Finding the right Aligner and Options for scChIPseq

Pacome Prompsy

October 22, 2019

Context

This document is here to present the reasons behind choosing the specific options and filterings in single-cell ChIP-seq data engineering pipeline, especially concerning the removal of multimapped reads. Two options are usually chosen concerning multimapped reads, e.g. reads that are very similar in sequence to multiple sequences in the genome :

- Either report one location in the genome in a pseudo-random manner
- Discard the read

In scChIP-seq we feel it is important to remove those reads that are potentiall false positive. Indeed, the **sparsity** & the "binary" nature of the data makes it that *adding a false positive in a region for a cell may have a strong impact on the clustering & differential analysis.* Removing those reads will ensure that all reads are true positive. More importantly, when not removing theses reads in scChIP-seq and depending on the mapper (bowtie, bowtie2, bwa, STAR) and mapper options, some ** narrow and high signal peaks** are appearing in the profiles (for H3K27me3 mark), fully composed of reads that are very similar in sequence to multiple locations in the genome. These peaks have a strong bias on the downstream analysis and drive the clustering of cells. These peaks are often located in close proximity to centromers and/or are located in repeat regions.

In Figure 1, the peak is located in the centromer of chromosome 1, in a highly repeated encode black region and has a very high signal 3198 compared to average signal below 100. Reads in this peak have a 90% or more match to many different locations in the genome.

The historical mapper used (in V1) was **bowtie1 (v1.2)** in single-end, using the options '-m 1' that successfuly discard any reads that is similar to more than one region in the genome. However, the evolution of the technology, notably the reduced size of cell barcode, allowed for mapping in paired-end mode and more precise duplicate removal. We found that bowtie1 is not efficient in mapping paired-end reads with different sizes (read1 101bp, read2 ~20bp). This is why we choose to benchmark other aligners and options and found that STAR aligner was the best fit for mapping paired-end reads with high mapping rate (~10-20% more mapped reads than bowtie). However, STAR mapper was not always sucessfully detecting & removing multimapped reads and narrow peaks composed of multimapped reads were biasing the analysis.

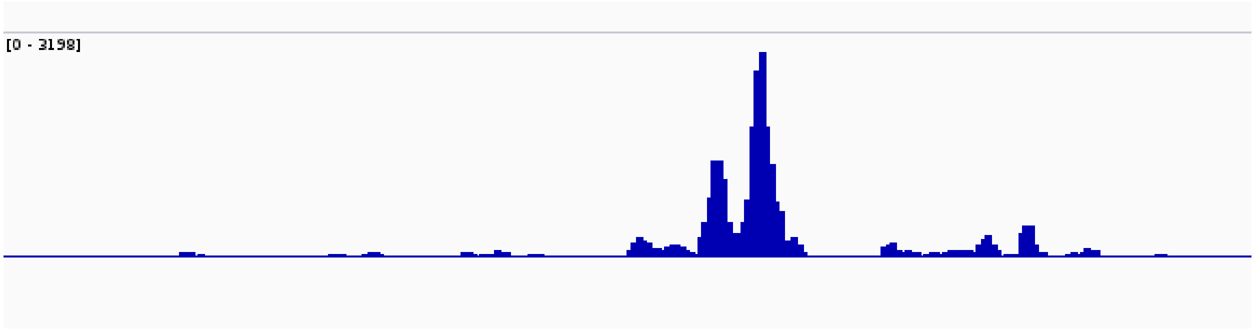

Figure 1: Example of multimapped peak on MM468-5FUR3 sample H3K27me3 mark on chr1, using STAR 'default' parameters .

