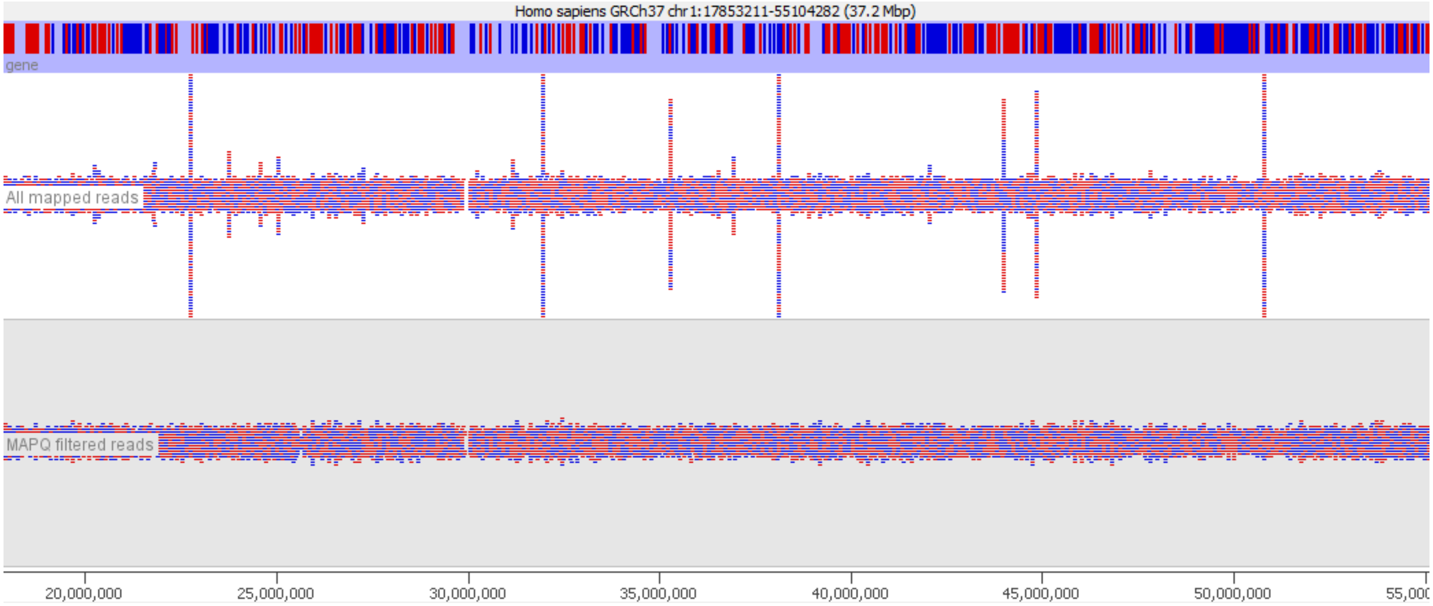
MAPQ values are really useful but their implementation is a mess

One of the standard fields in the SAM/BAM file format is the mapping quality (MAPQ) value. This value can be very useful to help filter mapped reads before doing downstream analysis – unfortunately the implementation of this value is in no way consistent between different aligners so it takes a fair bit of research to know how to use it appropriately. Mis-applying the filter could cause reads to be inappropriately excluded from an analysis.

Introduction

In various stages of the processing of an NGS dataset it can be useful to filter the data to remove poor quality reads. At an early stage this could be reads with poor quality base calls, but after mapping to a reference genome you may want to filter out alignments which show a poor match to the reference, or which could have mapped to a number of different places in the genome. These ambiguously mapped reads can add a lot of noise to an analysis and will tend to accumulate over repetitive regions. In the example below you can see the comparison of the reads from a standard bowtie2 mapping of a genomic dataset, and the result of applying a MAPQ filter to the data. The peaks over repetitive elements are largely suppressed in the filtered data.



