Methods

Trimming and aligning

We have used the scripts made by LM Safrané. The pipeline used here was almost the same as the one used for the Saccharomyces alignment. The first step was trimming the ends of the reads, and keep only the ones with an average quality of 15 or more. The tool ‘trimmomatic’ was used, with the parameter MINLEN 140, this time, keeping a read of minimum length of 140bp. The bwa tool was then used to index the reference fasta file (‘bwa index’ command) and ‘bwa mem’ in order to align all the reads against the reference fasta file, as done previously for the yeast project. The sam files that were produced by the bwa aligner were then sorted and converted into bam files through ‘Piccard tools’.

The new step of this pipeline, is the marking of duplicated reads, with ‘piccard –tools MarkDupilcates’, that takes a sorted bam file and will add information about reads that might come from the same DNA fragment, in order to avoid counting the information given by one fragment more than one time.

(<https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.4.0/picard_sam_markduplicates_MarkDuplicates.php>)

Haplotype Caller and Base Quality Score Recalibration

Since the genome of the arabidopsis is diploid, in order to analyse its sequenced data we had to use the ‘Haplotype Caller’ from the GATK4 pipeline, in order to have a first variant call. The vcf file was then split into two other vcf files, one containing only the Indel information, and one containing the SNPs, with the tool ‘SelectVariants’ from the GATK4 pipeline because the filtration methods differ. Then of course the new vcf files were filtered to get rid of all the false variants, through ‘VariantFiltration’ from GATK4-tools, using different parameters for indels and for snps. We have performed a first Base Quality Score Recalibration (BSQR) that takes into account the quality score assessed by the sequencer, when getting the vcf files.(pas sûr du tout) For this step, the the filtered vcf files were needed aswell. The ‘Haplotype Caller’ was then used on the newly recalibrated bam file, in order to get a new vcf without too much garbage information.

(cf BC.7107 - Mutation mapping I 2019 LMS.pdf)

Annotation

For the annotation step, SnpEff was used in order to annotate the vcf file, to get rid of all the synonymous and intergenic variants, since they bring no new information to our analysis, and also to keep only the variants that are found in less than 2 strains. At the end we got a nice annotated and filtered vcf file. The analysis was further done by using GeneSearch, and since the whole bam files were a bit too big to handle, we have used the shortened versions of the bam files, provided by the professors, containing only the information about the GAI gene on the first chromosome, and the SPY gene on the third chromosome.