ART: a next-generation sequencing read simulator

Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth

Supplementary Material

ART sequencing quality or error profiles

Data sources

ART's sequencing quality or error profiles were primarily derived from whole genome re-sequencing data of both the *Pichia Stipitis* genome and the human genome. Illumina and 454 sequencing data of *Pichia Stipitis* were provided by Marth's lab at Boston College and are publically available at NCBI SRA database (accession# SRX001281 and SRX001282); and those of the human genome were from the 1000 genomes project (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/).

Sequencing error profiles derived from *Pichia Stipitis* data

We estimated sequencing error rates of both Illumina and 454 platforms by the following procedure. First, we created 10 sample datasets, each having 20,000 reads that were randomly sampled from the whole read data of a specific platform. We mapped both forward and reserve strands of all reads from each sample dataset to the *Pichia Stipitis* reference genome from JGI at http://genome.jgi-psf.org/Picst3/Picst3.download.ftp.html with our modified ACANA alignment tool [1], which can identify all local optimal and suboptimal alignments. We kept only reads that were uniquely aligned to the reference genome with fewer than five mismatches. We calculated error rates at each position of sequencing reads for each sample dataset, and we reported sequencing error rates averaged over all sample datasets.

Illumina sequencing error profiles

As shown in **Figure** 1, the overall error rate of single-end 35bp reads from Illumina Genome Analyzer I (GA-I) is 2.26%. The dominant error type is base substitution that accounts for 98.2% of all errors, while insertion/deletion errors are rare and account for only 1.8% of all errors. The sequencing error rate is also position-dependent as the error rate increases significantly towards to the end of a read (**Figure 2**). For paired-end reads, the error profile of the first reads is the same or at least very close to that of single-end reads, however, the error profile of the second reads is significantly different. The overall error rate of the second reads is 3.35% from our estimation, and is higher than that of the first reads. The cumulative quality distributions (**Figure 3** and **Figure 4**) also show that the position-dependency structure of Illumina read quality profiles, and different error profiles between the 1st and 2nd reads.

454 sequencing error profiles

We evaluated sequencing error profiles of both GS 20 and GS FLX systems. The overall error rate of the GS FLX is 0.21%, which is markedly lower than the 2.20% error rate of the GS 20 system (**Figure 5**). The dominant errors are insertions/deletions as expected from 454 Pyrosequencing technologies. Almost all 454 sequencing errors are resulted from the over-call or under-call of number of bases in a homo-polymer, and both over-call and under-call rates are homopolymer-length dependent. Overall, GS FLX is less likely to have over-call errors than GS20 as shown in their homopolyer-length error profiles (**Figure 6)**. We also found that the base-call error rate of a 454 read was relatively flat across different positions, although its error rate slightly increased towards the end of the read (see the recalibrated read quality profile in **Figure 13** for example).

SOLiD sequencing error profile

The ART built-in SOLiD error profile was kindly provided by Dr. Heather E. Peckham at Applied Biosystems. The profile was based on 35bp mate-paired reads from SOLiD system 1.0. **Figure 7** shows the position-dependent error profiles, and **Figure 8** shows color-substitution type dependent error profiles.

Read quality profiles derived from 1000 Genomes project data

For each sequencing platform (i.e., Illumina, 454, and SOLiD), read data files were chosen randomly from a large pool of available sequence files at the NCBI ftp site of the 1000 Genomes Project. The data file selection process also took consideration to have a similar number of files from each of the sequencing centers. Note that not all participated sequencing centers generated sequencing data of all three platforms, so data source originators were not the same among the of the three sequencing platforms. 454 data were mainly from Baylor College of Medicine (BCM) and Washington University Genome Sequencing Center (WUGSC); Illumina data were from Broad Institute (BI), Beijing Genome Institute (BGI), Sanger sequence center (SC), and WUGSC; SOLiD data were from Applied Biosystem(AB), BCM, and WUGSC.

For Illumina and 454 platforms, we analyzed raw read data directly from sequencing machines, as well the corresponding GATK-recalibrated read data by the Broad institute [2]. Based on alignment, recalibration assigned a new quality score to each base in a read. Since recalibrated data were available only in the pilot 1 project, our Illumina and 454 sample data were all from the pilot. For SOLiD platform, only raw read data were available and our SOLiD data files were sampled from the whole project. We generated read quality profiles of each platform using empirical position-dependent quality score distributions, which summarized from all sample data files used. The number of data samples used to summarize each read quality profile ranged from 10 to 20.

We created read quality profile plots to make it easy to visualize and compare individual quality profiles from different platforms or different versions of the same platform. In all quality plots, the y-axis is average quality score; x-axis is base position, and each profile is represented by a LOWESS-smoothed line.

Illumina read quality profiles

We summarized quality profiles of paired-end reads of four different lengths: 36bp, 44bp, 50bp, and 75bp. 36bp and 44bp data were generated from Illumina Genome Analyzer I while 50bp and 75bp were sequenced from Illumina Genome II. **Figure 9** and **Figure 10** are the read quality profiles for 36bp and 44bp reads, and **Figure 11** and **Figure 12** are those for 50bp and 75bp reads, respectively. These plots show that there are substantial differences between raw quality and recalibrated quality profile for each read type.

454 read quality profiles

The only difference between single-end and paired-end reads in 454 sequencing is DNA library preparation. Single-end reads are sequenced directly from shotgun DNA fragments, while paired-end reads were sequenced from a specially prepared DNA library, where each fragment consists of two ends of a shotgun DNA fragment, joined by a special adapter. The special adapter is removed after sequencing, and the remaining two fragments are a pair of paired-end reads. Therefore, there is no essential difference in their position-dependent read quality profiles between 454 single-end and paired-end reads. **Figure 13** shows the 454 position-dependent read quality profiles for GS FLX system. Compared to the average recalibrated scores as shown in the Figure, the raw base quality scores are slightly higher at the beginning part of reads, and are slightly lower at the end part.

SOLiD read quality profiles

For SOLiD platform, we evaluated both SOLiD system version 2.0 and version 3.0. Since the recalibrated data were not available at the time of our evaluation, we did not compare a raw quality profile of each system with its recalibrated quality profile. **Figure 14** shows the read quality profile of 25bp mate-pared reads from SOLiD system 2.0, and **Figure 15** is the profile of 50bp mate-paired reads from SOLiD system 3.0.

An application example on performance assessment of alignment tools

As a demon example, we used simulated read data to assess and compare performance of three read mapping tools: MAQ 0.6.6 (http://maq.sourceforge.net/) [3], ELAND 2 from Illumina Genome Analyzer pipeline, and Mosaik 1.0 (http://bioinformatics.bc.edu/marthlab/Mosaik).