To aid with this task the SAM format specification (http://samtools.github.io/hts-specs/SAMv1.pdf) defines the mapping quality (MAPQ) value. In the spec the value is described as:

> MAPping Quality. It equals -10 log10 Pr {mapping position is wrong}, rounded to the nearest integer. A value 255 indicates that the mapping quality is not available.

So in the spec this is pretty clear, it's analogous to a Phred score in a fastq file in that it's a simple transformation of the probability that the mapped position reported is wrong.

In practice, unfortunately, the use of this value is much less clear. For many types of read mapper there is no sensible way to put a p-value on the likelihood that a reported mapped position is wrong so rather than stick to the published spec the aligners have taken the valid value range (0-255) and implemented their own scoring scheme on top of this.

In many cases the values encoded in the MAPQ value hold useful information about the reads and are a valuable resource when filtering data, but the variability with which the value is calculated means that it can be difficult to create pipelines which use this value and which are robust to changes in the aligner used.

Implementation

To try to make more sense of this we've gone through the documentation of a bunch of the most popular aligners to see how they make use of the MAPQ value.

Bowtie1

Reference (http://bowtie-bio.sourceforge.net/manual.shtml)

Bowtie1 sets the MAPQ value to 255 for uniquely mapped reads and 0 for multiply mapped reads, unless the `--mapq` flag was added when the program was launched, in which case the value specified will be used instead.

Bowtie2

Reference (http://biofinysics.blogspot.co.uk/2014/05/how-does-bowtie2-assign-mapq-scores.html)– this page has a great explanation for how alignments in bowtie2 are scored and MAPQ values are assigned.

Bowtie 2 uses a system of flag values for its mapped alignments based on the number of mismatches of various qualities, and the number of multi-mapping reads.

```
MAPQ >= X    #MM Q40   #MM Q20     #MM Q0    Description
0            5         7           15        All mappable reads
1            3         5           10        True multi w/ "good" AS, maxi of MAPQ >= 1
2            3         5           10        No true multi, maxi of MAPQ >= 2
3            3         5           10        No true multi,  maxi of MAPQ >= 3
8            2         4           8         No true multi, maxi of MAPQ >= 8
23           2         3           7         No true multi, maxi of MAPQ >= 23
30           1         2           4         No true multi, maxi of MAPQ >= 30
39           1         2           4         No true multi, maxi of MAPQ == 39*
40           1         2           4         No true multi, only true uni-reads
42           0         1           2         Only "perfect" true unireads
```

In the case of bowtie2 therefore you could use a MAPQ filter of >=40 to get reads which had only 1 convincing alignment, or a lower filter to allow multi-mapped reads where there was a secondary alignment with varying degrees of difference to the primary.

Bismark

The MAPQ values reported in Bowtie1 mode are always 255 (multiply aligning hits are not reported). In Bowtie 2 mode the MAPQ scores are re-calculated using the Bowtie2 scoring scheme.

BWA

BWA actually follows the SAM spec and reports Phred scores as MAPQ values. The calculation is based on the number of optimal (best) alignments found, as well as the number of sub-optimal alignments combined with the Phred scores of the bases which differ between the optimal and sub-optimal alignments.

Tophat

Reference (https://www.biostars.org/p/69773/)

Tophat uses flag values with specific meanings to populate the MAPQ value field. Older versions of tophat set all values to 255 (not available) but any recent version has used an updated scoring scheme.

- 50 = Uniquely mapping
- 3 = Maps to 2 locations in the target
- 2 = Maps to 3 locations in the target
- 1 = Maps to 4-9 locations in the target
- 0 = Maps to 10 or more locations in the target

There are however some caveats which come with these values!

