OPEN ACCESS Freely available online

PLOS ONE

MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping

Wan-Ping Lee[1]*, Michael P. Stromberg[1], Alistair Ward[1], Chip Stewart[1,2], Erik P. Garrison[1], Gabor T. Marth[1]

1 Department of Biology, Boston College, Chestnut Hill, Massachusetts, United States of America, **2** Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Abstract

MOSAIK is a stable, sensitive and open-source program for mapping second and third-generation sequencing reads to a reference genome. Uniquely among current mapping tools, MOSAIK can align reads generated by all the major sequencing technologies, including Illumina, Applied Biosystems SOLiD, Roche 454, Ion Torrent and Pacific BioSciences SMRT. Indeed, MOSAIK was the only aligner to provide consistent mappings for all the generated data (sequencing technologies, low-coverage and exome) in the 1000 Genomes Project. To provide highly accurate alignments, MOSAIK employs a hash clustering strategy coupled with the Smith-Waterman algorithm. This method is well-suited to capture mismatches as well as short insertions and deletions. To support the growing interest in larger structural variant (SV) discovery, MOSAIK provides explicit support for handling known-sequence SVs, e.g. mobile element insertions (MEIs) as well as generating outputs tailored to aid in SV discovery. All variant discovery benefits from an accurate description of the read placement confidence. To this end, MOSAIK uses a neural-network based training scheme to provide well-calibrated mapping quality scores, demonstrated by a correlation coefficient between MOSAIK assigned and actual mapping qualities greater than 0.98. In order to ensure that studies of any genome are supported, a training pipeline is provided to ensure optimal mapping quality scores for the genome under investigation. MOSAIK is multi-threaded, open source, and incorporated into our command and pipeline launcher system GKNO (http://gkno.me).

Citation: Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, et al. (2014) MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLoS ONE 9(3): e90581. doi:10.1371/journal.pone.0090581

Editor: Chuhsing Kate Hsiao, National Taiwan University, Taiwan

Received November 12, 2013; **Accepted** January 31, 2014; **Published** March 5, 2014

Funding: NIH: 5R01HG004719-04; NIH: 3U01HG006513-02S1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wanping.lee@bc.edu

Introduction

The widespread availability of next-generation sequencing platforms has revolutionized the life sciences through the ever more accessible ultra-high throughput DNA sequencing efforts [1]. Next-generation sequencing technologies including Illumina, Complete Genomics, and Applied Biosystems (AB) SOLiD have been driving the current market forward, whereas Pacific Biosciences SMRT [2], Ion Torrent [3], and Nanopores [4] are leading the development of third-generation sequencing instruments. These technologies bring novel opportunities for many applications including genetic variant discovery, epigenomic variant discovery, RNA-Seq and ChIP-Seq, but also provide complex computational challenges. The short reads generated by these technologies are generally aligned to a reference genome as an early step in many of the current analysis workflows and the alignment quality limits the accuracy of any downstream analysis. Large sequencing projects often use sequencing machines from multiple manufacturers for data generation and can also make use of legacy data. It is desirable that any researcher tasked with analyzing the available data need not learn the intricacies of multiple alignment software packages to utilize all of the available data. This is unnecessary, since, MOSAIK can, uniquely, accurately align sequencing data from all current and legacy platforms.

Current sequencing technologies typically generate on the order of hundreds of millions of short reads (of the order of a few hundred nucleotides or shorter) on a single run. In order to analyze all of these reads in a reasonable amount of computational time, the performance of reference-guided alignment programs is paramount. The memory footprint of these algorithms must also be well managed to allow their deployment beyond institutions with extremely expensive computational infrastructure. These goals must be met without compromising the accuracy of resulting alignments. Most existing aligners utilize hashing algorithms or the Burrows-Wheeler transform [5,6] to search exact matches (algorithms may be modified to allow few mismatches) as their first step to achieve high performance and optimize memory usage. Theoretically, hashing method outperforms BWT method for DNA database searching [7]. The hash-based aligners, Eland (AJ Cox, Illumina, San Diego), MAQ [8], mrFAST/mrsFast [9,10], SHRiMP [11,12], and ZOOM [13,14] hash reads and fit these hashes to the reference genome, while MOM [15], MOSAIK, PASS [16], ProbeMatch [17], SOAP [18], SRmapper [19], and STAMPY [20] hash the reference genome and store this for comparison with reads. Major Burrows-Wheeler transform (BWT) based aligners include BWA [21], Bowtie [22,23], segemehl [24] and SOAP2 [25]. In general, BWT-based aligners are sensitive but include a slow query operation (each FM-index query is slower than a hash query [26,27]). In regions with

genomic variation (e.g. those regions in which the investigator is usually most interested), maintaining good performance generally leads to lower sensitivity [19,28]. In addition, the Burrows-Wheeler transform method is less flexible than hash based methods. For example, it is more difficult for the Burrows-Wheeler transform to consider ambiguities by using IUPAC [29] ambiguity codes representing, for example, known SNPs. The main drawback of hash-based aligners is that they usually consume more memory than BWT-based aligners; however, as high-memory machines become cheaper, this becomes less of a problem. Currently, MOSAIK can be operated in a low-memory mode that keeps the memory footprint small (~8Gb for the human genome), ensuring that even for lower memory machines, MOSAIK can still be used with confidence.

Here, we introduce a reference-guided aligner, MOSAIK, that is highly sensitive, stable and flexible, whose utility on a range of different sequencing technologies has been demonstrated in the context of the 1000 Genomes Project [30,31]. In addition to MOSAIK's ability to map data from all major sequencing technologies, it has been developed to address many of the issues currently facing genome researchers. These developments are outlined here. The primary goal of any mapping software is to minimize alignment artefacts and increase alignment sensitivity and accuracy. To achieve this, MOSAIK uses a Smith-Waterman algorithm and is able to align reads to a genome using IUPAC ambiguity codes, ensuring that alignments against known *single-nucleotide polymorphisms* (SNPs) are not penalized. Using this method, MOSAIK achieves positive predictive values (PPVs) of 99.5% for all alignments and 100.0% for high confidence alignments (those with a mapping qualities larger than 20) in experiments on simulated data. In addition to providing the genomic coordinates of the read mapping, it also important to provide a measure of the confidence in this coordinate. For this purpose, MOSAIK uses a neural-network based training scheme to provide well-calibrated mapping quality scores. In our experiments, the correlation coefficient between the quality scores assigned by MOSAIK and the actual scores is 0.97. To ensure that studies of any genome are supported, MOSAIK provides a training pipeline to ensure optimal mapping quality scores for the genome under investigation. A major area of active investigation is the study of structural variation (SV). MOSAIK has been designed to aid and simplify the discovery of such variants. In particular, known-insertion sequences, for example, mobile element insertions (MEIs), can be included as part of the reference genome. This helps to minimize alignment artefacts, but MOSAIK also provides a host of valuable information to user on the paired-end reads that map to one of these sequences. When requested, MOSAIK also outputs all possible mapping locations for every read in a separate BAM file. This is essential for determining the mappability of the genome under study. The most recent versions of BWA, BOWTIE and MOSAIK are comparable in their run times, and STAMPY is approximately six times slower. Finally, MOSAIK is implemented in C++ as a modular suite of programs that is dual licensed under the GNU General Public License and MIT License. It is multi-threaded, open source, and incorporated into our command and pipeline launcher system GKNO (http://gkno.me).

Results

Alignments from all sequencing technologies

All of the available sequencing technologies use different techniques for library preparation, paired-end read protocols and DNA sequencing, resulting in a range of read lengths, fragment lengths, base quality assignments, as well as different error profiles. Additionally, not all technologies report their sequencing reads in the conventional basespace (strings of the A, G, C and T nucleotides) format. Notably, AB SOLiD uses a di-base encoding scheme known as colorspace and single-molecule sequencing technologies use dark bases [32] for bases not registered by the instrument. These facts mean that all of the currently available aligners are tailored for use on data from one, or a small number of the available technologies. MOSAIK is the only aligner that can be used in a consistent manner across most of these technologies.

