitive mode while maintaining the same sensitivity. With three computers and one-click run, the same job finished 4.1 hours by FANSe2 (Figure 3B and 3C). With the indel detection enabled, FANSe2 mapped these reads within 23.9 hours with three computers. FANSe2 with indel search mapped more reads than Bowtie2 and BWA, which were also enabled indel search (Figure 3B). Compared with one computer, three computers accelerated the mapping for 2.43x (Figure 3C). Note that this efficient parallelization was performed with user-friendly graphical user interface. In contrast, SHRiMP2 failed to run because of its high memory demand. Novoalign was unable to finish the task in 4 days (Figure 3B). These results showed that FANSe2, as a seed-based algorithm, is approaching the speed of BWT-based algorithms while maintaining similar or higher sensitivity when handling huge datasets.

Sensitivity and correctness of FANSe2 tested with simulated dataset

Practically, the raw error rates of the current next-generation sequencing platforms were reported as 0.26~13% [11], and the base calling step adds further 2.76~4.86% error rate [10]. Therefore, a mapping algorithm should reliably map reads containing at least such errors. To test the sensitivity and correctness of FANSe2 algorithm, we generated four simulated datasets, each containing 500,000 reads of 75-nt Illumina-like single-end reads, from the non-masked and masked human chromosome 1 genome sequence (hg19) and with substitution rate of 2% and 4%, respectively. For all four cases, the speed of traditional BWT-based algorithms (Bowtie, Bowtie2 and BWA) are generally faster than traditional seed-based algorithms (SHRiMP2 and novoalign). This coincides with the previous comparisons [6]. However, FANSe2 is just slightly slower than BWT-based algorithms in all cases, and is even faster than Bowtie2 when using the masked genome. In all four cases, FANSe2 mapped more reads than all other tested algorithms (Figure 4A). The sensitivity of FANSe2 increased slightly when allowing more errors in a read. When 7 errors were allowed, the sensitivity of FANSe2 reached 99.99% and 99.0% for 2% and 4% error rate, respectively. Again more than 99% of the reads were mapped using 14-nt seeds, exceeding the sensitivity of all other tested algorithms. Stepping down to 12-nt or lower hardly mapped more reads, thus is practically unnecessary. Note that the error allowance for the whole read cannot be explicitly set when using Bowtie2 and novoalign. Some reads mapped with 7 errors were found in their results, showing that this comparison is fair.

Next, we analyzed the correctness of FANSe2 mapping results using the previously described method [20], plotting the number of reads mapped to wrong locations against the number of reads mapped to correct locations (Figure 4B). In all cases, FANSe2 allowing 6~7 mismatches mapped more reads correctly to its original position than all other tested algorithms. For the non-masked genome, FANSe2 allowing 7 mismatches mapped 2.2% and 6.9% more reads to their correct positions than Bowtie2 at 2% and 4% error rate, respectively. Meanwhile, FANSe2 mapped 41.0% and 44.5% less reads than Bowtie2 to their wrong positions. Applying increasing mapping quality threshold, only the uniquely mapped reads were kept. Bowtie2 decreased the wrongly mapped reads in the cost of discarding a considerable fraction of mappable reads. At the threshold of mapping quality of 5, FANSe2 mapped 4.5% and 27.8% more uniquely-mapped reads to its correct place than Bowtie2 at 2% and 4% error rate, respectively. BWA performed more robust than Bowtie2 in this test, as increasing the mapping quality threshold do not decrease the number of mapped reads dramatically. However it still mapped less reads than FANSe2. Novoalign mapped comparable number of reads as Bowtie2 and BWA, however it mapped 2~3 times more mapped to wrong places than Bowtie2 and BWA, and increasing the mapping quality threshold almost do not increase the correctness. Bowtie and SHRiMP2 mapped considerably less reads than the other algorithms, especially at 4% error rate.