1. Tophat has the option to restrict reporting of hits using the -g parameter and unfortunately the calculation of MAPQ values appears to happen after this filtering resulting in all hits being given a MAPQ of 50 (https://groups.google.com/forum/#!topic/tuxedo-tools-users/i5Xp0srRSIY). This means that to see meaningful MAPQ values you have to set -g to at least 2 (you can then later filter on the primary alignment flag to remove the secondary alignments).
2. Tophat uses a dual mapping strategy where it first tries to align to a transcriptome and only if it doesn’t get a good hit there will it search the entire genome. When you have a read which is uniquely mapped within the transcriptome, but has multiple hits within the genome as a whole the hit will be reported as unique and given a MAPQ of 50, which can result in artefacts in downstream analyses (https://sequencing.qcfail.com/articles/mapping-to-a-transcriptome-can-incorrectly-report-reads-as-mapping-uniquely/).

STAR

Reference (https://rna-star.googlecode.com/files/STARmanual_2.3.0.1.pdf)

Star uses a similar scoring scheme to tophat except that the value for uniquely mapped reads is 255 instead of 50.

> The mapping quality MAPQ (column 5) is 255 for uniquely mapping reads, and int(-10*log10(1-1/[number of loci the read maps to])) for multi-mapping reads. This scheme is same as the one used by Tophat…

HiSat2

Reference (https://github.com/infphilo/hisat2/blob/master/unique.h)

The HiSat2 manual (https://ccb.jhu.edu/software/hisat2/manual.shtml#sam-output) helpfully has no information at all on the meaning of the MAPQ values it assigns. The code which generates it (https://github.com/infphilo/hisat2/blob/master/unique.h) though at least gives some better clues. It looks like the MAPQ value is based on two factors – whether the aligner finds more than one hit, and whether the best hit it finds is a perfect match. It then generates a set of MAPQ values based on the degree to which an alignment is perfect, and the difference between the best alignment and the second best one. The scoring matrix can be seen here (ftp://ftp.cs.washington.edu/courses/cse427/bin/bt221/bowtie2-master/unique.cpp).

In effect it seems that the score for a perfect unique alignment is 44. A perfect alignment with a secondary hit will scale down from 42 to 2. An imperfect unique alignment scales down from 43 to 0. An imperfect primary alignment with a secondary alignment scales between 30 and 0.

A pragmatic level to filter at would therefore seem to be somewhere around 40 to get only very good, unique alignments.

Novoalign

Reference (http://www.novocraft.com/documentation/novoalign-2/novoalign-reference-manual/)