In addition to the second-generation technologies, Illumina, Roche 454 and AB SOLiD, MOSAIK can also be deployed on third-generation technologies, in particular, Pacific Biosciences and Ion Torrent reads. MOSAIK uses the same algorithmic approach for all sequencing technologies, however, since the characteristics of each technology are different, the resultant alignment rates vary, as shown in Table 1. These alignment rates were generated using Illumina paired-end (PE), single-end (SE) and Roche 454 SE reads generated using the MASON read simulator (http://www.seqan.de/projects/mason/) as well as Illumina and AB SOLiD reads from the Han Chinese in Beijing (CHB) population from the 1000 Genomes Project. For the third-generation technologies, we used *E. coli* reads provided by Ion Torrent (http://www.iontorrent.com/applications-pgm-accuracy/) and *V. cholerae* reads provided by Pacific Biosciences (ftp://ftp.ncbi.nlm.nih.gov/sra/Submissions/SRA026/SRA026766/provisional/SRX032454/SRR075103/).

In general, sequencing reads containing fewer sequencing errors have higher alignment rates, e.g. Illumina reads, and longer or paired-end reads require more time to align. That paired-end reads take additional time is not unexpected. If one of the reads in a pair cannot be mapped unambiguously, additional searches are performed guided by the mapped mate in the pair. The additional processing time results in more accurate alignments as well as a lower fraction of unaligned reads. AB SOLiD reads are aligned in colorspace (converting to basespace prior to alignment loses all of the benefits of colorspace), but additional processing is required due to the required conversion of the alignments into basespace post-alignment. These experiments show that MOSAIK works well for existing sequencing technologies.

Highly accurate alignments on simulated data

To investigate the accuracy of reads aligned using MOSAIK, we simulated a total of 12 million Illumina paired-end reads from chromosome 20 of the Hg19 human genome using the MASON read simulator. Reads of length 76 and 100 basepairs were simulated with a haplotype SNP rate of 0.1%. The reads were aligned against the entire human genome using BWA-0.5.9, BOWTIE-2.0-beta5, STAMPY-1.0.13, and MOSAIK-2.1.78. The default parameter settings were used for all of the aligners. The positive predictive value of each aligner was then calculated as the number of correctly placed reads (the genomic coordinate of the mapped read agreed with the known location of the read from MASON) divided by the total number of mapped reads. Notice that an alignment is considered incorrect as the aligned position is out the 20 bp tolerant window and thus alignments with more than 20 bp unmapped bases may be considered as incorrect. We choose 20 bp as the tolerant window since on the dataset most of alignments contain fewer than 20 bp clipped bases (see supplemental Figure S1).

Figure 1 shows the positive predictive value (PPV, the number of correctly mapped reads divided by the total number of mapped reads) of the aligners as a function of mapping quality cutoffs (complete information is shown in Figure S2). At a mapping quality cutoff of twenty, for example, the PPV is calculated using

Table 1. Summary of the alignment accuracies achieved by MOSAIK for reads generated from different sequencing technologies.

Technologies	Aligned (%)	Speed (reads/second)	Read lengths [min;max]	Reference genome	Dataset
Illumina; PE	99.98	83.95	100; 50	Human hg19	MASON simulated
Illumina; SE	99.75	153.98	100; 76; 50	Human hg19	MASON simulated
Illumina; PE/SE	91.48	147.42	81; 76; 51; 45; 41	Human hg19	CHB population in 1000G
454; SE	99.42	8.018	400.673 [266;529]	Human hg19	MASON simulated
Ion Torrent	77.02	20.85	223.99 [59;398]	*E. coli* strain 536	Ion Torrent released
SOLiD	55.64	126.81	50	Human hg19	CHB population in 1000G
Pacific Biosciences*	85.79	0.69	698.61 [48;6084]	*V. cholerae* 4,033,464 bp.	Pacific Biosciences released

*The parameter set "-hs 10 -mmp 0.5 -act 15" was used as opposed to the default values "-hs 15 -mmp 0.15 -act 55".
With the exception of the Pacific Biosciences data, all alignments were generated using MOSAIK's default parameters.
doi:10.1371/journal.pone.0090581.t001

only those reads with mapping quality values greater than or equal to twenty. It can be seen that the PPVs of BWA and MOSAIK are comparable and are significantly better than those achieved by BOWTIE and STAMPY. For mapping quality cutoffs smaller than five, BWA is more accurate (fewer incorrect alignments among total mapped alignments) than MOSAIK, however, MOSAIK is the most accurate as the mapping quality cutoff is increased. For a mapping quality cutoff of twenty (a common cutoff employed by downstream analysis tools that only wish to consider confidently aligned reads), the PPVs of MOSAIK, BWA, BOWTIE and STAMPY are 100.00%, 99.99%, 99.79% and 99.63% respectively. These results are summarized in Table 2.

Figure 2 shows receiver operating characteristic (ROC) curves for the same data. The total number of mapped reads (x axis) is plotted against the number of incorrectly mapped reads (y axis). Each point on the curve represents the number of alignments whose mapping qualities are greater than or equal to the indicated value. MOSAIK has a relatively smooth curve, ensuring that downstream tools that employ mapping quality cutoffs (i.e. ignoring all reads with mapping qualities less than the cutoff) do not incur extremely large changes in the number of reads while progressively increasing the cutoff. Conversely, the other aligners do not share this property. For example, consider the BWA alignments. By decreasing the mapping quality cutoff from 30 to 29, the number of incorrectly mapped reads increases by 308.56% while for MOSAIK, the increase is a much more modest 6.25%. Downstream analysis tools require a useful mapping quality scale, so that excluding lower quality reads improves the specificity of the analysis results. The dynamic range demonstrated by MOSAIK is therefore a very valuable result for these tools.

Mapping quality calibration

The Phred mapping quality score present in the standard SAM/BAM format represents the probability that the read was mapped incorrectly and is defined as:

$$Q = -10\log_{10}P, \qquad \text{(Equation 1)}$$

where Q is the Phred score and P is the probability that the read was misaligned. For example, a read assigned a Phred mapping quality score of 30 has a 1 in 1000 chance of being misaligned.

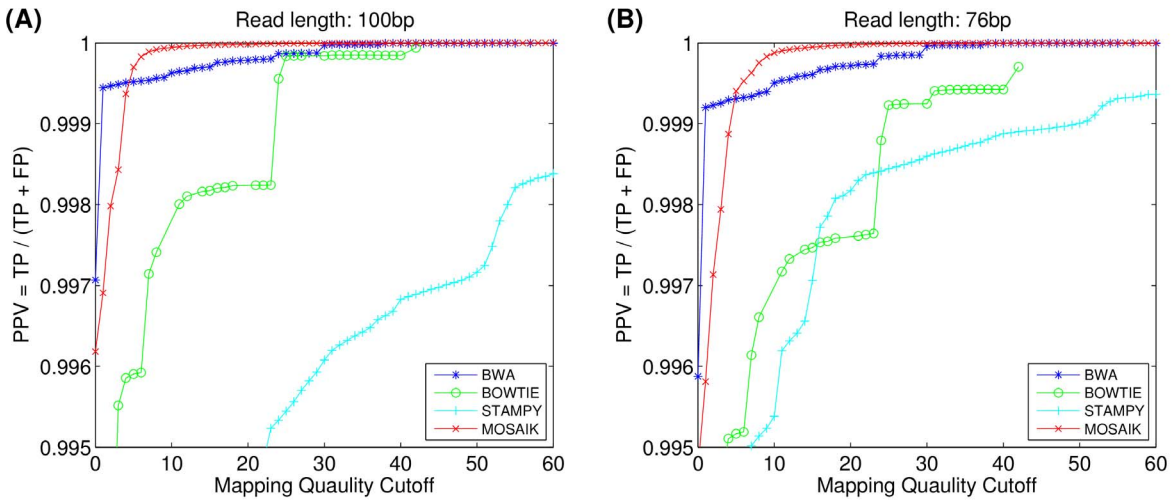

Figure 1. The positive predictive value of aligners (the number of correctly mapped reads divided by the total number of mapped reads) as a function of mapping quality threshold. Datasets in (A) 100 bp and (B) 76 bp read lengths. PPV, TP, and FP stand for positive predictive value, true positive and false positive, respectively.
doi:10.1371/journal.pone.0090581.g001

Table 2. The positive predictive values (the number of correctly mapped reads divided by the total number of mapped reads) in terms of mapping quality cutoffs.