As repetitive sequence creates challenges to correct read mapping, masked genome sequence is widely used in major studies to improve the efficiency of sequence alignment (e.g. NCBI BLAST) [21], polymorphism and mutation discovery [28,29], genome annotation and comparison [30–33], etc. In clinical diagnosis procedures, such as the non-invasive prenatal diagnosis based on next-generation sequencing, mapping reads to masked human genome is also used as a standard [34–37]. Therefore, we also performed read mapping tests using the masked genome sequence provided by UCSC Genome Browser. Compared to the non-masked tests, the sensitivity and correctness of all algorithms increased slightly, because the masked genome sequence is free of repetitive regions. Nevertheless, the scenario remains similar to the non-masked tests: FANSe2 has higher sensitivity while maintaining the correctness.

Experimental validation of the RNA-seq mapping result by FANSe2

The robust sensitivity and correctness of FANSe2 maximizes the usage of data in sequencing datasets. This advantage may be more significant when dealing with RNA-seq data that is more error-prone than DNA-seq. In our previous work, we had shown that BWA and BLAT lose mappable reads in low abundance mRNA unproportionally in a prokaryotic system [5]. We next tested FANSe2 and Bowtie2 with our previously reported mRNA-seq dataset (75 nt single-end reads) of human lung cancer cell line A549 [22]. Aiming at quantitative profiling of known mRNAs, we mapped the reads to RefSeq human RNA reference sequence and the splice variants were merged. Previous study showed that mapping to mature mRNA sequence avoided the error of mapping splice junction reads when using genomic sequence as reference and should be preferentially used for RNA-seq, unless novel splice junctions are to be detected [38,39]. Additionally, protein coding mRNAs consist only a small proportion of the genomic sequence, reducing the computational demand dramatically. Therefore this is an efficient strategy that is widely used by the community [23,38–43]. Genes with less than 10 reads mapped were considered as unreliable quantified genes and removed [23]. We found that the gene expression quantitation of the two algorithms in general coincide for the genes that were identified by both algorithms (Figure 5A).

We next experimentally investigated the genes that were solely identified by FANSe2 (Table 1) or Bowtie2 (Table 2) to check the possible false positives and false negatives. The abundances of the top five RNAs that solely identified by FANSe2 range from 1.47 to 5.77 rpKM (Table 1). They were all validated by RT-PCR with clear bands on the gel at the estimated sizes (Figure 5B). Although the primer specificity of SPIN2A was not high enough so that additional bands appeared in addition to the strongest and expected band, SPIN2A has been detected by microarray in lung
Figure 5. Experimental examination of the results of FANSe2 and Bowtie2. (A) Quantification of mRNA from A549 cells using the mapping results of FANSe2 and Bowtie2. The mRNA sequencing dataset was mapped to the human RefSeq RNA reference sequences and the quantification was performed using the standard rpkM method. (B) RT-PCR validation of mRNAs that were detected by FANSe2 but not by Bowtie2 (see Table 1). 15 μl PCR product were loaded for each lane and resolved on a 3% agarose gel. The bands with the expected product size were marked with stars.
The expected product sizes were noted below. (C) RT-PCR validation of mRNAs that were detected solely by Bowtie2 (See Table 2). Two RNAs detected solely by FANSe2 (LOC647859 and PPIAL4F) were loaded as positive control. 7 μl PCR product were loaded for each lane and resolved on a 2.7% agarose gel. The bands with the expected product size were marked with stars. The expected product sizes were noted below. A faint band appeared at ~200 bp in the lane of BCL2L2-PABPN1 but is quite different than the expected product size.

