**From FASTQ to GVCF**

*Script name:* radseq\_paired\_end\_fastq\_to\_gvcf\_jobarray2.pl

What it does: get one GVCF file ready for SNP calling per accession from radseq paired end fastq.gz files.

*Synopsis:*

***perl radseq\_paired\_end\_fastq\_to\_gvcf\_jobarray2.pl -r musa.fasta -x fq.gz -cu musa\_acuminata –q 64***

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

/!\ quality base of radseq paired end (in our case) is 64 (and not 33)

/!\ FASTQ.GZ files have to be present in the current directory in the format Acc\_name\_1.fq.gz, Acc\_name\_2.fq.gz

/!\ 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 (fq.gz)

-cu (string): cultivar

-q (int): base quality (64 in our case)

*All the steps in detail*

* Control the quality of the raw fastq file with FASTQC <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
* 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.

--quality-base=64: Assume that quality values are encoded as ascii(quality + QUALITY\_BASE). The default (33) is usually correct, except for reads produced by some versions of the Illumina pipeline, where this should be set to 64. (Default: 33)

-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 again the quality with FASTQC after the trimming of the reads
* Check homogeneity of the paired fastq files

This is done with the perl script **compare\_fastq\_paired\_v5.pl** which should be in the same directory as ***radseq\_paired\_end\_fastq\_to\_gvcf\_jobarray2.pl***

* Mapping with BWA with default parameters

<http://bio-bwa.sourceforge.net/>

* Add read group and sort BAM with PicardTools

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

* Index BAM with Samtools

<http://samtools.sourceforge.net/>

* 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>

* Create a GVCF file with HaplotypeCaller from GATK

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

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