MQ cutoffs	30		20		10		0	
Read lengths	**100**	**76**	**100**	**76**	**100**	**76**	**100**	**76**
BWA	**1**	**1**	0.9998	0.9997	0.9996	0.9995	**0.9971**	**0.9959**
BOWTIE	0.9998	0.9992	0.9982	0.9976	0.9980	0.9972	0.9823	0.9819
STAMPY	0.9961	0.9986	0.9945	0.9982	0.9897	0.9954	0.9813	0.9909
MOSAIK	**1**	**1**	**1**	**1**	**0.9999**	**0.9999**	0.9962	0.9947

doi:10.1371/journal.pone.0090581.t002

MOSAIK's mapping qualities are obtained using a neural network that approximates the error function when provided with features such as best and second best Smith-Waterman alignment scores, read entropy, number of potential mapping locations and hashes. For paired-end reads, the fragment length of mapped paired end reads is also used in the neural network to produce more precise mapping quality calculations. MOSAIK embeds the Fast Artificial Neural Network (FANN) library (http://leenissen.dk/fann/wp/), which implements multilayer artificial neural networks in C, supporting both fully connected and sparsely connected networks, to calculate Phred score for each alignment.

The default neural network provided with MOSAIK was generated by training on the human genome. The first step involves simulating reads and then aligning them to the human reference genome to obtain MOSAIK's behaviour such as best and second best Smith-Waterman scores, and numbers of obtained mappings and hashes. Then, the neural network was trained based on MOSAIK's behaviour.

Figures 3(A) and 3(B) compare the actual (calculated using Equation (1)) and the assigned mapping quality scores. Both, BOWTIE and MOSAIK produce very accurate Phred score mapping qualities across the whole quality score spectrum. The Pearson correlation coefficients between the assigned and actual quality scores are shown in Table 3. MOSAIK has an average (across all read lengths investigated) correlation coefficient of 0.9698, compared with 0.9061, 0.9207, and 0.8652 for BWA, BOWTIE, and STAMPY respectively.

Retraining Mapping-Quality Neural Network for *E. coli* Alignment

The genomes of different species differ in many respects including sequence content (base composition as well as relative frequency of repeat or low-complexity sequence) as well as the size of the genome. Most aligners, including MOSAIK, are general programs that can operate on any given reference genome, however, in general, the properties of the genome under investigation are ignored. MOSAIK provides a retrainable mapping-quality pipeline to generate applicable neural networks for different genomes or sequencing technologies. This means that the calibration of the mapping quality scores remains of a very high quality, regardless of reference genomes.

To demonstrate the merit of the retrainable mapping-quality pipeline, we used 6 million simulated paired-end reads from the *E. coli* genome to train a neural network (see supplemental method (A): Retraining Mapping Quality Neural Network). An additional independent set of 6 million simulated *E. coli* paired-end reads were then generated and aligned to the *E. coli* genome using multiple aligners. The assigned and actual mapping quality scores are plotted for all aligners in Figure 3(C). There are two sets of

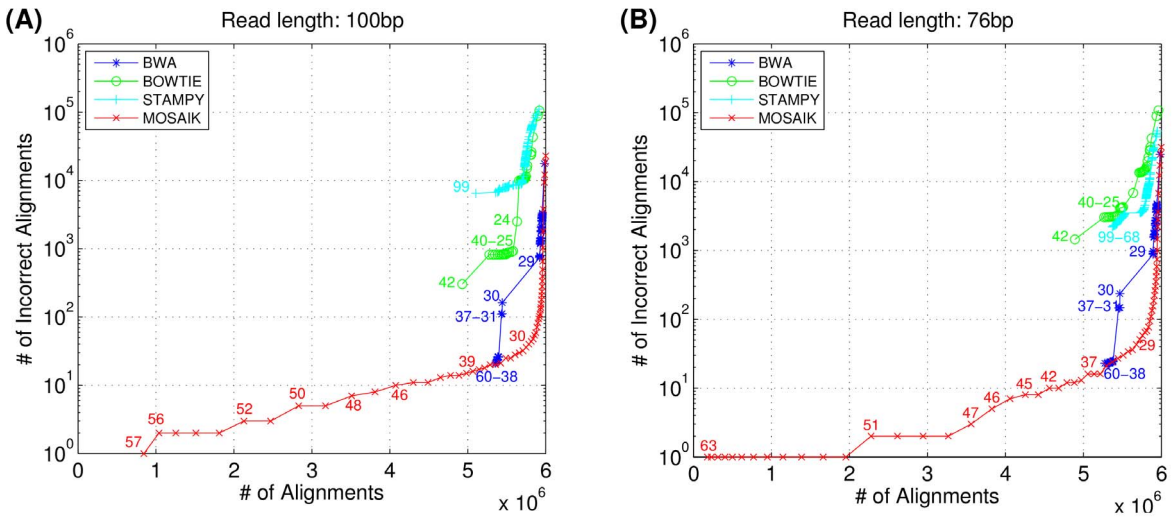

Figure 2. The receiver operating characteristic (ROC) curves; Datasets in (A) 100 bp and (B) 76 bp read lengths. Each point represents the total numbers of alignments whose mapping qualities are greater than the indicated value. MOSAIK has a relatively smooth curve, ensuring that downstream tools that employ mapping quality cutoffs (i.e. ignoring all reads with mapping qualities less than the cutoff) do not incur extremely large changes in the number of reads while progressively increasing the cutoff.
doi:10.1371/journal.pone.0090581.g002

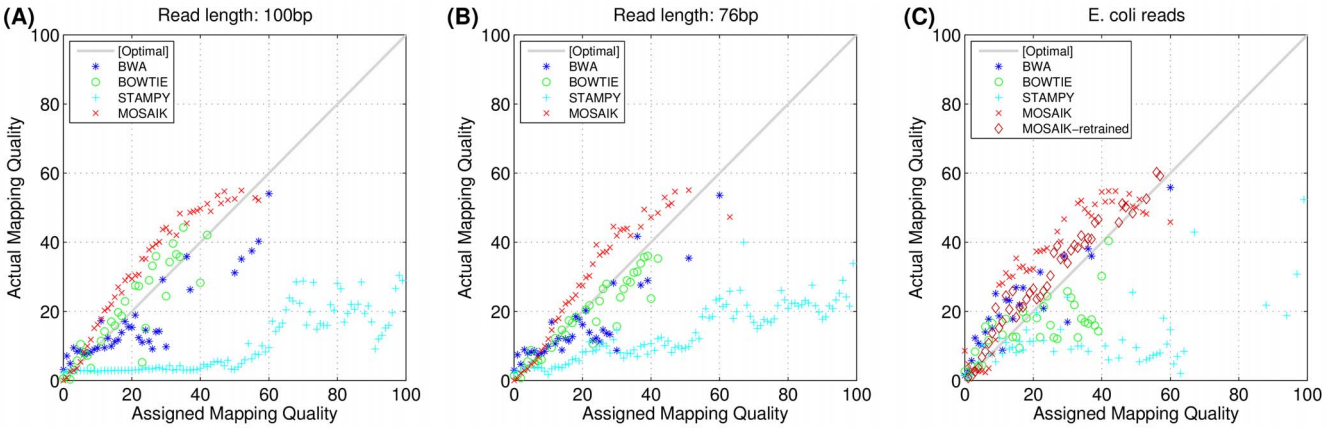

Figure 3. The correlations between the aligners' assigned and actual mapping qualities. Phred score scheme. (A) and (B) simulated datasets in 100 bp and 76 bp read lengths. (C) *E. coli* simulated dataset in which "MOSAIK" is MOSAIK's default mapping-quality network trained by human genome while "MOSAIK-retrained" is the retrained mapping-quality network by using *E. coli* simulation and *E. coli* genome. The detailed numbers of the Pearson's correlation coefficients are given in Table 3.
doi:10.1371/journal.pone.0090581.g003