Simulation of benchmark datasets

The reference for the simulation was based on the chromosome 17 of the human genome reference (NCBI build 36). We prepared a mutated sequence of the chromosome 17 by changing SNP alleles on the chromosome. The SNP data were from NCBI dbSNP database. In total, 307,416 bases were changed in the mutated sequence compared to the original reference sequence. We then used ART to generate five datasets of 32bp Illumina single-end reads from the mutated reference. Each dataset is of 5X read coverage of the reference. Each dataset was evaluated independently, and we reported the average performance on the five datasets for each tool.

Read mapping

For all read mapping, we used the original chromosome 17 sequence from NCBI as the mapping reference. We ran MAQ and ELAND with their defaulting settings for single-end read mapping, and ran Mosaik with its default settings except with the “multi” option for single-end read mapping.

Performance comparison

The performance comparison showed that MAQ aligned much more reads than both ELAND and Mosaik (**Figure 16**). Compared to MAQ and Mosaik, ELAND aligned only a small proportion of reads with three or more mismatches (**Figure 16**), and had a low percentage of reads aligned correctly among those aligned (**Figure 17**). Among the aligned reads with 2 or fewer mismatches, Mosaik and ELAND had a slightly higher proportion of reads aligned correctly than MAQ (**Figure 17**). With regards to the running time, ELAND was about 4 times faster than both MAQ and Mosaik.

References

1. Huang, W., D.M. Umbach, and L. Li, *Accurate anchoring alignment of divergent sequences.* Bioinformatics, 2006. **22**(1): p. 29-34.
2. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
3. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.

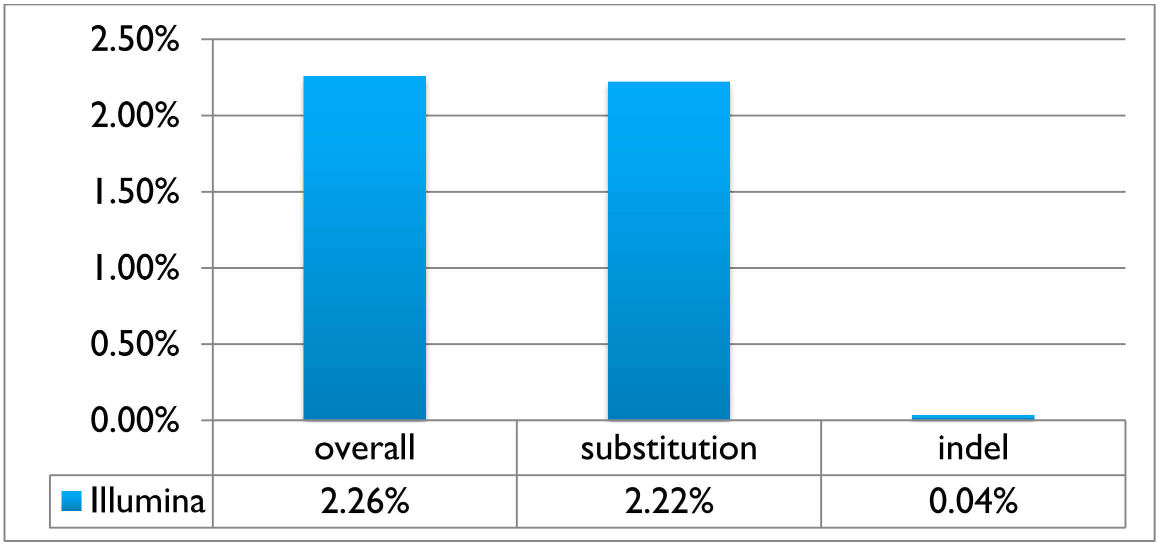


Figure 1: the overall sequencing error rates of Illumina 36bp reads from GA-I system.

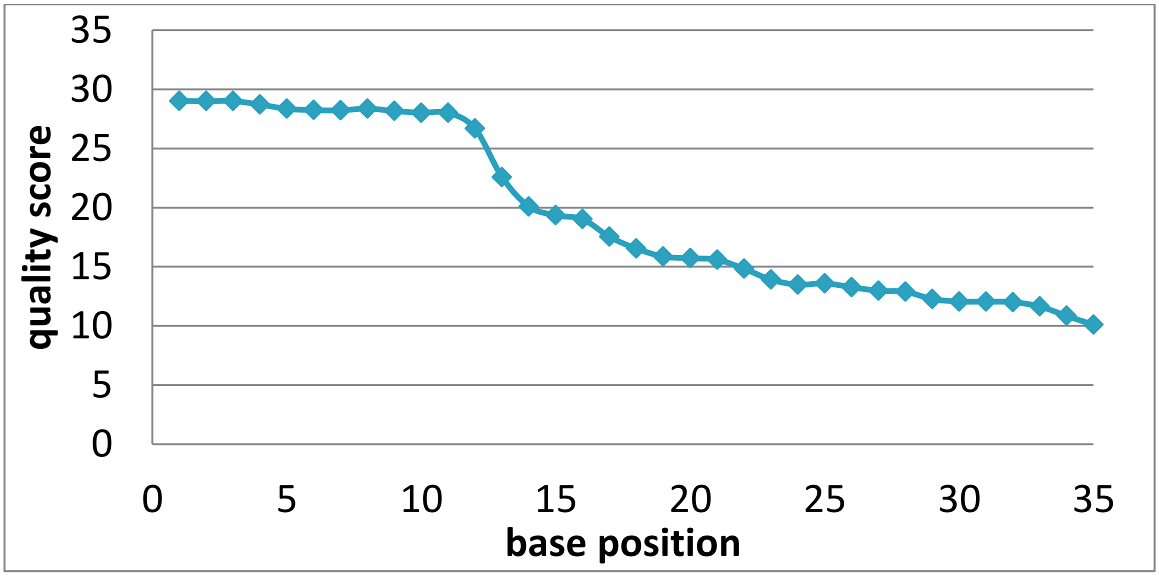


Figure 2: The base quality profile of Illumina 35bp single-end reads from GA-I system.