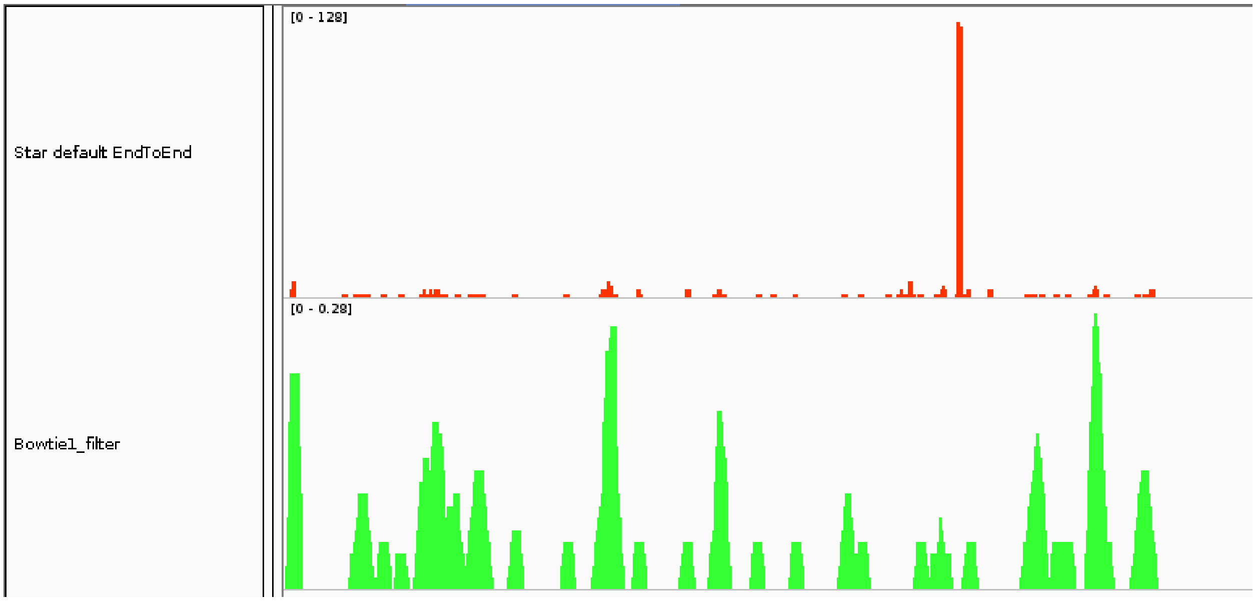

Figure 2: Difference of mapping between Bowtie -m1 filter & STAR EndToEnd default mapping for a strange peak

Tuning of STAR parameters

In order to keep the higher mapping rate as well as convenient paired-end mapping, we decided to keep STAR aligner and tried to tune the parameter & filtering. To do that, we mapped the same dataset (MM468-5FUR3) with either STAR varying the parameters or Bowtie -m1 (ground truth, no multimappers).

After looking into STAR 2.6 parameters and discussing with A. Doblin, the creator & maintainer of STAR rna-star, few parameters were found to have an impact on multimapped reads :

- alignEndsType : either Local to allow soft clipping of reads or EndToEnd to prevent soft clipping, default Local.
- winAnchorMultimapNmax : “max number of loci anchors are allowed to map to”, default 50. This number is actually the maximum number of times an seed (e.g. small part of the read) can be “anchored” to a sequence in the genome. For example, if a seed is 30bp and it matches 51 sequences in the genome, only the first 50 will be fetched and elongated. If this is number is too low, there is a chance that a read is found unique because the seed was not anchored to all the possible locations in the genome, and therefore one of the 50 sequence may outscore the 50 others, therefore be considered as “unique”.
- outFilterMultimapScoreRange : the score range below the maximum score for multimapping alignments, default 1. Defines what is called “multimapped” or not.

In order to benchmark those 3 parameters, we retrieved reads from either “normal” regions (wide region without any outlier peak) or “mulimapped” (strange) regions in the STAR EndToEnd mapping. Those reads were then realigned, varying each parameter. The goal here was to find the set of parameters minimizing the unique mapping of “strange” peaks reads while maximizing the mapping of “normal” regions.