data for MOSAIK: the first (red crosses) is generated using the default neural network trained on the human genome, and the second (dark red diamond) uses the neural network retrained on the *E. coli* genome. It is clear that the mapping qualities generated by the retrained neural network for MOSAIK are the best calibrated, although the data using the human genome trained neural network is still of a high quality. Also of note, Figures 3(A) and 3(B) show that BOWTIE has quite well calibrated mapping qualities for mapping to the human genome, however, when applied to *E. coli*, the calibration is noticeably worse.

MOSAIK accurately accounts for short INDELs

MOSAIK uses a Smith-Waterman (SW) algorithm as the final polishing step to produce pairwise read alignments, which is the preferred choice for aligning gapped (short INDELs) sequences since it seeks all possible frames of alignment with all possible gaps. To assess the sensitivity of different aligners to short INDELs, we simulated Illumina paired-end reads containing 1-14 bp INDEL events that are generated by a genome simulator, MUTATRIX (https://github.com/ekg/mutatrix). For each INDEL length, we introduced an average of 100 events, with approximately 800 spanning reads (see supplemental Figure S4).

Figures 4(A) and 4(B) plot the sensitivity (number of correctly mapped reads divided by the total number of simulated reads) as a function of the INDEL length. An alignment is considered correct when it is mapped to the correct position as well as contains the simulated variant. Alignments containing the correct variants can facilitate downstream variant detectors detecting variants depending on alignments and need no any realignment step which is timing consuming. MOSAIK is the most sensitive aligner considered here when considering deletions. When considering insertions, MOSAIK's sensitivity is comparable to, but slightly worse than those of STAMPY and BOWTIE. It is clear from Figures 4(A) and 4(B) that MOSAIK is the only mapper considered here that is highly sensitive to both insertion and deletion polymorphisms. We understand that some aligners tend to report partial alignments that may not contain variants but are mapped to right places. Those alignments still provide values for variant detections. We thus change the criteria of correct alignments used in Figures 4(A) and 4(B). In figures 4(C) and 4(D), an alignment is considered a correct mapping when it is entirely or partially mapped to the correct positions. The four aligners achieve 96% sensitivity based on the criteria.

Table 3. Pearson's correlation coefficients of mapping qualities.

Read lengths	100	76	*E. coli*
BWA	0.8987	0.8625	0.8936
BOWTIE	0.9027	0.9449	0.6989
STAMPY	0.8317	0.8818	0.5262
MOSAIK	**0.9609**	**0.9497**	0.8881
MOSAIK-retrained	–	–	**0.9749**

doi:10.1371/journal.pone.0090581.t003

Effect of mapping errors on SNP studies

Aligners provide information on where reads map in the human genome along with information on the confidence of the mapping, however, they do not themselves weigh evidence for genetic variants in the genome being studied. Dedicated variant callers use the information provided by mapper in statistical models to determine if there is enough evidence to report a difference with respect to the reference genome. To determine the effect of the mapping on single nucleotide polymorphism (SNP) discovery, we simulated 1,486 SNPs on the human genome chromosome 20 using MUTATRIX. We then used MASON to generate 12 million reads (with read lengths of 76 and 100 basepairs) from this mutated chromosome generated by MUTATRIX. The same four aligners were then used to align these reads back to the entire human reference genome and the variant callers FREEBAYES [33] and SAMTOOLS [34] were used to call SNPs. Figure 5 shows the variant callers sensitivity to SNPs as a function of the false discovery rate (FDR) (the complete information is shown in Figure S3). The points on the curves are generated by only considering SNP calls with variant quality scores (provided by the variant caller) greater than a specific cutoff. Moving from lower-left to upper-right, SNP calls with lower quality scores are cumulatively being included. Both FREEBAYES and SAMTOOLS produce lower sensitivity calls on the BOWTIE alignments and have a lower FDR on BWA and MOSAIK alignments. It is clear from both Figures 5(A) and 5(B) that the