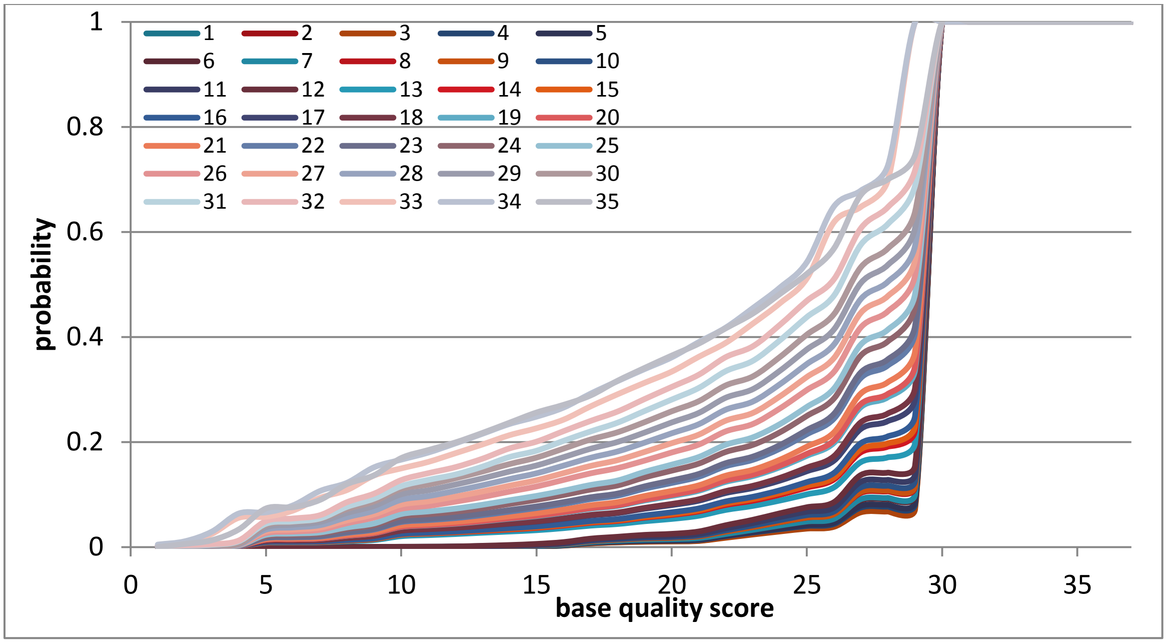


Figure 3: The cumulative base quality distributions of the 1st reads of Illumina 35bp single-end reads. A total of 35 lines of different colors represent 35 base quality distributions at different positions as indicated in the legend.

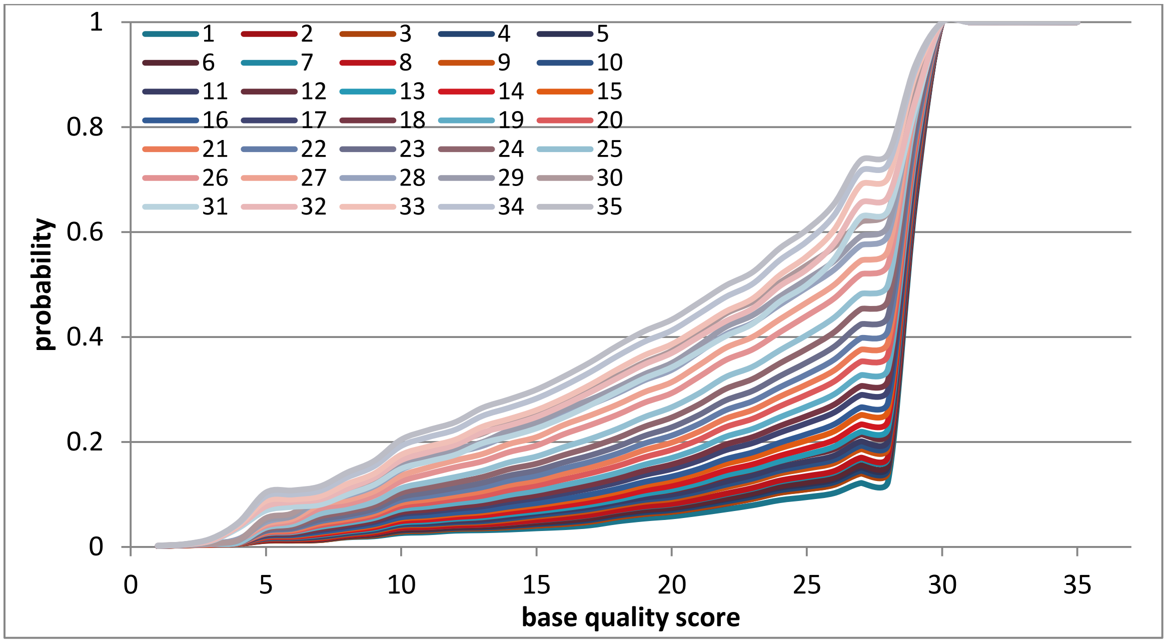
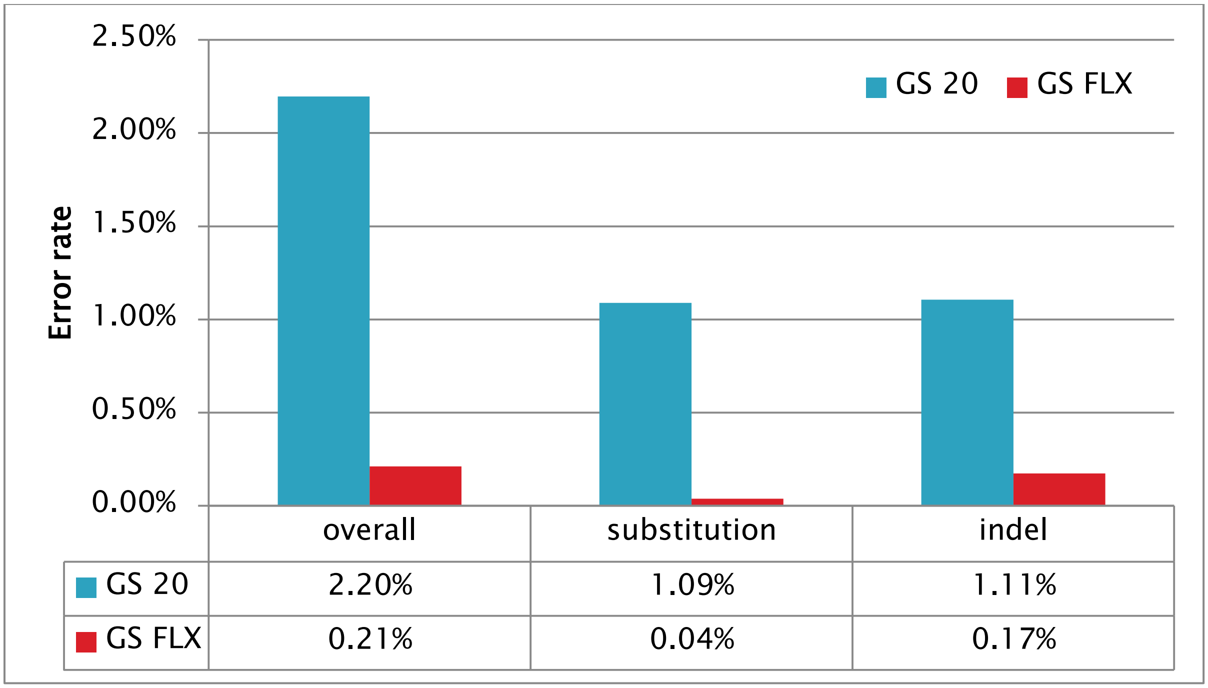


Figure 4: The cumulative base quality distributions of the 2nd reads of Illumina 35bp single-end reads. A total of 35 lines of different colors represent 35 base quality distributions at different positions as indicated in the legend.



	overall	substitution	indel
🟦 GS 20	2.20%	1.09%	1.11%
🟥 GS FLX	0.21%	0.04%	0.17%

Figure 5: Comparison of 454 read error rates between 454 GS 20 and GS FLX systems.