Table 1. The top five RefSeq RNAs that are exclusively identified by FANSe2 in A549 mRNA-seq dataset.

| Gene name | RefSeq-ID | Read count | rpKM | Whole Transcriptome qPCR Primer Database ID | Expected product size (bp) | Validated (Figure 5) |
|-----------|-----------|------------|------|--------------------------------------------|---------------------------|---------------------|
| PCDHG3    | NM_018924 | 105        | 1.47 | PCDHG3_unc003jw2.1_2_2                     | 111                       | Yes                 |
| SPIN2A    | NM_019003 | 60         | 2.92 | PB *                                       | 103                       | Yes                 |
| LOC647859 | NR_026578 | 52         | 5.77 | OCLN_unc11trc1.1_2_1_2                    | 78                        | Yes                 |
| PNMA6A    | NM_032882 | 45         | 1.95 | PB *                                       | 74                        | Yes                 |
| PPIAL4F   | NM_00116426 | 44      | 3.60 | PB *                                       | 181                       | Yes                 |

*PB: primer pair not available in whole transcriptome qPCR primer database. The primers are automatically designed using NCBI-PrimerBLAST. Please refer to Table S2 in File S1 for details.
Table 2. The top 20 RefSeq RNAs that are exclusively identified by Bowtie2 in A549 mRNA-seq dataset.

| Gene name                  | RefSeq-ID | Read count | rpkM | Identical to gene ** | Whole Transcriptome qPCR Primer Database ID | Expected product size (bp) | Validated (Figure 5) |
|----------------------------|-----------|------------|------|----------------------|-------------------------------------------|--------------------------|----------------------|
| SENP3-EIF4A1               | NR_037926 | 9440       | 146.46 |                      |                                           |                          |                      |
| LIZP6                      | NM_00128619 | 3264    | 54.95 |                      | MTPN                                      |                          |                      |
| RPL36A-HNRNPHP2            | NM_001199973 | 2653    | 65.24 |                      | RPL36A-HNRNPHP2_uc022cag.1_3_1_2          | 93                       | Failed              |
| RPS10-NUDT3                | NM_001202470 | 2240    | 57.96 |                      | PB *                                      | 993                      | Failed              |
| SLMO2-ATP5E                | NR_037929 | 1490     | 94.87 |                      |                                           |                          |                      |
| BLOC1SS-TXNDC5             | NR_037616 | 1336     | 26.06 |                      |                                           |                          |                      |
| BCL2L2-PABPN1              | NM_001199864 | 1297    | 39.82 |                      | BCL2L2-PABPN1_uc001wjd.4_2_2_1            | 109                      | Failed              |
| HSPE1-MOB4                 | NM_001202485 | 1119    | 17.10 |                      | HSPE1-MOB4_uc021wum.1_1_1_2               | 77                       | Failed              |
| HNRNPU1L2-BSCL2            | NR_037946 | 1092     | 17.79 |                      |                                           |                          |                      |
| HIF1A-AS2                  | NR_045406 | 985      | 31.40 |                      |                                           |                          |                      |
| ATP6V1G2-DDX39B            | NR_037853 | 969      | 27.49 |                      |                                           |                          |                      |
| COMMD3-BMI1                | NM_001204062 | 865    | 16.69 |                      | COMMD3-BMI1_uc009xkg.3_4_2_2              | 98                       | Failed              |
| DNAJC25-GNG10              | NM_001125 | 675      | 29.45 |                      | PB *                                      | 979                      | Failed              |
| URGCP-MRPS24               | NM_001204871 | 644    | 50.25 |                      | PB *                                      | 574                      | Failed              |
| PGK                        | NM_001042616 | 607    | 29.27 |                      | PYURF                                     |                          |                      |
| RBM14-RBM4                 | NM_001198845 | 559    | 22.65 |                      | PB *                                      | 1056                     | Failed              |
| SNRPN                      | NM_022807 | 559      | 20.88 |                      | SNURF                                     |                          |                      |
| RGD6                       | NM_001122363 | 535    | 4.58  |                      | PB *                                      | 315                      | Failed              |
| C7orf55-LUC7L2             | NM_00124584 | 532    | 12.66 |                      | PB *                                      | 1169                     | Failed              |
| SYNJ2BP-COX16              | NM_001202547 | 526    | 17.59 |                      | SYNJ2BP-COX16_uc021w2m.1_1_2_1            | 111                      | Failed              |

*PB: primer pair not available in whole transcriptome qPCR primer database. The primers are automatically designed using NCBI-PrimerBLAST. Please refer to Table S2 in File S1 for details.