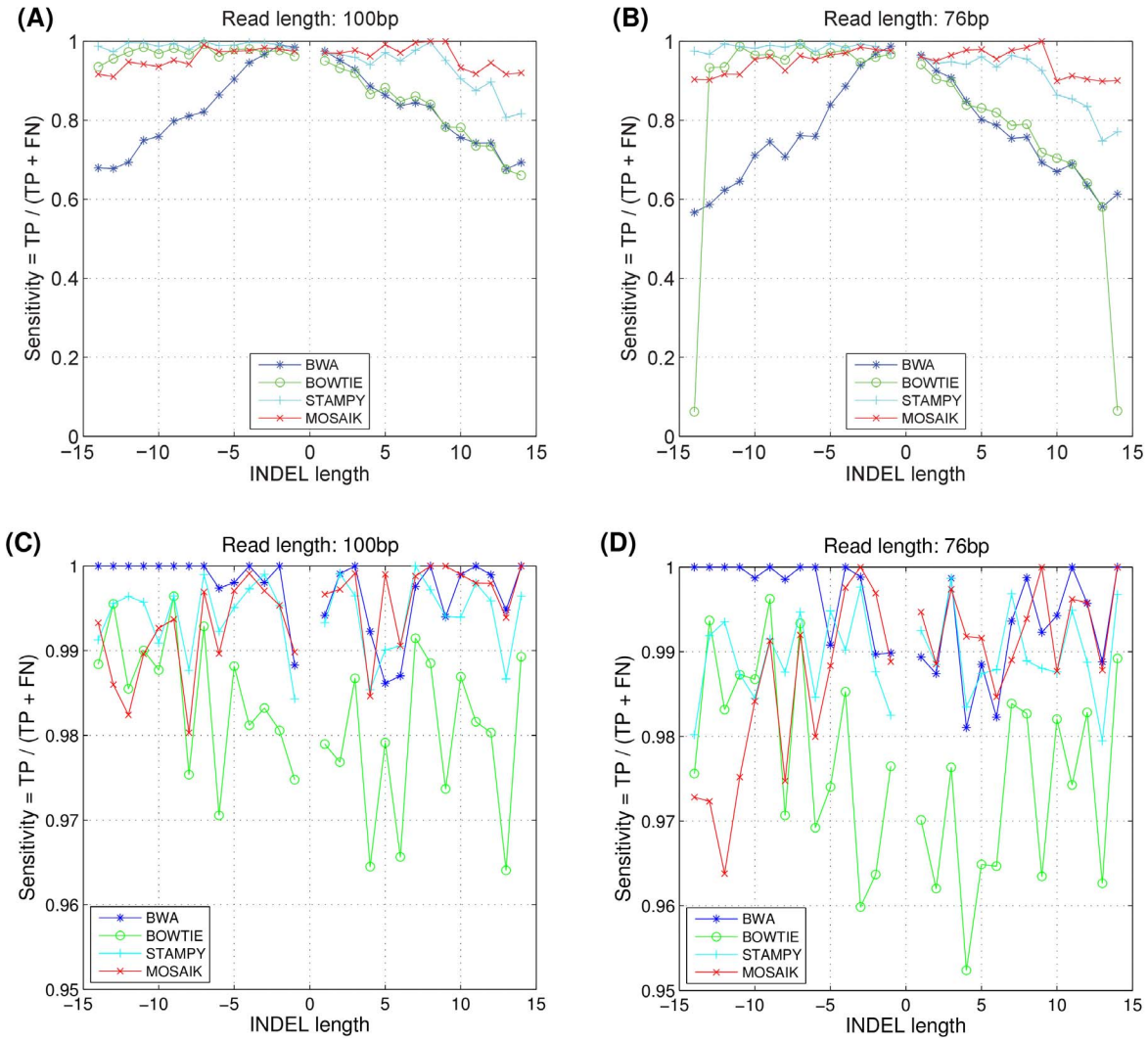

Figure 4. The sensitivities of simulated reads spanning INDELs, which is defined as the number of correct mapped reads divided by the number of simulated reads for each INDEL length. In (A) and (B), the alignments are considered correct as they cross INDELs, while in (C) and (D), the alignments are considered correct as they are entirely or partially mapped to the correct positions. TP and FN are "true positive" and "false negative" respectively.
doi:10.1371/journal.pone.0090581.g004

most sensitive SNP calls are produced when using the MOSAIK alignments, although the BWA alignments are also of a high quality. It is also worth noting that the SNP calls produced by FREEBAYES are more sensitive than those produced by SAMTOOLS regardless of the mapper used.

Support for mobile element insertion

Detecting structural variations using NGS data is a more complex task than detection of short variants and often requires or would benefit from information over and above that ordinarily required for small variant detection. An increasing number of SV detection algorithms are being developed and, in order to increase the effectiveness of these algorithms, MOSAIK has been developed to provide as much relevant and useful information as possible.

There are many genetic sequences that can be considered distinct from the standard set of chromosomes in the genome under investigation. These can include repetitive sequences such as mobile elements [35], viruses (e.g. human endogenous retroviruses [36]), known novel insertions [37,38] or bacterial contaminants [39] amongst others. MOSAIK provides support for an additional reference genome file containing any genetic sequences provided by the investigator. The advantages of this are two-fold: a) reads originating from contaminants will map to the additional sequences, rather than a lower quality mapping to the best location in the standard reference genome. These sequences essentially act as a sink to catch all the reads that do not originate from the standard reference, reducing the number of mismapped reads that variant detectors have to contend with. b) Reads mapping to repetitive elements (e.g. ALU or LINE elements) are identified as mapping to the additional reference sequence. MOSAIK reports the coordinates of the best mapping in reference genome coordinates, but also includes an additional tag in the BAM file (appearing as ZA in the BAM file), indicating that the read maps to one of the additional reference sequences. Our MEI detector, TANGRAM (https://github.com/jiantao/Tangram) looks for read pairs with one mate uniquely aligned to the genome and the other mate falls within a mobile element

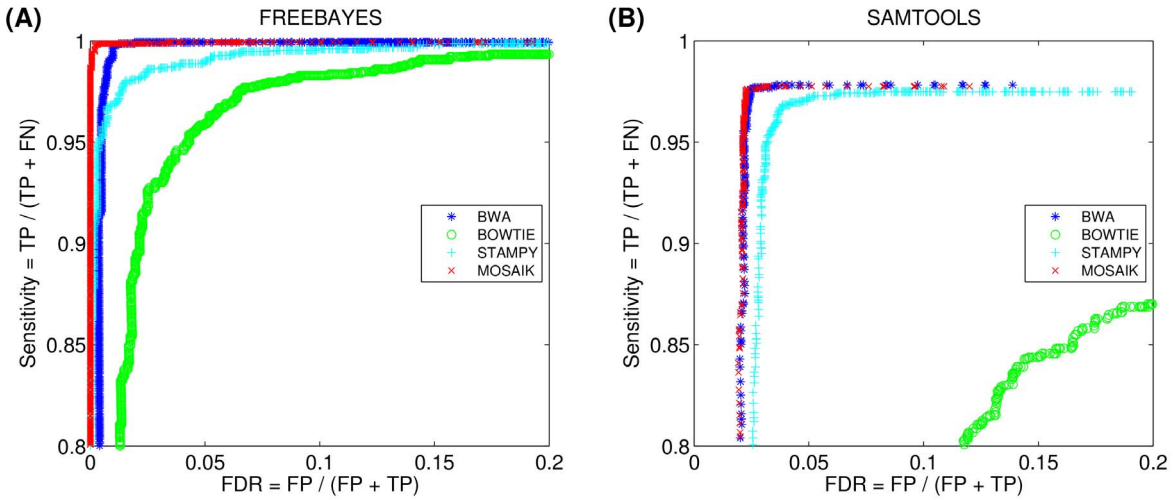

Figure 5. The receiver operating characteristic (ROC) curves of SNPs called by FREEBAYES and SAMTOOLS. The points on the curves are sorted by called qualities and the points closer to the upper-right corner have higher called qualities. The true positive (TP), false positive (FP), and false negative (FN) are calculated by intersecting SNPs called on each aligner's alignments and gold SNPs called on the simulated alignments.
doi:10.1371/journal.pone.0090581.g005

reference sequence. Only relying on this information provided by MOSAIK, the sensitivity of MEI detection can achieve 84%.

Applications

SNP and INDEL Analyses in the 1000 Genomes Project. The 1000 Genomes Project is in the process of using second-generation sequencing instruments to study human genetic variation at the population level. The Phase I [31] release, based on a population of 1,092 sample individuals in 14 populations includes approximately 38 million SNPs, 1.4 million bi-allelic INDELs and 14,000 large deletions. These calls were generated from approximately 966 billion reads and 64 trillion base pairs of human DNA and were sequenced using Illumina, AB SOLiD and Roche 454 for both low-coverage whole-genome and exome targeted sequencing data. A collaborative effort between Boston College and the National Center for Biotechnology Information (NCBI) used MOSAIK to align all of the reads from all of these machines, and served as the official primary alignment set for the exome sequencing data [40] and an alternative alignment set for the low-coverage.

Based on the MOSAIK alignments, SNP, MNP (multi-nucleotide polymorphism) and INDEL calls were generated using the FREEBAYES Bayesian variant calling software. 33,324,407 SNPs were detected in the autosomes of the 1,092 samples, of which, only 23.8% were previously known sites (contained in dbSNP). The transition/transversion (ts/tv) ratio for these sites was 2.12 (2.1 for novel sites and 2.17 for known sites). The Illumina exome data yielded 344,781 SNPs with a ts/tv ratio of 3.18 (3.09 for the novel sites and 3.52 for the known sites) and 22.1% of the exome sites were previously known. The SOLiD exome data yielded 176,637 SNPs with a ts/tv ratio of 3.34 (3.22 for novel sites and 3.58 for known sites). The ts/tv ratios are in accordance with expectations for both the low-coverage and the exome SNPs.

Other SNP Studies. In addition to the 1000 Genomes Project, MOSAIK is widely used for other human clinical genome studies, such as human cancer studies [41–46]. MOSAIK is also used for other species genome studies including model species [47,48], HIV [49–52], parasites [53–55], plants [56–58], and other animals [59,60].

Human Mobile Element Insertion Discovery

In addition to short variants, the 1000 Genomes Project aims to characterize larger structural variations present in the human population. By augmenting the reference genome with known mobile element insertions (MEI), the MOSAIK alignments were able to provide a host of information about their distribution in the human population. As part of the pilot phase of the project, 7,380 MEI polymorphisms were detected using the whole-genome sequencing data [61]. This sample set included 60 samples of European origin (CEU), 59 African (YRI) and 60 Asian samples from Japan and China (CHB/JPT). The FDRs for Alu, L1, and SVA insertions were 2%, 17%, and 27% respectively.