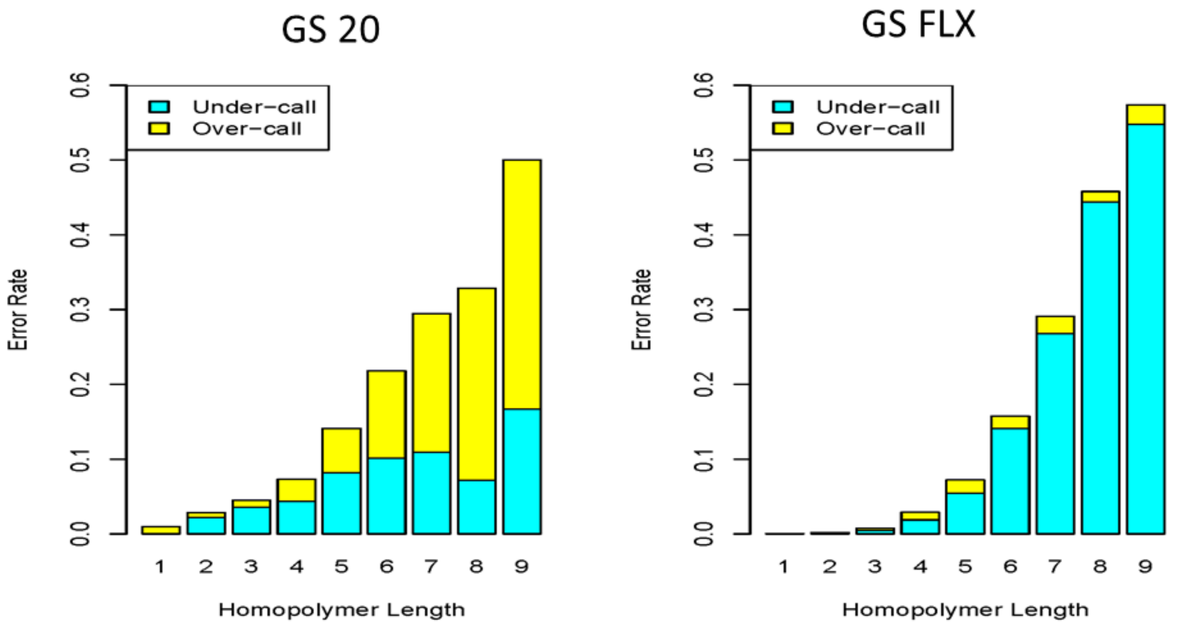


Figure 6: 454 homopolymer-length dependent base-call error profiles of 454 reads. The left panel is the error profile of the GS 20 system, and the right panel is the one for GS FLX system.

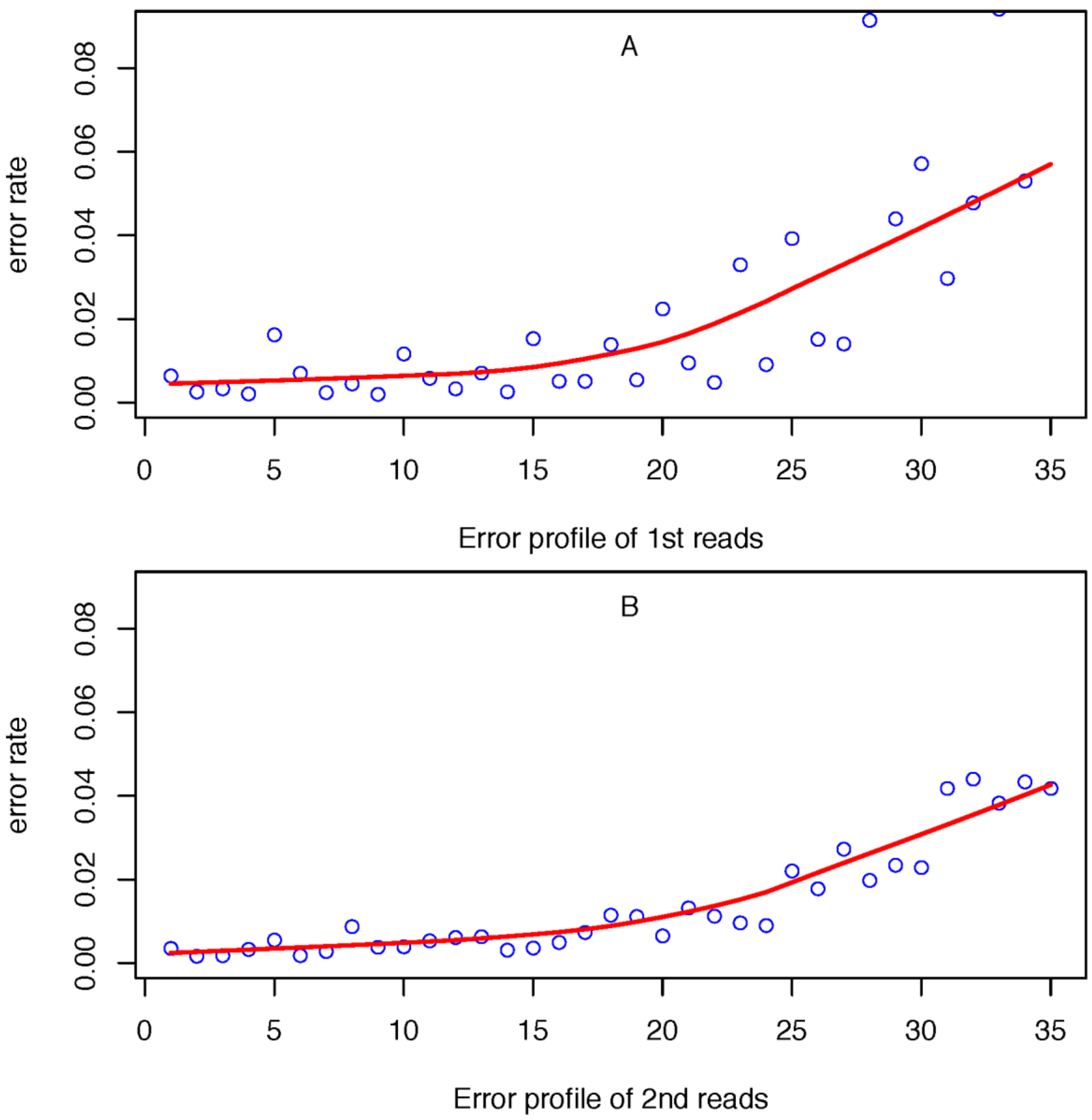


Figure 7: The overall sequencing error profile of 35bp mate-paired reads from SOLiD system 1.0. The x-axis is the base position within a read. The blue circles are error rates at individual positions, and the red lines are LOWESS-smoothed error profiles. The upper panel **A** is the error profile of the 1st reads, and the bottom panel **B** is the one of the 2nd reads.

