**From FASTQ to BAM (for RNAseq)**

*Script name:* rnaseq\_single\_end\_fastq\_to\_bam\_job\_array.pl

What it does: get a BAM file ready for SNP calling per accession from RNAseq single end fastq files.

*Synopsis:*

***perl rnaseq\_single\_end\_fastq\_to\_bam\_job\_array.pl -g genome/dir/path -r musa.fasta -x fastq -cu musa\_acuminata***

/!\ The jobarray was done with SGE on a cluster

/!\ FASTQ files have to be present in the current directory

/!\ Change the parameters –t (5th line of the script) to the number of accessions you have to treat in the current directory

*Parameters:*

-r (string): reference fasta filename

-x (string): file extension (fastq)

-cu (string): cultivar

-g (string): location of index files for from STAR mapper

*All the steps in detail*

* Trim low quality ends and remove adapters (illumina) with cutadapt <http://cutadapt.readthedocs.io/en/stable/guide.html>

Fixed parameters:

-b AGATCGGAAGAGC (universal sequence for illumina). Sequence of an adapter that was ligated to the 5’ or 3’ end. The adapter himself is trimmed and anaything that follow too if located at 3’ end.

-O 7: Minimum overlap length. If the overlap between the read and the adapter is shorter than LENGTH, the read is not modified. This reduces the no. of bases trimmed purely due to short random adapter matches.

-m 30: Discard trimmed reads that are shorter than 30.

-q 20,20: Trim low-quality bases from 5' and/or 3' ends of reads before adapter removal. If one value is given, only the 3' end is trimmed. If two comma-separated cutoffs are given, the 5' end is trimmed with the first cutoff, the 3' end with the second. The algorithm is the same as the one used by BWA (see documentation).

* Control the quality of the raw fastq file with FASTQC <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
* Mapping with STAR in 2-pass mode

<https://github.com/alexdobin/STAR>

* Add read group and sort BAM with AddOrReplaceReadGroups from PicardTools

<https://broadinstitute.github.io/picard/>

* Mark duplicate reads and index BAM with MarkDuplicates from PicardTools

<https://broadinstitute.github.io/picard/>

* Split ‘N CIGAR’ reads with SplitNCigarReads from GATK

<https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_rnaseq_SplitNCigarReads.php>

* Realign indels with IndelRealigners from GATK (2 steps)

<https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_indels_IndelRealigner.php>

Remarks: All the fixed parameters can be changed directly in the script.