Discussion

MOSAIK is a highly sensitive, stable and flexible reference-guided read mapper which supports most existing sequencing technologies. While MOSAIK is extremely accurate (positive predictive values achieve 99.5% for all alignments and 100.0% for alignments whose mapping qualities are larger than 20 on simulated data), not all reads are aligned with equal confidence. The mapping qualities that MOSAIK provides are generated using a retrainable neural network and are a very good representation of the probability of the alignment being incorrect. In fact, the correlation coefficient between MOSAIK assigned and the actual mapping qualities is 0.97. The retraining pipeline ensures optimized mapping quality score schemes for any genome being studied. For example, when considering aligning against the *E. coli* genome, the correlation coefficient increases from 0.89 to 0.97 when using the human and the *E. coli* neural nets respectively. By using the Smith-Waterman algorithm, MOSAIK is very effective at mapping reads containing short INDELs and the experiments demonstrate that the sensitivity of INDEL mappings is greater than 90%. Additionally, MOSAIK provides explicit support for SV detections.

Most SV detectors make extensive use of information from paired end reads [62–64]. If the two mates in a pair map to greatly separated locations (often the case when the read pair spans or falls within a structural variant), multiple searches through the BAM files are required to assemble all of the information about both

mates. This can be a lengthy task, severely impacting the performance of SV detectors. The ZA tag provides a host of information about the reads mate, including the location, mapping quality, number of mappings for the mate, which ensures that these searches are not required, created vast increases in the efficiency of the SV detectors using this information.

The other utility for SV detections is reporting all possible mappings. Many genomes contain regions that are considered unmappable, usually due to the presence of low complexity DNA. Depending on the algorithms employed, NGS reads can still map to these regions; however, it is often prudent to omit these reads from variant detection. Instead of discarding reads mapping to multiple locations or picking the best quality alignment, MOSAIK records all locations to which a read maps (given the constraints imposed by the selected parameters) and records them in a separate BAM file. Since the number reads mapping to multiple locations as well as the number of entries for each multiply aligned read can be extremely large, the resulting BAM file has the potential to be excessively large. By default, MOSAIK omits much of the read specific information (e.g. read name, sequence and error information), allowing for effective compression of the file after positional sorting, resulting in very small BAM files. The information contained in these BAM files allows easy identification of genomic regions where many individual reads are aligning. These regions are those that can be considered unmappable, since reads hitting these regions are also able to align to other genomic regions. Thus they provide a guide to the mappable genome which can greatly aid in variant discovery.

The default parameters used by MOSAIK were optimized using simulated Illumina datasets from the human genome. They were generated to provide a balance between mismatches and gaps in the alignments, leading to balanced calling of SNPs and INDELs by variant callers. For the experimenter only interested in a specific variant type, it is possible to modify the parameters to provide alignments more sensitive for the variant type of interest. For example, if INDEL discovery is paramount, reducing the Smith-Waterman penalty for the creation and extension of gaps in alignments will lead to a greater likelihood that INDELs will be discovered.

MOSAIKs memory footprint depends on the size of the reference hash-table which, in turn, depends on the hash (*k*-mer) size as well as the length of the reference sequence. For the human genome using the default value of $k = 15$, MOSAIK requires approximately 20Gb of memory. For machines with less available RAM, MOSAIK can be run in a low-memory mode that performs alignments chromosome by chromosome. This reduces the required memory to 7Gb, which makes MOSAIK accessible to most machines.

Improvements in the computational performance can be achieved at the expense of decreased sensitivity, but ongoing development (including replacing the traditional Smith-Waterman algorithm with a *single-instruction-multiple-data* (SIMD) Smith-Waterman algorithm [65,66]) provides significant performance improvements. Initial testing of the SIMD Smith-Waterman algorithm demonstrate a twofold speed up [65]. Further improvement is achieved by reducing the number of applications of the Smith-Waterman algorithm for each read. Reads that originate in highly repetitive sequence can produce tens of thousands of candidate loci (see supplemental Figure S6) in the genome and the Smith-Waterman algorithm is applied to each one of these regions. This is extremely computationally intensive with very little benefit to the alignment sensitivity. As a result, if there are greater than a preset number (the default is 200) of potential mapping loci, MOSAIK only invokes the Smith-Waterman algorithm on the top 200 loci. MOSAIK then reports the most confident alignment from all of the regions in which the Smith-Waterman algorithm was applied. Supplemental Table S1 demonstrates that MOSAIK 2.2.3 (with these modifications) is of the order of five times faster than version 2.1.78 (without the modifications). Importantly, these modifications do not adversely impact the sensitivity of MOSAIK.

Methods

Overview

MOSAIK is a hash-based aligner and it hashes reference sequences as its first step. MOSAIK splits the reference sequences into overlapping contiguous *k*-mers (hashes) and stores the positions of each hash in a hash table data structure that guarantees *O(1)* lookups. Then, MOSAIK hashes each read in the same hash size and looks hashes up in the hash table to obtain the genomic positions of the hashes of a read. Next, nearby hash positions are consolidated as a hash region (hashes of a read may be clustered as several hash regions) where a Smith-Waterman algorithm is applied to align the read to the local region of a genome reference sequence as a final "polishing" step. For paired-end reads, each end-mate of a read is mapped separately. For some cases, that may be one end-mate aligned well and the other one failing to be aligned. The well-aligned mate can be used to try and rescue the unaligned mate using knowledge of the approximate fragment length used in the paired-end read generation.

Processing Reference Sequences

MOSAIK can handle a nearly unlimited number of reference sequences, however, the maximum aggregated reference length is four billion bases. Alignments to the human transcriptome using more than 95,000 individual reference sequences are easily handled. The available hash sizes are 4–32.

MOSAIK supports the full set of IUPAC ambiguous nucleotide characters. This allows users to use reference sequences that have been masked by confirmed dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) calls. The ambiguity codes minimize the alignment bias that might be caused when aligning to reference sequences containing SNPs. For considering IUPAC, MOSAIK substitutes ambiguous codes with all of the alternative bases represented by the ambiguity code and stores the resulting hashes in the hash table. In order to avoid increasing the size of the jump database dramatically, the ambiguity codes N and X are not considered when hashing the reference sequences.

Clustering Hashes

MOSAIK supports various read formats (SRF, FASTA, FASTQ, Bustard, and Gerald). In each case, the reads are split into a set of overlapping hashes and the genomic positions of each hash are queried from the stored reference hash table. A modified AVL tree [67] is employed to handle and cluster nearby hash positions to form a hash region. The clustering algorithm considers sequencing errors, SNPs and single base INDELs. For example, consider a 35 bases read split into hashes of 15 bases. The first hash consists of the first 15 bases in the read. The second hash consists of bases 2–16 in the read and so on. The read consists of 22 individual hashes, each of which is associated with positions within the reference genome. If the read can be aligned perfectly to the somewhere in the reference genome (i.e. there are no sequencing errors or variations), each of the 22 hashes will have a reference genome position offset by a single base (i.e. if the first hash in the read is associated with the reference position *x*, the second hash with the reference position *x+1* etc.). The AVL tree will consolidate those hits into a single alignment candidate region

(see Supplemental Figure S5(A)). The presence of a single sequencing error will ensure that 15 of the hashes (each hash overlapping the error), will not be associated with the correct genomic coordinate. Since the clustering algorithm considers sequencing errors, however, an alignment candidate region is still present in the AVL tree (see Supplemental Figure S5(B)).