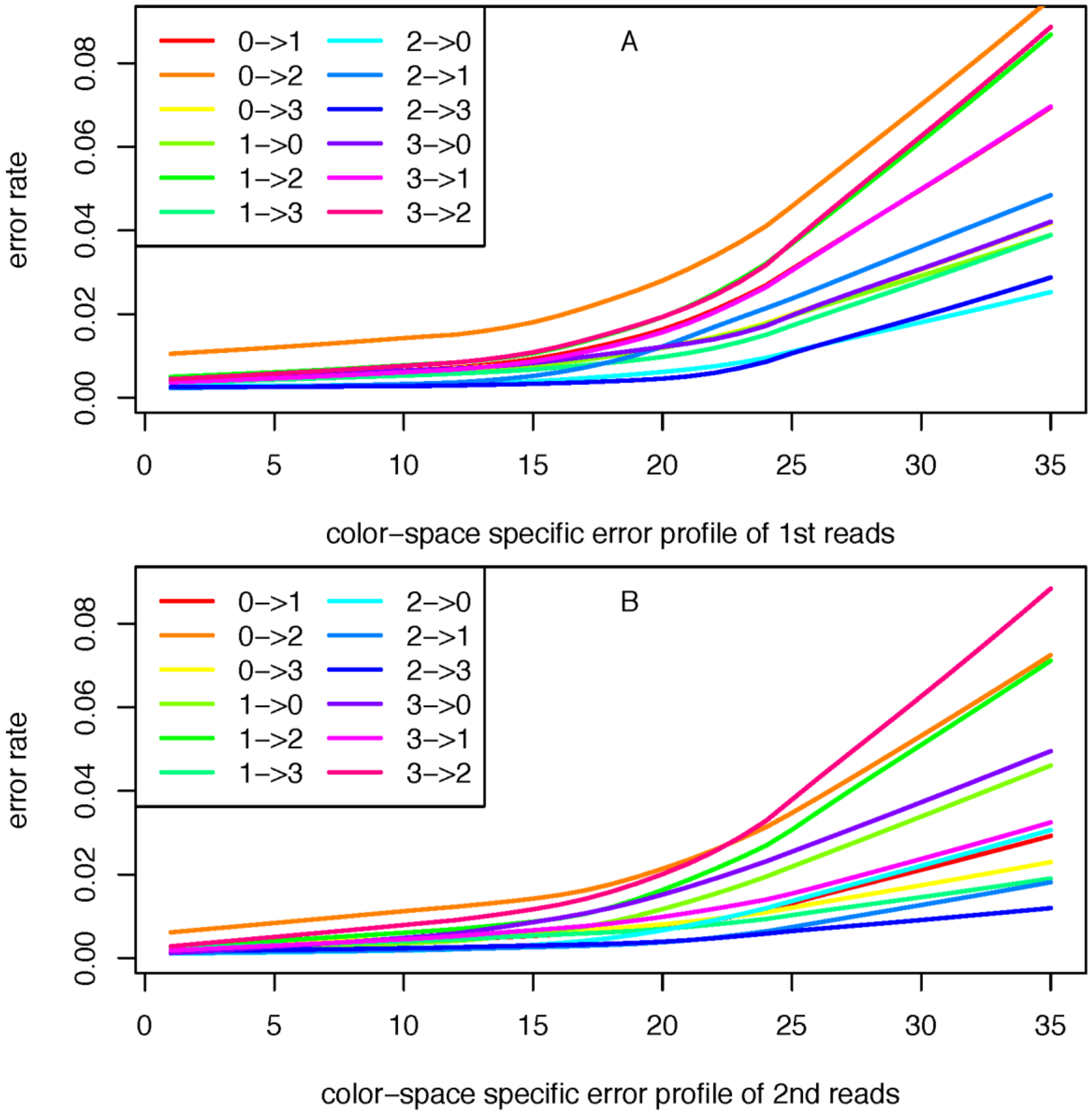


Figure 8: The color-specific error profiles of 35bp mate-paired reads from SOLiD system 1.0. The x-axis is the base position within a read. The lines in different colors stand for profiles of different types of color substitution errors. For example, 0->1 stands for the substitution error type that color base 0 is replicated with the wrong color base "1". All profiles were LOWESS-smoothed. The upper panel **A** is the error profile of the 1st reads, and the bottom panel **B** is the one of the 2nd reads.

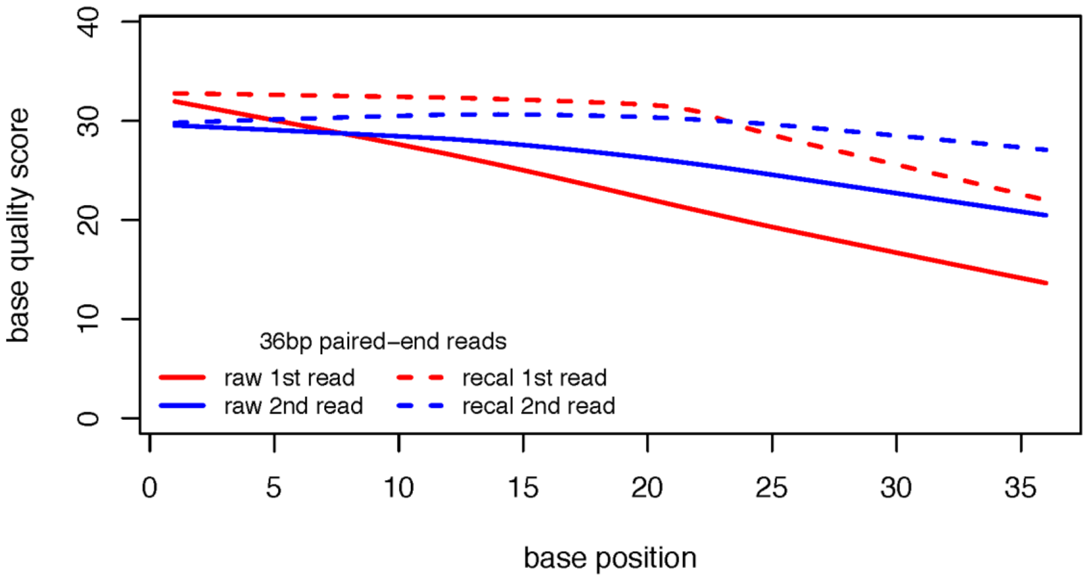


Figure 9: Read quality profiles of 36bp paired-end reads from Illumina Genome Analyzer I. The solid lines are raw quality profiles and dotted-line ones are recalibrated quality profiles. The red ones are for the 1st read, and blue ones are for the 2nd read.

