MAQ: Mapping and Assembly with Qualities

Heng Li[1], Jue Ruan[2] and Richard Durbin[1]

[1]The Wellcome Trust Sanger Institute; [2]Beijing 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.

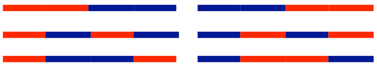
MAQ differs from most existing alignment software in several aspects. First, it calculates a mapping quality for each alignment, measuring the probability of the alignment being wrong. This greatly helps accurate variant calling. Second, it maps every read that has a match, placing repetitive reads randomly 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

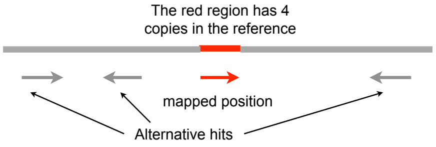
28bp seed. Eland like seed indexing. Guarantee to find 2-mismatch seed hit.



Scoring hits and random mapping

A hit is scored as $2^{24}\cdot q+h$, where q 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 are multiple equally best hits due to genomic repeats.

The number of hits a read has implies the reference copy number of the region where the read is mapped:



Mapping quality

Mapping quality is the phred-scaled probability (Ewing and Green, 1998) that a read alignment may be wrong. Given L-long reference x and l-long read z, the probability that z is mapped at u is:

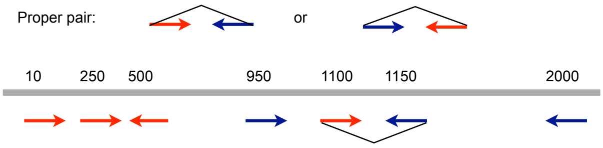
$$p_s(u|x,z)=\frac{p(z|x,u)}{\sum_{v=1}^{L-l+1}p(z|x,v)}$$

where $p(z|x,u)$ equals the product of the error probabilities of mismatched bases. The mapping quaity is:

$$Q_s(u|x,z)=-10\log_{10}\left[1-p_s(u|x,z)\right]$$

Paired-end (PE) alignment

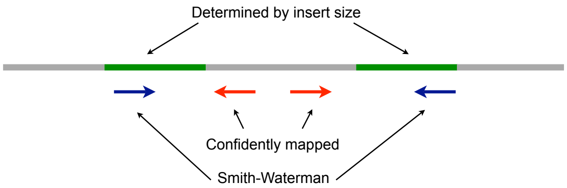
Hit to a read found on the *forward* strand: keep the position in a 2-element queue
Hit to a read found on the *reverse* strand: check the positions in the queue of its mate



Scan at 1150bp, read1 queue: (250,1100); read2 queue: (950). Find read2 hit and check distance: (250,1150) and (1100,1150).

Short indels

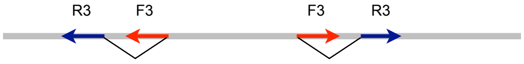
For read pairs where only one end maps, use gapped Smith-Waterman alignment for the other end in a restricted region.



SOLiD alignment

Map in color space:

1. The complement of a colour is itself, and therefore a colour read only needs to be reversed.
2. The correct orientation for a read pair is:



Decoding colour sequence

Given the reference sequence $b_1\ldots b_{l+1}$ and the colour read sequence $\hat{c}_1\ldots\hat{c}_l$, let $f_i(\hat{b}_i)$ be the best decoding up to position *i*. Then:

$$f_1(\hat{b}_1)=q_0\cdot(1-\delta_{\hat{b}_1,b_1})$$

$$f_{i+1}(\hat{b}_{i+1})=\min_{\hat{b}_i}\left\{f_i(\hat{b}_i)+q_0\cdot(1-\delta_{b_{i+1},\hat{b}_{i+1}})+q_i\cdot\left[1-\delta_{\hat{c}_i,g(\hat{b}_i,\hat{b}_{i+1})}\right]\right\}$$

where $g(\cdot,\cdot)$ translates adjacent nucleotides to the corresponding colour.

Consensus calling

Given data D at an aligned position, calculate $P(D|\langle b_1,b_1\rangle)$, $P(D|\langle b_2,b_2\rangle)$ and $P(D|\langle b_1,b_2\rangle)$, assuming the sample is diploid. Knowing the prior of a heterozygote $\langle b_1,b_2\rangle$, we can calculate the posterior probability of each genotype. The consensus $\hat{g}$ is the genotype that maximizes the posterior probability and its quality is $-10\log_{10}[1-P(\hat{g}|D)]$.

Results

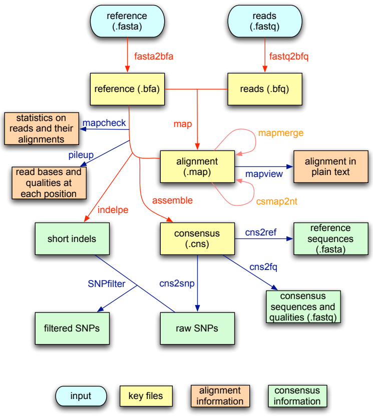
Designed for human genome resequencing

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.
3. Compressed binary consensus file: 4 bytes per nucleotide on the reference.
4. Quickly retrieval of reads in any regions (via maqview).
5. Quick and compact alignment viewer (maqview)

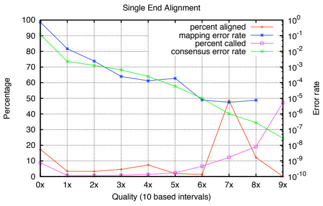
Simulation

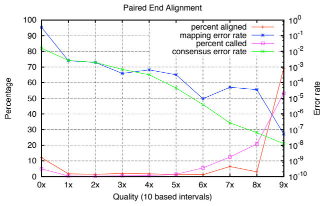
100 million 35bp read pairs on human chrX. Added 0.1% substitutions and 0.01% 1bp indels, assuming diploid sample. Mapped to the human reference genome.

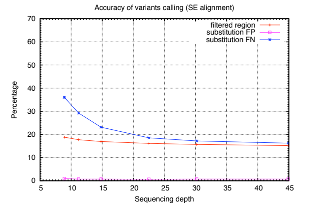
MAQ Work Flow

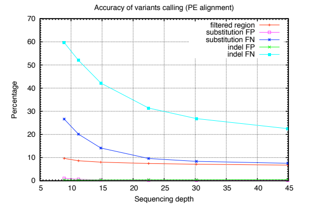


Accuracy of alignments, consensus and variants calls









Maq has been used for several projects including 1000genome 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 Pleasance, Ken Chen, David Spencer, LaDeana Hillier and all the MAQ users for their valuable feedback as MAQ has matured, and thank Klaudia Walter and the members of the Durbin research group for their helpful comments. This work is funded by the Wellcome Trust.

References

Cox A.J. (2007) Ultra high throughput alignment of short sequence tags. *Unpublished*.

Ewing B. and Green P. (1998) Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res.*, **8**:186–194.

Li H., Ruan J. and Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Submitted.*