Figure 3 & 4 show that increasing outFilterMultimapScoreRange from 1 to 2 is the main source of improvement, along with increasing winAnchorMultimapNmax from 50 to 1000. The difference between Local & EndToEnd is not that great, however EndToEnd mapping allows to reduce to ~0% the strange peak unique mapping while Local mode is stucked at ~25%. The optimal parameters were thus set : alignEndsType = EndToEnd mode, winAnchorMultimapNmax=1000, outFilterMultimapScoreRange=2.

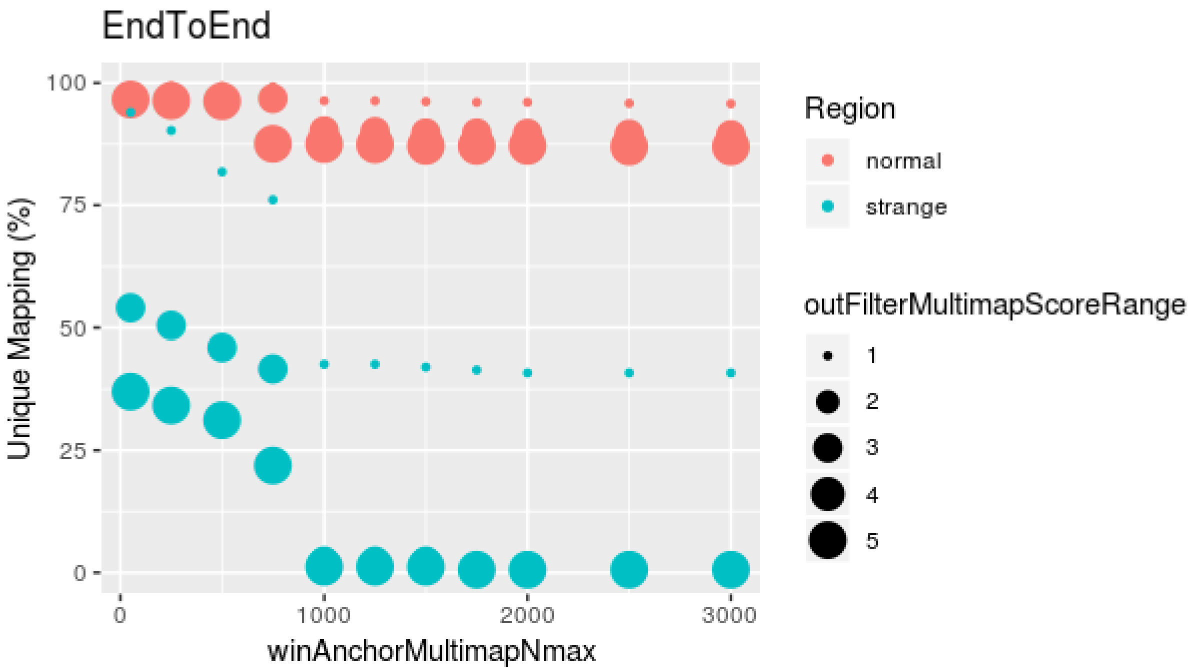

Figure 3: Benchmark of parameters in EndToEnd mode