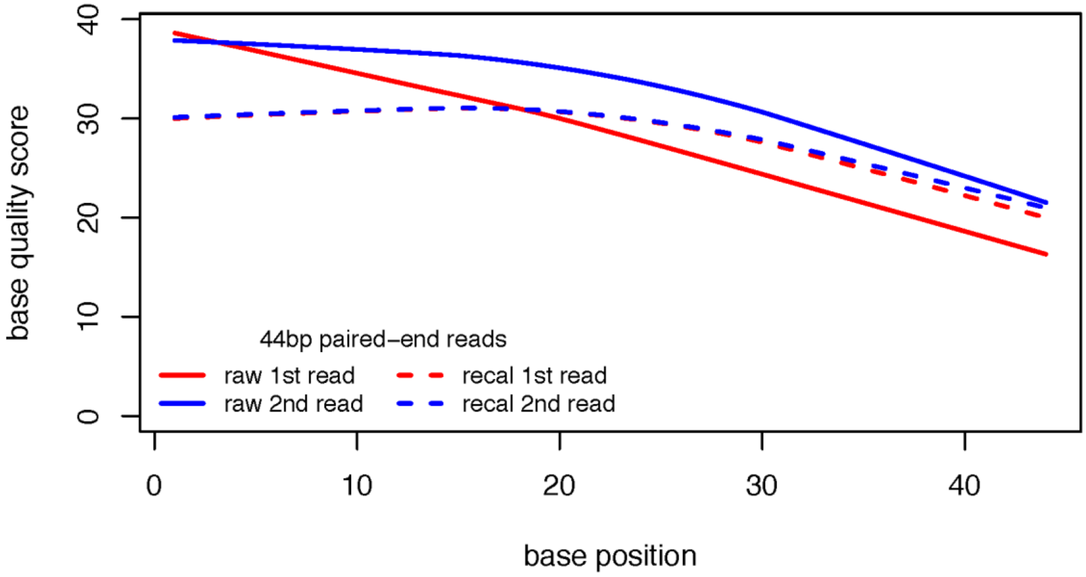


Figure 10: Read quality profiles of 44bp paired-end reads from Illumina Genome Analyzer I. The solid lines are raw quality profiles and dotted-line ones are recalibrated quality profiles. The red ones are for the 1st read, and blue ones are for the 2nd read.

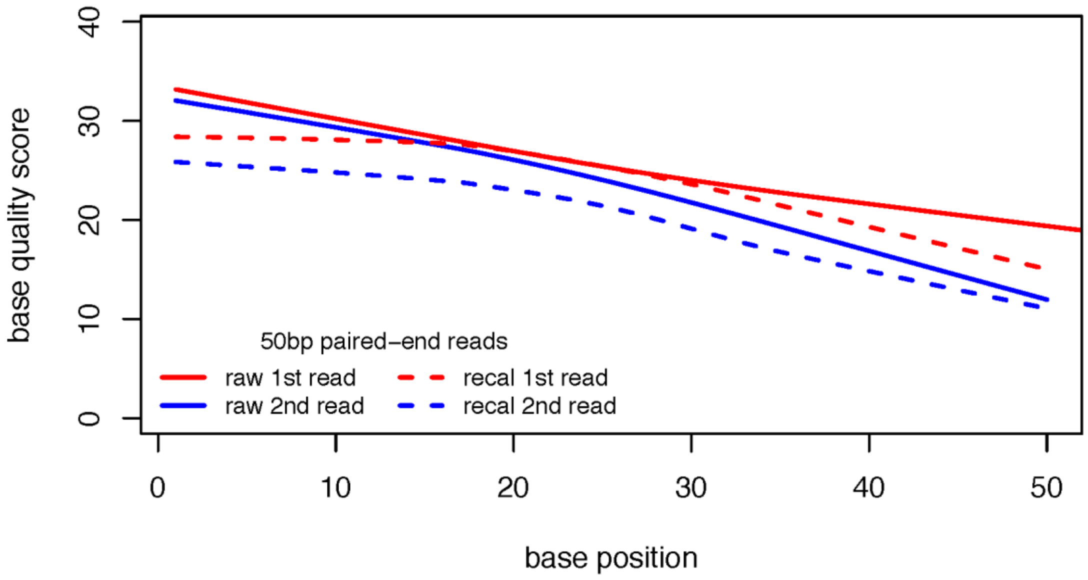


Figure 11: Read quality profiles of 50bp paired-end reads from Illumina Genome Analyzer II. The solid lines are raw quality profiles and dotted-line ones are recalibrated quality profiles. The red ones are for the 1st read, and blue ones are for the 2nd read.