**Identical to gene: two genes have identical sequence and thus are non-distinguishable by the mapping algorithm or RT-PCR.

doi:10.1371/journal.pone.0094250.t002
both techniques reaches only $R = 0.52$–$0.66$ [18,19]. The false negative rates of Illumina NGS platform was as high as 12%, much higher than the microarray platforms (0.97%–3.1%) [17]. This is also consistent with other studies [18]. Considering the enormous throughput and dynamic range of NGS, this low correlation and high false negative rates is not likely to be caused by the throughput, but by the data processing. As experimentally shown in this study, mapping to RNA reference sequences especially requires the sensitivity and correctness of the mapping algorithm. Algorithms lacking robustness can result in numerous false positives and false negatives in gene identifications and may affect the gene quantification (Figure 5). Also, FANSe2 showed remarkable better consistency to the microarray data than most other algorithms, bridging the gap between the NGS and microarray and leading to better reproducibility and confidence, which is in great demand for the NGS-based studies (reviewed in [45]).

Furthermore, almost all mapping algorithms offer numerous parameters, and small alteration of parameters may lead to significant change of result. This already leads to low reproducibility and low robustness of many next-generation sequencing studies (reviewed in [45]). In contrast, the parameter settings of FANSe2 almost did not affect the sensitivity and correctness (Figure 4B), providing a remarkable simplicity and robustness of usage.

Previous algorithms require more memory for larger reference genomes. For human genome, 3–14 GB memory is usually required [46]. To reduce the memory consumption when parallelized, some common data, e.g. the reference sequences and the index, need to be shared by multiple CPU cores, increasing the risk of access contention, i.e. simultaneous access of the same data by different CPU cores. This may trigger an unpredictable error or needs additional handling, leading to reduced stability or speed. This problem remains as an open challenge in computational science [47–49]. In contrast, FANSe2’s memory consumption is almost independent of the reference genome size, since it splits the large reference genome into user-specified segments (Figure 3A). Reducing the segment size can significantly reduce the memory demand without impairing the result, facilitating the parallelization especially in normal office computers. The small and user-adjustable memory consumption also allows parallelization of multiple processes instead of threads, since there is no need to share any common data in the memory, thus eliminating the instability caused by access contention.

Importantly, this merit makes FANSe2 the first algorithm with the feature of flexible, scalable and almost maintenance-free parallelization across multiple computers, efficiently utilizing the computational power of inexpensive office computers and even laptop computers. With just three computers, FANSe2 mapped...
608 million reads to human genome within 4.1 hours. This might be the first mapping algorithm that matches the speed of the coming generation of sequencers like Ion Torrent Proton P2 (860 M reads in several hours) running in normal computers. There is no need for expensive, exclusive and maintenance-intensive clusters or workstations.

FANSe2 runs under various operating systems including Windows, providing user-friendly graphical user interfaces, bringing convenience to the biological researchers who are not familiar with computational issues. With simple online video tutorials, everyone knows how to install and use it in 15 minutes. The ability of mapping billions of reads in hours using normal office computers with robust accuracy makes FANSe2 a good candidate to remove the bottleneck of data processing pipeline, leading to much faster, more reliable and quantitative analysis, able to handle the future sequencing applications.