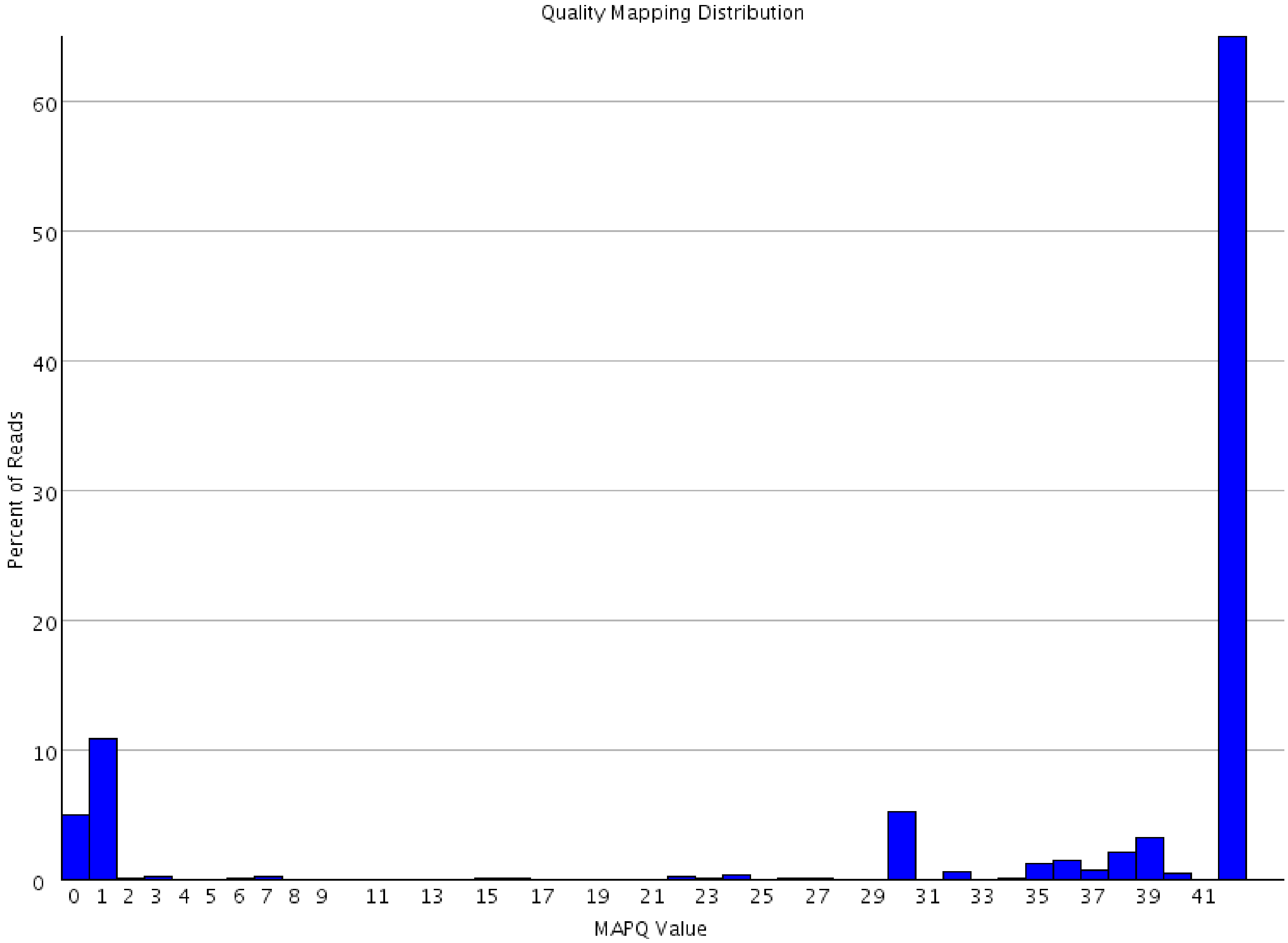
Novoalign creates proper probabilistic MAPQ scores, based on the primary and secondary alignments. It also tries to take into account the likelihood that a read might have come from a region of the genome which was not present in the assembly. The full description can be found in sections 4.3.2 of the manual (http://www.novocraft.com/userfiles/file/Novocraft.pdf). The MAPQ values are capped at 70.

GSNAP

Since GSNAP is a popular aligner (albeit one that we don't personally use) I tried to find the details of how it calculates MAPQ scores, but failed. The documentation (http://research-pub.gene.com/gmap/src/README) has no information on this, and although I can find the source file which reports the MAPQ values (https://github.com/julian-gehring/GMAP-GSNAP/blob/master/src/mapq.c) it has no useful comments and a fairly complex schema so I gave up. If anyone wants to provide a summary of how this works I'll be happy to add it.

Diagnosis

If you're not sure what MAPQ scoring scheme is being used in your own data then you can plot out the MAPQ distribution in a BAM file using programs like BamQC. This will at least show you the range and frequency with which different values appear and may help identify a suitable threshold to use.



Summary

MAPQ values are a useful and important metric in BAM files. Most aligners will report alignments which are of poor quality either due to high numbers of mismatches, or the presence of high quality secondary alignments and the MAPQ value is an easy filter to remove these. We can see from the data above that the documented meaning of this value is not followed by many of the most common aligners, and is calculated on a different basis in pretty much all of them.

What this means in effect is that before applying MAPQ filtering to your data (which you should) you need to consult the documentation for the aligner you are using to find out what value would be appropriate. Generic pipelines should be aware that there is no common standard for fixing a MAPQ threshold.