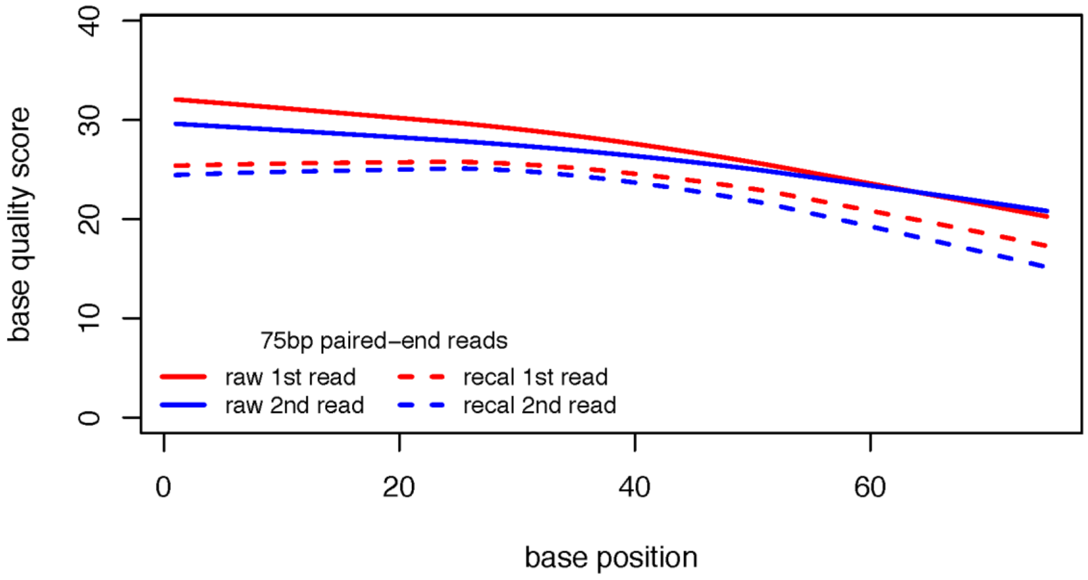


Figure 12: Read quality profiles of 75bp paired-end reads from Illumina Genome Analyzer II. The solid lines are raw quality profiles and dotted-line ones are recalibrated quality profiles. The red ones are for the 1st read, and blue ones are for the 2nd read.

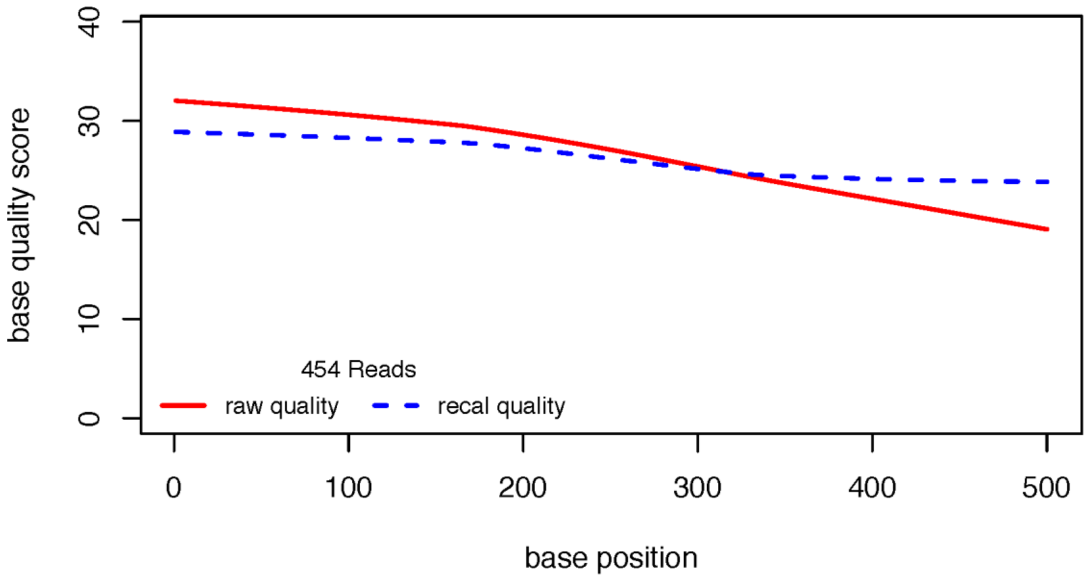


Figure 13: Read quality profiles of 454 reads from 454 GS LFX system. The red-solid line is the raw quality profile and blue-dotted line is the recalibrated quality profile.

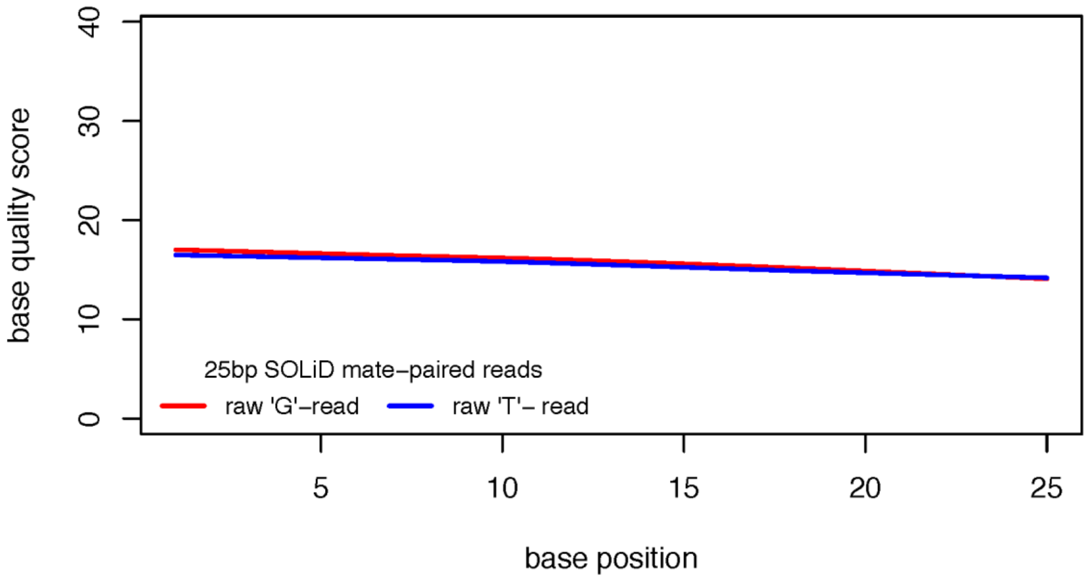


Figure 14: Read quality profiles of 25bp mate-paired reads from SOLiD system 2.0. The red line is for the 'G' (1st) read, and blue one is for the 'T' (2nd) read.