**References**

1. Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. Nat Biotechnol 27: 455–457.
2. Li H, Homer N (2011) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11: 473–483.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcription by RNA-Seq. Nat Methods 5: 621–628.
4. Myers E (1986) An ONTAND difference algorithm and its variations. Algorithmica 1: 251–266.
5. Zhang G, Fedoryn I, Kirchner S, Xiao C, Valleriani A, et al. (2012) FANSe: an accurate algorithm for quantitative mapping of large scale short reads. Nucleic Acids Res 40: e63.
6. Fonscra NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28: 3169–3177.
7. Schbath S, Marinus V, Zytynicki M, Fayolle J, Louis V, et al. (2012) Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. J Comput Biol 19: 796–813.
8. Glisl Ö, Pritchard JK, Thornton K (2009) Characterizing natural variation using next-generation sequencing technologies. Trends Genet 25: 463–471.
9. Makridakis NM, Phipps T, Srivastav S, Reichardt JK (2009) PCR-free method detects high frequency of genomic instability in prostate cancer. Nucleic Acids Res 37: 7441–7446.
10. Ledergerber C, Desaiom C (2011) Base-calling for next-generation sequencing platforms. Brief Bioinform 12: 489–497.
11. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.
12. Ruffalo M, LaFramboise T, Koyuturk M (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 27: 2790–2796.
13. Peng Z, Cheng Y, Tan RC, Kang L, Tian Z, et al. (2012) Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol 30: 253–260.
14. Ieda K, Jin H, Zha JK (2009) Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in Arabidopsis thaliana. BMC Genomics 10: 155.
15. Almazò M, Barillaricono V, Alonso-Lehmann RN, Menendez-Arias L (2013) Altered error specificity of Rnase H-deficient HIV-1 reverse transcriptases during DNA-dependent DNA synthesis. Nucleic Acids Res 41: 4601–4612.
16. Parmar VN, Nelson SF (2009) BEAST: an alignment model for large scale genome resequencing. PLoS One 4: e7677.
17. Willenbrock H, Salmonon J, Sokale R, Barlow KR, Hansen TN, et al. (2009) Quantitative miRNA analysis by microarray sequencing with next-generation sequencing. Rna 15: 2029–2034.
18. Gilt A, Dvinge H, Salmonon Divon M, Osborne M, Kutter C, et al. (2010) Systematic comparison of microarray profiling, high throughput sequencing, and next-generation sequencing technologies for measuring differential microRNA expression. Rna 16: 991–1006.
19. Kelly AD, Hill KE, Carroll M, Hu L, Wang YE, et al. (2013) Next-generation sequencing and microarray-based interrogation of microRNAs from formalin-fixed, paraffin-embedded tissue: preliminary assessment of cross-platform concordance. Genomics 102: 8–14.
20. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.
21. Morgulis A, Gertz EM, Schaffner AA, Agarwala R (2006) WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22: 134–141.
22. Wang T, Cui Y, Jin J, Guo J, Wang G, et al. (2013) Translating mRNAs strongly correlate to proteins in a multicvariate manner and their translation ratios are phenotypic prototype. Nucleic Acids Res.
23. Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics 10: 221.
24. Zeisel A, Vizitskaya A, Bosel Ben-Moshe N, Domany E (2013) An accessible database for mouse and human whole transcriptome 2PRC primers. Bioinformatics 29: 1355–1356.
25. Su Z, Li Z, Chen T, Li Q-Z, Fang H, et al. (2011) Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. Chemical research in toxicology 24: 1489–1495.
26. Izriazy RA, Behdad BM, Collins F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e13.
27. Prasad A, Kumar SS, Dessinoc M, Jaquet V, Bleuler S, et al. (2013) Global regulary architecture of human, mouse and rat tissue transcriptomes. BMC genomics 14: 716.
28. Evrony GD, Cai X, Lee E, Hills LB, Elhorsy PC, et al. (2012) Single-neuron sequencing analysis of Ii retrotransposition and somatic mutation in the human brain. Cell 151: 483–496.
29. Hillier LW, Marsh GT, Quinlan AR, Dooling D, Evelow G, et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. Nat Methods 5: 133–139.
30. Jex AR, Liu S, Young ND, Hall RS, et al. (2011) Ascaris suum draft genome. Nature 479: 529–533.
31. Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, et al. (2013) Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetus griseus draft genome. Nat Biotechnol 31: 759–765.
32. Zhang G, Liu X, Quan Z, Cheng S, Xu X, et al. (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nat Biotechnol 30: 549–554.
33. Li Z, Zhang Z, Yan P, Huang S, Fei Z, et al. (2013) RNA-Seq improves annotation of protein-coding genes in the cucumber genome. BMC Genomics 12: 540.
34. Chiu RW, Chan KC, Gao Y, Lau VY, Zheng W, et al. (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A 105: 20458–20463.
35. Canick JA, Kloza EM, Lambert-Messerlian GM, Haddow JE, Ehrich M, et al. (2012) DNA sequencing of maternal plasma to identify Down syndrome and other trisomies in multiple gestations. Prenat Diagn 32: 730–734.
36. Palomaki GE, Deciu C, Kloza EM, Lambert-Messerlian GM, Haddow JE, et al. (2012) DNA sequencing of maternal plasma to identify Down syndrome and other trisomies in multiple gestations. Prenat Diagn 32: 730–734.
37. Huang X, Liu S, Ouyang J, Huang X, et al. (2014) RNA-Seq analysis of L1 retrotransposition and somatic mutation in the human genome. PLoS One 9: e88487.
38. Szabo Z, Brown GD, Celis M, et al. (2014) Systematic comparison of RNA-Seq and microarray expression profiling of mouse brain. PLoS One 9: e88388.
39. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafina V (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics 27: 1068–1073.

