MAQ: Mapping and Assembly with Qualities

Heng Li¹, Jue Ruan² and Richard Durbin¹
¹The Wellcome Trust Sanger Institute; ²Beijing Genomics Institute

Abstract

The vast numbers of very short reads produced by new sequencing technologies such as Illumina GA and AB SOLiD pose a great challenge to data analysis. MAQ, which stands for Mapping and Assembly with Qualities, is a software for mapping short reads to diploid mammalian-sized genomes and calling variants.

MAQ differs from most existing alignment software in several aspects. First, it calculates a mapping quality for each alignment, measuring the probability of the alignment being wrong. This greatly helps accurate variant calling. Second, it maps every read that has a match, placing repetitive reads randomly amongst equally good alternatives, but with a low mapping score, instead of discarding them. This avoids any ambiguity in defining "unique," and provides more data for the subsequent analysis. Third, MAQ aligns mate-pair reads in a sliding window, effectively examining proper paired positions with little computational effort, and allowing reads in repeats to be aligned with high confidence if their mates are in unique regions. Furthermore, MAQ calls the diploid genotype at each base position in the reference with a Phred-like quality, allowing the user to control the sensitivity/specificity trade-off, and facilitating the use of MAQ calls in downstream analysis.

MAQ has been used at multiple sites for human chromosome resequencing, human structural variation analysis, and multiple smaller genome variation studies. In addition to its key functionalities, MAQ can also call short indels, work with SOLiD data, simulate reads and comes with a fast graphic visualizer for the read alignments. Most of its functions come with user friendly interfaces and are well documented. It is available from http://maq.sourceforge.net.

Methods

Indexing reads and scanning the reference

250bp seed. Eland like seed indexing. Guarantee to find 2-mismatch seed hit.

Scoring hits and random mapping

A hit is scored as 2nd-pH, where p is sum of qualities of mismatched bases and H is a 24-bit integer from hashing the coordinate of the hit and read identifier. As a result, MAQ randomly maps a read if there are multiple equally best hits due to genomic repeats.

Decoding colour sequence

Given the reference sequence \( z_1 \cdots z_L \) and the colour read sequence \( c_1 \cdots c_L \), let \( f_i(z_i) \) be the best decoding up to position \( i \): 

\[
f_i(z_i) = \min_{k} \left\{ f_{i-1}(z_{i-1}) + q_i(c_i | z_{i-1} = k, z_{i}) \right\}
\]

where \( q \) translates adjacent nucleotides to the corresponding colour.

Results

Designed for human genome resequencing

1. In alignment, 6 CPU hours and 800MB memory per 1 million read pairs. Easy parallelization on clusters.
2. Compressed binary alignment file: 1 byte per nucleotide on reads.
4. Quickly retrieval of reads in any regions (via mapview).
5. Quick and compact alignment viewer (mapview).

Conclusion

MAQ has been used for several projects including 1000 genome and Cancer Genome Project at this meeting.

Acknowledgements

We are grateful to Tony Cox, Keira Cheetham, Richard Carter and David Bentley from Illumina for beneficial discussions on consensus genotype calling. We also thank Erin Plesas, Ken Chen, David Spencer. LaDeana Hiller and all the MAQ users for their valuable feedback as MAQ has matured, and thank Claudia Walter and the members of the Durbin research group for their helpful comments. This work is funded by the Welcome Trust.

References