Privacy - Terms

March 17, 2016 | Simon Andrews (https://sequencing.qcfail.com/author/simon/)

- Technologies : All Technologies (https://sequencing.qcfail.com/technologies/all-technologies/)
- Applications : All Applications (https://sequencing.qcfail.com/applications/all-applications/)
- Software : BamQC (https://sequencing.qcfail.com/software/bamqc/)
- Software : SeqMonk (https://sequencing.qcfail.com/software/seqmonk/)

7 thoughts on "MAPQ values are really useful but their implementation is a mess"

(https://sites.google.com/site/brancolaboratory/home)

Miguel Branco (https://sites.google.com/site/brancolabor

Perfect timing. I have just encountered this issue w values in Tophat when -g 1 and hadn't figured out a Thanks!

March 17, 2016 at 1:34 pm (https://sequencing.qcfail.com/articles/mapq-valu useful-but-their-implementation-is-a-mess/#comm
Reply (/articles/mapq-values-are-really-useful-but implementation-is-a-mess/?replytocom=16#respo

(http://www.bioinformatics.babraham.ac.uk)

Felix Krueger (http://www.bioinformatics.babraham.ac.uk)

As a workaround in our Clusterflow pipeline we run Tophat with `-g 2` and filter the data subsequently to remove secondary alignments with `samtools view -F 0x100`.

March 17, 2016 at 2:22 pm (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/#comment-17)
Reply (/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/?replytocom=17#respond)

Roy

() When using Hisat2 2-2.0.4, I get mapq values of 0,1 and 60. I just sampled around 1 mil lines. I took it that 60 might be uniquely mapped.

June 3, 2016 at 6:06 pm (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/#comment-29)
Reply (/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/?replytocom=29#respond)

(http://www.bioinformatics.babraham.ac.uk)

Simon Andrews (http://www.bioinformatics.babraham.ac.uk)

The latest beta of hisat changed the scoring scheme so that 60 is now the high score rather than 255 – this is actually better since 255 is a reserved number when you have no sensible value to put in. I'll update the article when I get a chance to reflect this.

June 4, 2016 at 10:33 am (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/#comment-30)
Reply (/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/?replytocom=30#respond)

Richard

()

Thanks for this! I was looking for information on Mapping quality for BWA and STAR and found everything I needed on one page.

June 21, 2018 at 9:14 pm (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/#comment-71)
Reply (/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/?replytocom=71#respond)

Kalium

()

I am using sRNA datasets and my MAPQ values are 255. Can I considered them as a good alignment from bowtie?

March 15, 2019 at 12:47 pm (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/#comment-103)
Reply (/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/?replytocom=103#respond)

(http://www.bioinformatics.babraham.ac.uk)

Felix Krueger (http://www.bioinformatics.babraham.ac.uk)

I believe Bowtie did not support the concept of assigning MAPQ values, so it uses 255 for each alignment. You thus need to find a different way to judge your aligment quality (e.g. the number of mismatches, and the uniqueness (parameters -m and –strata))

March 19, 2019 at 10:07 am (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/#comment-104)
Reply (/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/?replytocom=104#respond)

Leave a Reply

Your e-mail address will not be published.

Comment

Your Comment

Name

Your Name

Email

Your E-Mail

Website

Your Website

□**Save my name, e-mail, and website in this browser for the next time I comment.**

ERROR for site owner: Invalid key type

reCAPTCHA
Privacy - Terms

Post Comment

Privacy - Terms

Mailing List (https://sequencing.qcfail.com/mailing-list/)

About Us (https://sequencing.qcfail.com/about/)

Suggest a Post (https://sequencing.qcfail.com/suggest/)

Contact Us (https://sequencing.qcfail.com/contact/)

Babraham Bioinformatics (http://www.bioinformatics.babraham.ac.uk/)

Life (https://www.scilifelab.se/)