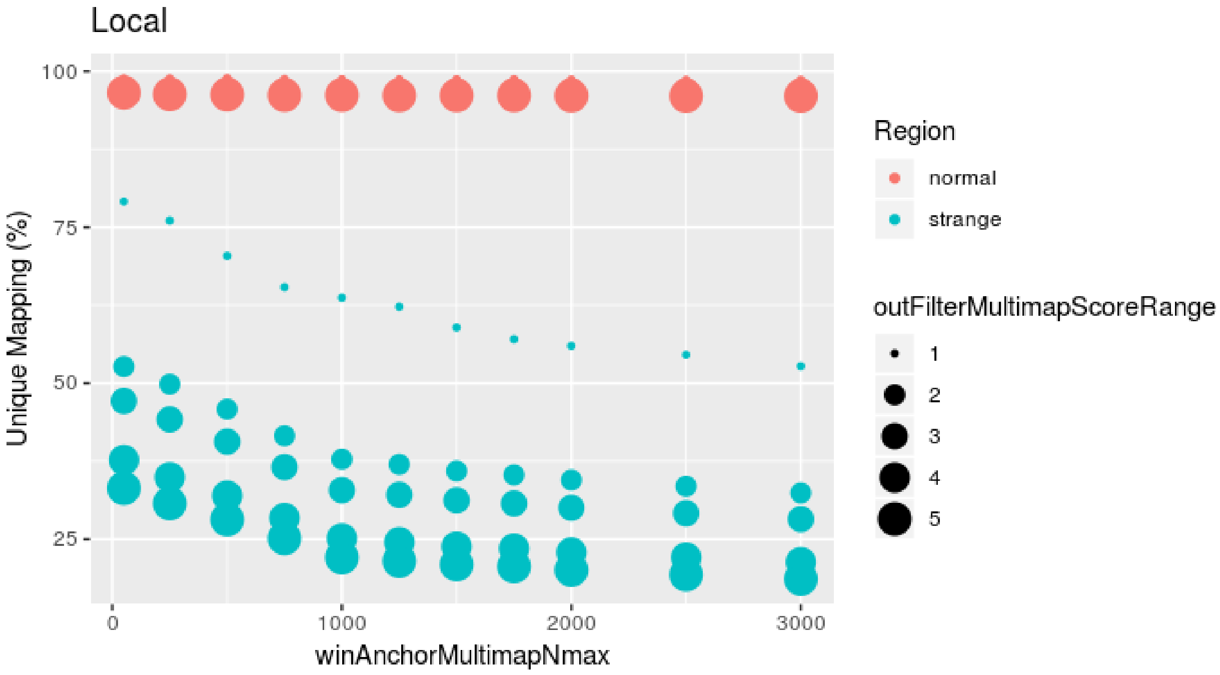

Figure 4: Benchmark of parameters in Local mode

Filtering out with Encode Black Regions

We realigned the dataset using STAR optimal parameters, and compared to bowtie -m1 alignment.

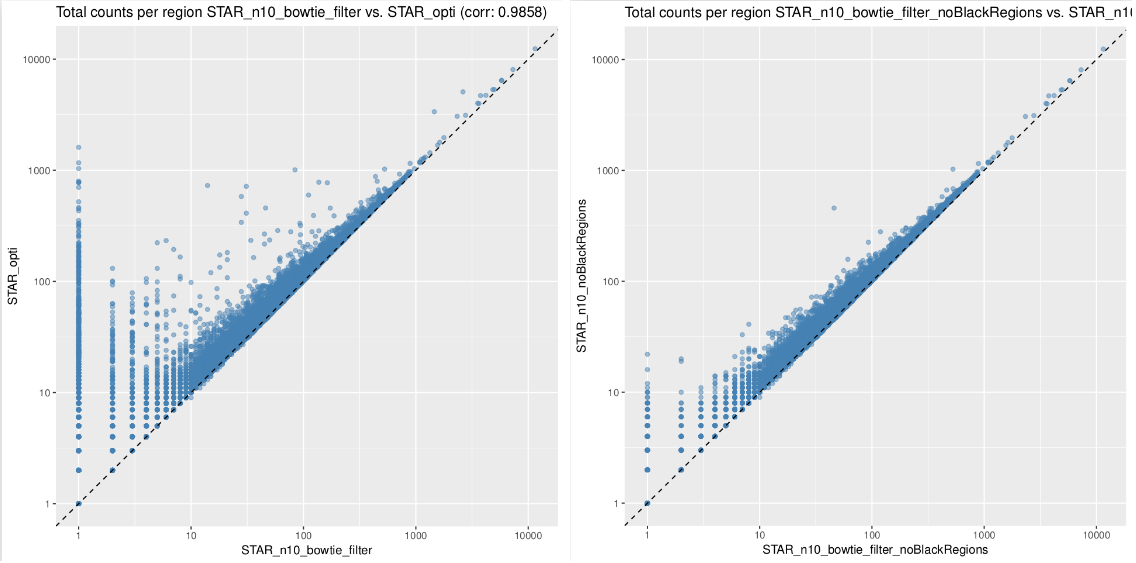

Figure 5 (left) shows that mapping with STAR there are still plenty of regions (50kb windows) that contain high amount of reads in STAR but not in bowtie (multimapped peaks). As we observed that these peaks are often present in Encode black region, we tried to filter using ENCODE black regions :