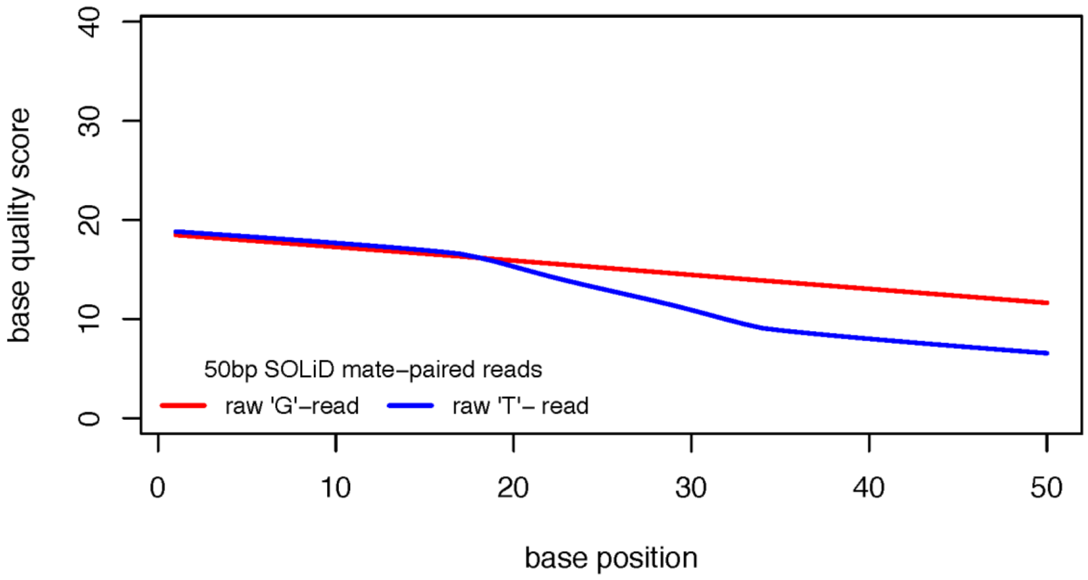


Figure 15: Read quality profiles of 50bp mate-paired reads from SOLiD system 3.0. The profiles are based on the raw quality scores of SOLiD mate-paired reads. The red line is for the 'G' (1st) read, and blue one is for the 'T' (2nd) read.

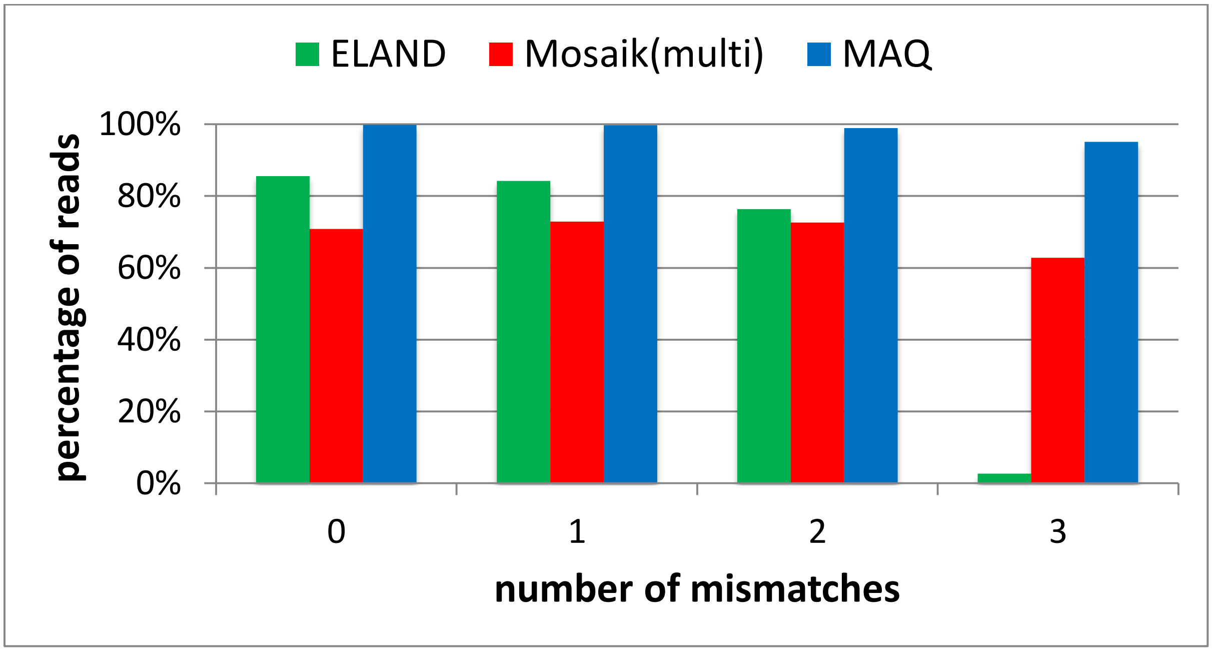


Figure 16: Comparison of the proportion of reads aligned by each alignment tool.

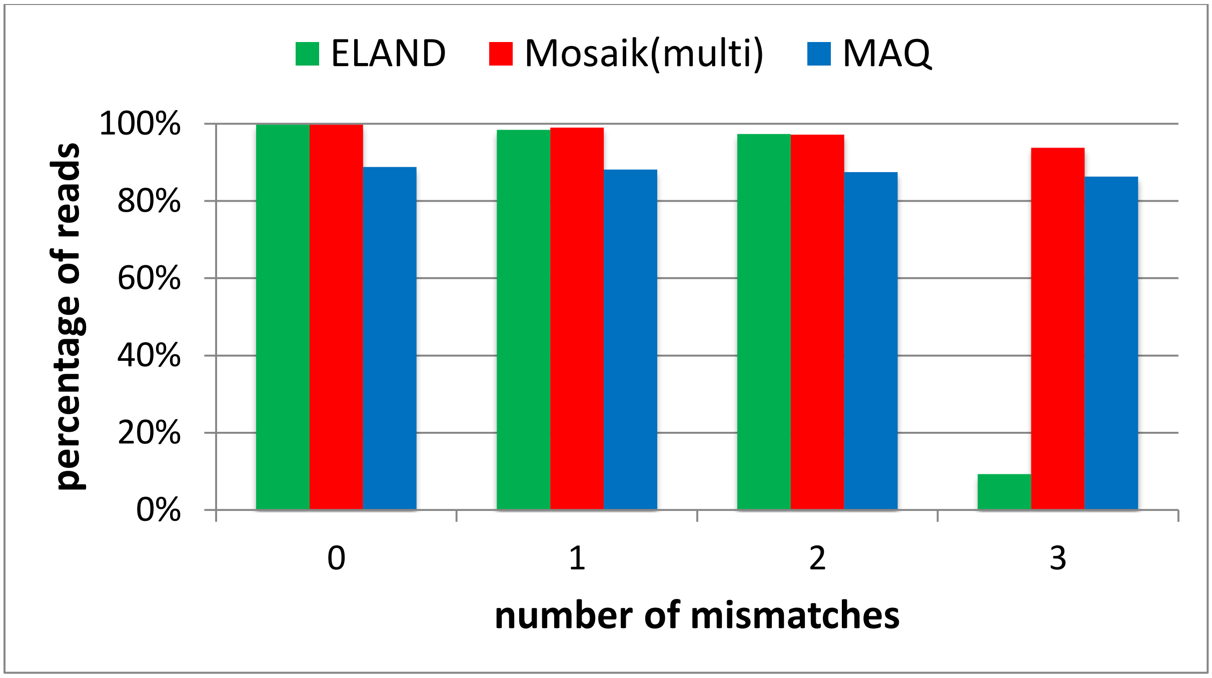


Figure 17: Comparison of the proportion of correctly aligned reads among those aligned