QuorUM (QUality Optimized Reads from the University of Maryland): An error correction software for second generation sequencing reads.

Guillaume Marçais, Aleksey Zimin, Jim Yorke
Institut for Physical Science and Technology
University of Maryland, College Park, MD 20742-2431

Abstract.

Illumina Sequencing data can provide high coverage of a genome by relatively short (100-150bp) reads at a low cost. The base error rates in the reads vary greatly along the read sequence. Low quality bases can be trimmed off or error-corrected. Trimming, especially based on the quality scores, can eliminate large amounts of useful sequence in the reads. Error correction is an alternative to trimming that takes advantage of the high coverage of the genome to make the high confidence base corrections in the reads. Error correction allows for more effective use of the sequencing data thus reducing the coverage depth requirements in sequencing projects and increasing the quality of the resulting assemblies. A conservative error correction is also useful for re-sequencing projects: error corrected reads map to the reference much better than the original sequenced reads. The vast majority of errors in Illumina data are substitutions, where a base is read incorrectly. Also, the sequence quality is usually higher on the 5' end of a read and deteriorates toward the 3' end of the read. Each base in the read has a quality score associated with it. Based on these observations we developed an error correction technique called QuorUM (Quality Optimized Reads from the University of Maryland, pronounced Quorum). QuorUM corrects substitution errors in Illumina reads with high throughput: 4 billion bases per hour on a 48–core computer. We present the comparisons and timings of QuorUM performance to other available error correction techniques: ECHO, HiTEC, Coral and Quake.

Methods

The error correctors are compared on the Illumina reads from the PE library of the bacteria *Rhodobacter sphaeroides* (accession SRX033397). The dataset contains 2 million reads 101bp in length. For the comparison we use the Error Correction Evaluation Toolkit from the Aluru HPC Group (Error correction statistics) and our own metrics (Chimeric reads, Idealized contig sizes and Percentage of initial bases aligned).
ECHO, HiTEC and Coral only correct reads while Quake and Quorum also trim reads.

Chimeric reads

A read is considered chimeric if it two parts (at least 20 bases long) of the read aligns to far apart region of the finished sequence and the read does not align as a whole to the finished sequence.
The error correctors that do not perform trimming (ECHO, HiTEC and Coral) provide only a small improvement or even create new chimeric reads. Only Quake and QuorUM provide a significant improvement.

Corrector	# of chimeric reads
None	1555
ECHO	1764
HiTEC	3810
Coral	1253
Quake	302
Quorum	**111**

Error correction statistics

The sequencing reads are aligned against the finished sequence using BWA. Then, the error corrected reads are compared to their corresponding uncorrected reads to measure the number of errors properly corrected.
In the following table, the ratio "error after / error after" is the percentage of errors that are remaining after the error correction procedure. The mistakes column represent, in percentage of the total number of errors, the number of new errors introduced.
QuorUM corrects the largest number of errors initially present in the sequencing reads while introducing relatively few new errors.

Corrector	error after / error before %	Mistakes %
ECHO	40.7	17.1
HiTEC	33.7	19.8
Coral	32.5	18.3
Quake	89.8	**1.29**
QuorUM	**24.9**	2.47

Idealized contig sizes

The reads are mapped to the finished sequence allowing for at most 1 error. An idealized contig is a stretch of sequence covered by reads overlapping by at least 10 bases. A read can be placed multiple times. The idealized N25, N50 and N90 are approximations of the best assembly obtainable with these error corrected reads.
QuorUM error corrected reads give larger idealized contig sizes for all three measures.

Corrector	N25 kb	N50 kb	N90 kb
None	20.7	12.1	3.54
ECHO	33.6	20.6	5.74
HiTEC	15.6	9.76	2.75
Coral	32.2	18.0	5.48
Quake	20.3	12.7	3.38
QuorUM	**42.4**	**26.3**	**7.10**

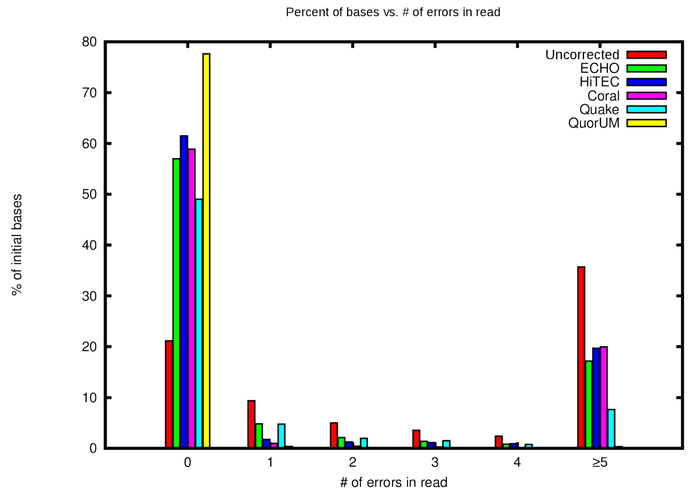
Scalability

Every error corrector tested, except HiTEC, can work in multi-threaded mode to take advantage of multi-CPU/multi-core machines. For the following table, the timing was performed on a 16-core AMD 2.9Ghz computer with 128GB of RAM running Linux 2.6.35.6.
QuorUM scales well in that setting.

Corrector	1 thread (minutes)	16 threads (minutes)	Speed up
ECHO	151	141	1.07
HiTEC	42	-	
Coral	87	10	8.70
Quake	**9**	**5**	1.80
QuorUM	61	**5**	**12.2**

Percentage of initial bases aligned

The error corrected reads are aligned against the finished sequence. The number of errors in a read is defined as the length of the read minus the number of unaligned bases (i.e. mismatches and indels). The following plot represents the number of bases (as a percentage of total number of bases in the uncorrected reads) which aligned in reads with 0, 1, 2, 3, 4, or more than 5 errors.
Despite the trimming, QuorUM output the most sequence which aligns to the finished genome. As seen on the following table, almost all of the bases in QuorUM's error corrected reads are in perfect reads (reads with zero errors).



Percentage of error corrected bases that aligned against the finished genome which are in error free reads.

Corrector	% bases in perfect reads
None	27.4
ECHO	68.4
HiTEC	71.3
Coral	73.0
Quake	74.6
QuorUM	**98.9**

Conclusion

QuorUM correct and trims Illumina reads with high accuracy and high speed on multi-core machines. The code is available under an open source license at: ftp://ftp.genome.umd.edu/pub/quorum/.

References

Kao, W.-C., Chan, A. H., & Song, Y. S. (2011). ECHO: A reference-free short-read error correction algorithm. Genome research, 21(7), 1181-1192. doi:10.1101/gr.111351.110

Ilie, L., Fazayeli, F., & Ilie, S. (2011). HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics, 27(3), 295-302.

Salmela, L., & Schröder, J. (2011). Correcting errors in short reads by multiple alignments. Bioinformatics, 27(11), 1455. Oxford Univ Press.

Kelley, D. R., Schatz, M. C., & Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol, 11(11), R116. doi:10.1186/gb-2010-11-11-r116

Yang, X. (2012). A survey of error-correction methods for next-generation sequencing. Briefings in Bioinformatics.

This project is supported by the following grants: NIH R01HG002945, USDA sub-award 201015739-01, USDA CSREES 2008-04049.