As we can see in Figure 6 (right), the number of outlier regions in STAR with optimal parameters compared to bowtie is not as important as before filtering.

Therefore, we choose to constantly filter the BAM files (before count) with the ENCODE v2 black regions.

Reducing the number of allowed mismatches

We tried to reduce the number of mismatches from default 10 to 5, but the number of outlier region does not disminish greatly while the total count is slightly decreased, therefore we chose to keep the max number of mismatches allowed to 10

Other parameters specified

Three other parameters are present in STAR command line, explained below:
–peOverlapNbasesMin 10 : allows overlapping of reads by 10 bases. This option allow to retrieve short mono-nucleosomes (<120bp) where read1 sequence and read2 sequence are overlapping.

–alignIntronMax 1 : for mapping DNA, doesn't allow splicing of sequence.

–alignMatesGapMax 450 : Maximum distance between read1 and read2. Allows at maximum 450bp between each mates (tri-nucleosomes)

Conclusion

We successfuly optimized parameters in order to map efficiently in paired-end using STAR 2.6 aligner in order to not get a high number of multimappers and multimapped peaks that could bias downstream single-cell analysis while still reaching higher level of mapping.

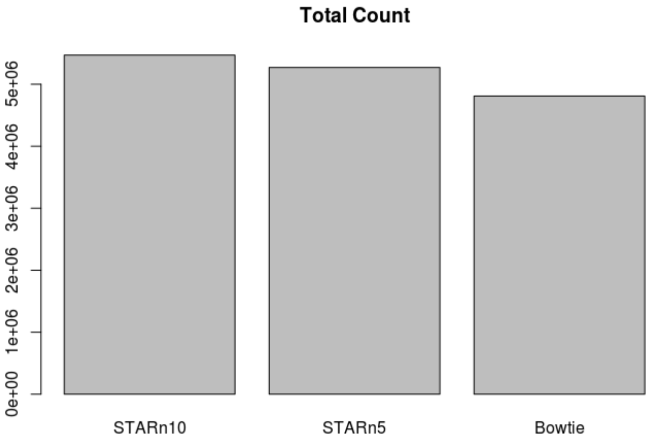

Figure 5: Total read count with STAR 10 mismatches, STAR 5 mismatches and Bowtie -m1

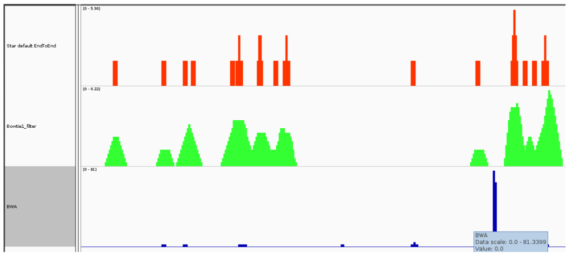

Figure 6: BWA mapping : an example of strange peak not present in STAR with non optimal parameter or Bowtie1 -m1

STAR :
–alignEndsType EndToEnd –outFilterMismatchNmax 5 –outFilterMultimapScoreRange 2 –winAnchorMultimapNmax 1000 –alignIntronMax 1 –peOverlapNbasesMin 10 –alignMatesGapMax 450 –limitGenomeGenerateRAM 25000000000 –outSAMunmapped Within
Filter with ENCODE black regions v2.

Supplement : example of peak using BWA

When testing with BWA, the alignment produced an even higher number of multimapped peaks. An example of such peak is presented on Figure 7.