**Supporting Information**

File S1. File S1 includes the following: Table S1. Six datasets downloaded from DDBJ Sequence Read Archive to analyze the error distribution. Table S2. The gene-specific PCR primers for validation of gene identifications. The genes identified solely by Bowtie2 are shaded as gray, and the genes identified solely by FANSe2 are not shaded. Table S3. Mapping programs tested in this study. Table S4. Test parameters for Figure 4. (PDF)

**Acknowledgments**

We thank Dr. Hanqi Yin (Shanghai Biotechnology Corporation) for his help on analyzing the microarray data.

**Author Contributions**

Conceived and designed the experiments: GZ CLX. Performed the experiments: XLL JYZ JJJ. Analyzed the data: GZ CLX ZBM. Wrote the paper: GZ QYH.

**FANSe2: A Robust and Fast Mapping Tool for Next Generation Sequencing**

The ability of mapping billions of reads in hours using normal office computers with robust accuracy makes FANSe2 a good candidate to remove the bottleneck of data processing pipeline, leading to much faster, more reliable and quantitative analysis, able to handle the future sequencing applications.
40. Rowley JW, Oler AJ, Tolley ND, Hunter RN, Low EN, et al. (2011) Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes. Blood 118: e101–e111.
41. McDanel TD, Smith TPL, Harhay GP, Wiedmann RT (2012) Next-Generation Sequencing of the Porcine Skeletal Muscle Transcriptome for Computational Prediction of MicroRNA Gene Targets. PLoS ONE 7: e42039.
42. Ramskold D, Luo S, Wang Y-C, Li R, Deng Q, et al. (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotech 30: 777–782.
43. Xiao W, Tran B, Staadt LM, Schmitz R (2013) High-Throughput RNA Sequencing in B-Cell Lymphomas. Lymphoma: Springer. pp. 295–312.
44. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, et al. (2014) Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res 42: D926–932.
45. Nekrutenko A, Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet 13: 667–672.
46. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967.
47. Chang DY-c, Lawrie D (1982) Performance of multiprocessor systems with space and access contention. Urbana, Ill.: Dept. of Computer Science, University of Illinois at Urbana-Champaign. 80 p.
48. Zhuravlev S, Blagodurov S, Fedorova A (2010) Addressing Shared Resource Contention in Multicore Processors via Scheduling. Asplos Xv: Fifteenth International Conference on Architectural Support for Programming Languages and Operating Systems: 129–141.
49. Chen KY, Chang JM, Hou TW (2011) Multithreading in Java: Performance and Scalability on Multicore Systems. Ieee Transactions on Computers 60: 1521–1534.