Applying Smith-Waterman Algorithm

After identifying alignment candidate regions, MOSAIK employs a Smith-Waterman algorithm to align reads to the alignment candidate regions. The Smith-Waterman algorithm, which was invented over 30 years ago, is still regarded as the most accurate pairwise alignment algorithm and the preferred choice for aligning gapped sequences since it seeks all possible frames of alignment with all possible gaps. Specifically, the alignments are performed using the Smith-Waterman-Gotoh alignment algorithm [68,69].

The time complexity of the Smith-Waterman algorithm is $O(n^2)$, which may render the mapper useless due to poor performance. To address this, a banded Smith-Waterman algorithm [70] has been implemented to improve the performance. According to our experiments, the runtimes for aligning Illumina and Roche 454 data are reduced by approximately 3× and 8× respectively. The further development of using SIMD SW promises significant performance improvements.

Rescuing Paired-End Mates

Each mate in a paired-end read is initially aligned individually. There are various factors that lead to some reads failing to be aligned to the reference. In the case of paired-end reads, the aligned mate can be used to try and rescue the unaligned mate using knowledge of the approximate fragment length used in the paired-end read generation. A local alignment search algorithm has been implemented which performs a Smith-Waterman algorithm in the region proximal to the aligned mate. If the read exhibits the expected strand, orientation, and fragment length, the read is considered rescued. Even if both mates in the pair are successfully aligned, the local alignment search may still be triggered, if the alignments are inconsistent with the expected fragment length.

The number of mates rescued by the local alignment search depends largely on the read lengths considered. With increasing read length, the aligner is less likely to miss a potential alignment and therefore fewer alignments are rescued.

Handling AB SOLiD reads

AB SOLiD reads are represented in the colorspace rather than in the more conventional basespace. Most downstream applications do not support colorspace and thus alignments require conversion to basespace for maximum utility. MOSAIK is equipped to align colorspace reads against a colorspace reference and then convert the resulting alignments into basespace. The di-base quality conversion algorithm uses the minimum of the two qualities that overlap a nucleotide in basespace. This approach allows users to specify parameters, such as the maximum number of mismatches. Additionally, it enables users to merge aligned SOLiD datasets with datasets from other sequencing technologies.

Known-Sequence Insertion Detections

MOSAIK is aware of user-specified insertion sequences, e.g. mobile element insertions. When the insertion sequences are provided, the reference hashes are prioritized such that alignment to the given insertion sequences are attempted prior to alignment to the genome reference. An additional tag in the BAM file (the ZA tag) then indicates any alignments of a read hitting the given insertion sequences. Since MEIs are repetitive elements, a read from an MEI can be mapped to several locations within the genome (potentially hundreds of locations). The ZA tag then populated with valuable information about the reads mate, including location, mapping quality and number of mapping locations for the mate. This information ensures that multiple BAM search operations (which can be lengthy for large BAM files) can be avoided. The downstream MEI detector can detect MEI by using ZA tag easily.

Supporting Information

Figure S1 The distributions of alignments' softclips.
(TIF)

Figure S2 The complete information of Figure 1. The positive predictive value of aligners (the number of correctly mapped reads divided by the total number of mapped reads) as a function of mapping quality threshold. Datasets in (A) 100 bp and (B) 76 bp read lengths. PPV, TP, and FP stand for positive predictive value, true positive, and false positive, respectively.
(TIF)

Figure S3 The complete information of Figure 5. The receiver operating characteristic (ROC) curves of SNPs called by FREEBAYES and SAMTOOLS. The points on the curves are sorted by called qualities and the points closer to the upper-right corner have higher called qualities. The true positive (TP), false positive (FP), and false negative (FN) are calculated by intersecting SNPs called on each aligner's alignments and gold SNPs called on the simulated alignments.
(TIF)

Figure S4 The short INDELs that are inserted for investigating the aligners' abilities for them, and the read coverage for each length INDEL.
(TIF)

Figure S5 MOSAIK hash clustering. (A) The read uniquely aligns perfectly to the references, all hashes will succeed in finding the adjacent reference locations and the AVL tree will consolidate those hashes into one alignment candidate region. (B) However, if only one hash succeeds in finding the proper reference location because of sequencing errors, an alignment candidate region is still present in the AVL tree.
(TIF)

Figure S6 The distribution of candidate loci in the genome of reads. MOSAIK applies a Smith-Waterman algorithm to each candidate locus of a read to generate an alignment. Therefore, the number of candidate loci is equal to the number of executed the Smith-Waterman algorithm. The mhp of MOSAIK is the maximum number of investigated hash positions per 15-mer.
(TIF)

Method S1 The methods of (A) Retraining Mapping Quality Neural Network and (B) Detecting Specified Insertion Sequences.
(PDF)

Table S1 The runtime of each mapper for aligning six million 100 bp reads. The version without '*' are the exact version of each mapper for which we report performance comparisons. For up to date information, we also report speed for the current version (indicated by '*') of each software.

STAMPY is a single-threaded program and thus the runtime of using 4 cpus is not available.
(PDF)

Author Contributions

Conceived and designed the experiments: WL AW EG GM. Performed the experiments: WL. Analyzed the data: WL AW. Contributed reagents/materials/analysis tools: MS CS EG. Wrote the paper: WL AW. Started the project: MS.

References

1. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327: 78–81. doi:10.1126/science.1181498.
2. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138. doi:10.1126/science.1162986.
3. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al. (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475: 348–352. doi:10.1038/nature10242.
4. Schneider GF, Dekker C (2012) DNA sequencing with nanopores. Nat Biotechnol 30: 326–328. doi:10.1038/nbt.2181.
5. Burrows M, Burrows M, Wheeler DJ (1994) A block-sorting lossless data compression algorithm.
6. Cox AJ, Bauer MJ, Jakobi T, Rosone G (2012) Large-scale compression of genomic sequence databases with the Burrows-Wheeler transform. Bioinformatics. doi:10.1093/bioinformatics/bts173.
7. Boytsov L (2011) Indexing methods for approximate dictionary searching. J Exp Algorithmics 16: 1.1. doi:10.1145/1963190.1963191.
8. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851–1858. doi:10.1101/gr.078212.108.
9. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061–1067. doi:10.1038/ng.437.
10. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, et al. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Methods 7: 576–577. doi:10.1038/nmeth0810-576.
11. Rumble SM, Lacroute P, Dalca A V, Fiume M, Sidow A, et al. (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol 5: e1000386. doi:10.1371/journal.pcbi.1000386.
12. David M, Dzamba M, Lister D, Ilie L, Brudno M (2011) SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics 27: 1011–1012. doi:10.1093/bioinformatics/btr046.
13. Lin H, Zhang Z, Zhang MQ, Ma B, Li M (2008) ZOOM! Zillions of oligos mapped. Bioinformatics 24: 2431–2437. doi:10.1093/bioinformatics/btn416.
14. Zhang Z, Lin H, Ma B (2010) ZOOM Lite: next-generation sequencing data mapping and visualization software. Nucleic Acids Res 38: W743–8. doi:10.1093/nar/gkq538.
15. Eaves HL, Gao Y (2009) MOM: maximum oligonucleotide mapping. Bioinformatics 25: 969–970. doi:10.1093/bioinformatics/btp092.
16. Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, et al. (2009) PASS: a program to align short sequences. Bioinformatics 25: 967–968. doi:10.1093/bioinformatics/btp087.
17. Kim YJ, Teletia N, Ruotti V, Maher CA, Chinnaiyan AM, et al. (2009) ProbeMatch: rapid alignment of oligonucleotides to genome allowing both gaps and mismatches. Bioinformatics 25: 1424–1425. doi:10.1093/bioinformatics/btp178.
18. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713–714. doi:10.1093/bioinformatics/btn025.
19. Gontarz PM, Berger J, Wong CF (2013) SRmapper: a fast and sensitive genome-hashing alignment tool. Bioinformatics 29: 316–321. doi:10.1093/bioinformatics/bts712.
20. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 936–939. doi:10.1101/gr.111120.110.
21. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. doi:10.1093/bioinformatics/btp324.
22. Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinforma Ed board Andreas D Baxevanis al Chapter 11: Unit 11.7.
23. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–360. doi:10.1038/nmeth.1923.
24. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, et al. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5: e1000502. doi:10.1371/journal.pcbi.1000502.
25. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967. doi:10.1093/bioinformatics/btp336.
26. Ferragina P, Manzini G (2005) Indexing compressed text. J ACM 52: 552–581. doi:10.1145/1082036.1082039.
27. Ferragina P, Manzini G (2001) An experimental study of an opportunistic index: 269–278.
28. Mahmud MP, Wiedenhoeft J, Schliep A (2012) Indel-tolerant read mapping with trinucleotide frequencies using cache-oblivious kd-trees. Bioinformatics 28: i325–i332. doi:10.1093/bioinformatics/bts380.
29. Tipton KF (1994) Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions. Eur J Biochem 223: 1–5.
30. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi:10.1038/nature09534.
31. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi:10.1038/nature11632.
32. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, et al. (2008) Single-molecule DNA sequencing of a viral genome. Science 320: 106–109. doi:10.1126/science.1150427.
33. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing: 9.
34. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993. doi:10.1093/bioinformatics/btr509.
35. Prak ET, Kazazian HH (2000) Mobile elements and the human genome. Nat Rev Genet 1: 134–144. doi:10.1038/35038572.
36. Griffiths D (2001) Endogenous retroviruses in the human genome sequence. Genome Biol 2: reviews1017.1–reviews1017.5. doi:10.1186/gb-2001-2-6-reviews1017.
37. Costantini M, Bernardi G (2009) Mapping insertions, deletions and SNPs on Venter's chromosomes. PLoS One 4: e5972. doi:10.1371/journal.pone.0005972.
38. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254. doi:10.1371/journal.pbio.0050254.
39. Osoegawa K, Mammoser AG, Wu C, Frengen E, Zeng C, et al. (2001) A bacterial artificial chromosome library for sequencing the complete human genome. Genome Res 11: 483–496. doi:10.1101/gr.169601.
40. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, et al. (2011) The functional spectrum of low-frequency coding variation. Genome Biol 12: R84. doi:10.1186/gb-2011-12-9-r84.
41. Su X, Zhang L, Zhang J, Meric-Bernstam F, Weinstein JN (2012) PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics 28: 2265–2266. doi:10.1093/bioinformatics/bts365.
42. Roberts KG, Morin RD, Zhang J, Hirst M, Zhao Y, et al. (2012) Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. Cancer Cell 22: 153–166. doi:10.1016/j.ccr.2012.06.005.
43. Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, et al. (2012) An in-depth map of polyadenylation sites in cancer. Nucleic Acids Res 40: 8460–8471. doi:10.1093/nar/gks637.
44. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, et al. (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods 8: 652–654. doi:10.1038/nmeth.1628.
45. Chung CC, Ciampa J, Yeager M, Jacobs KB, Berndt SI, et al. (2011) Fine mapping of a region of chromosome 11q13 reveals multiple independent loci associated with risk of prostate cancer. Hum Mol Genet 20: 2869–2878. doi:10.1093/hmg/ddr189.
46. Goya R, Sun MGF, Morin RD, Leung G, Ha G, et al. (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26: 730–736. doi:10.1093/bioinformatics/btq040.
47. Cridland JM, Thornton KR (2010) Validation of rearrangement break points identified by paired-end sequencing in natural populations of Drosophila melanogaster. Genome Biol Evol 2: 83–101. doi:10.1093/gbe/evq001.
48. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. Nat Methods 5: 183–188. doi:10.1038/nmeth.1179.
49. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, et al. (2012) Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog 8: e1002529. doi:10.1371/journal.ppat.1002529.
50. Malboeuf CM, Yang X, Charlebois P, Qu J, Berlin AM, et al. (2012) Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. Nucleic Acids Res 41: e13. doi:10.1093/nar/gks794.

51. Campbell MS, Mullins JI, Hughes JP, Celum C, Wong KG, et al. (2011) Viral linkage in HIV-1 seroconverters and their partners in an HIV-1 prevention clinical trial. PLoS One 6: e16986. doi:10.1371/journal.pone.0016986.
52. Wilen CB, Wang J, Tilton JC, Miller JC, Kim KA, et al. (2011) Engineering HIV-resistant human CD4+ T cells with CXCR4-specific zinc-finger nucleases. PLoS Pathog 7: e1002020. doi:10.1371/journal.ppat.1002020.
53. Farrell A, Thirugnanam S, Lorestani A, Dvorin JD, Eidell KP, et al. (2012) A DOC2 protein identified by mutational profiling is essential for apicomplexan parasite exocytosis. Science 335: 218–221. doi:10.1126/science.1210829.
54. Dark MJ, Al-Khedery B, Barbet AF (2011) Multistrain genome analysis identifies candidate vaccine antigens of Anaplasma marginale. Vaccine 29: 4923–4932. doi:10.1016/j.vaccine.2011.04.131.
55. Dark MJ, Lundgren AM, Barbet AF (2012) Determining the repertoire of immunodominant proteins via whole-genome amplification of intracellular pathogens. PLoS One 7: e36456. doi:10.1371/journal.pone.0036456.
56. Iorizzo M, Senalik DA, Grzebelus D, Bowman M, Cavagnaro PF, et al. (2011) De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC Genomics 12: 389. doi:10.1186/1471-2164-12-389.
57. Neves L, Davis J, Barbazuk B, Kirst M (2011) Targeted sequencing in the loblolly pine (Pinus taeda) megagenome by exome capture. BMC Proc 5: O48. doi:10.1186/1753-6561-5-S7-O48.
58. Cannon CH, Kua C-S, Zhang D, Harting JR (2010) Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. Mol Ecol 19 Suppl 1: 147–161. doi:10.1111/j.1365-294X.2009.04484.x.
59. Aslam ML, Bastiaansen JW, Elferink MG, Megens H-J, Crooijmans RP, et al. (2012) Whole genome SNP discovery and analysis of genetic diversity in Turkey (Meleagris gallopavo). BMC Genomics 13: 391. doi:10.1186/1471-2164-13-391.
60. Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA (2011) Sequencing and characterization of the guppy (Poecilia reticulata) transcriptome. BMC Genomics 12: 202. doi:10.1186/1471-2164-12-202.
61. Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, et al. (2011) A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. PLoS Genet 7: 1.
62. Tae H, McMahon KW, Settlage RE, Bavarva JH, Garner HR (2013) ReviSTER: an automated pipeline to revise misaligned reads to simple tandem repeats. Bioinformatics 29: 1734–1741. doi:10.1093/bioinformatics/btt277.
63. David M, Mustafa H, Brudno M (2013) Detecting Alu insertions from high-throughput sequencing data. Nucleic Acids Res: gkt612–. doi:10.1093/nar/gkt612.
64. Xing J, Witherspoon DJ, Jorde LB (2013) Mobile element biology: new possibilities with high-throughput sequencing. Trends Genet 29: 280–289. doi:10.1016/j.tig.2012.12.002.
65. Zhao M, Lee W-P, Garrison EP, Marth GT (2013) SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. PLoS One 8: e82138. doi:10.1371/journal.pone.0082138.
66. Farrar M (2007) Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics 23: 156–161. doi:10.1093/bioinformatics/btl582.
67. Adel'son-Vel'skii GM, Landis EM (1962) An algorithm for the organization of information. Sov Math Dokl 3: 263–266.
68. Smith TF, Waterman MS (1981) Indentification of common molecular subsequences. J Mol Biol 147: 195–197.
69. Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162: 705–708.
70. Chao KM, Pearson WR, Miller W (1992) Aligning two sequences within a specified diagonal band. Comput Appl Biosci 8: 481–487.