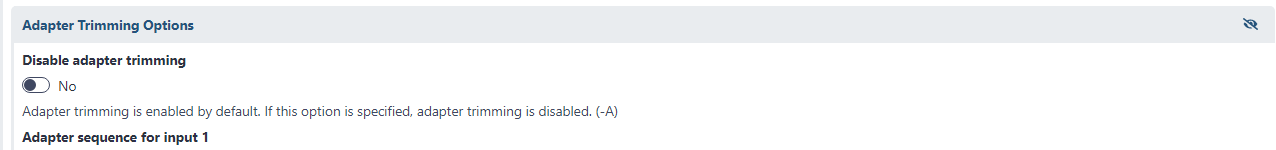
**STR profiling for cell lines**

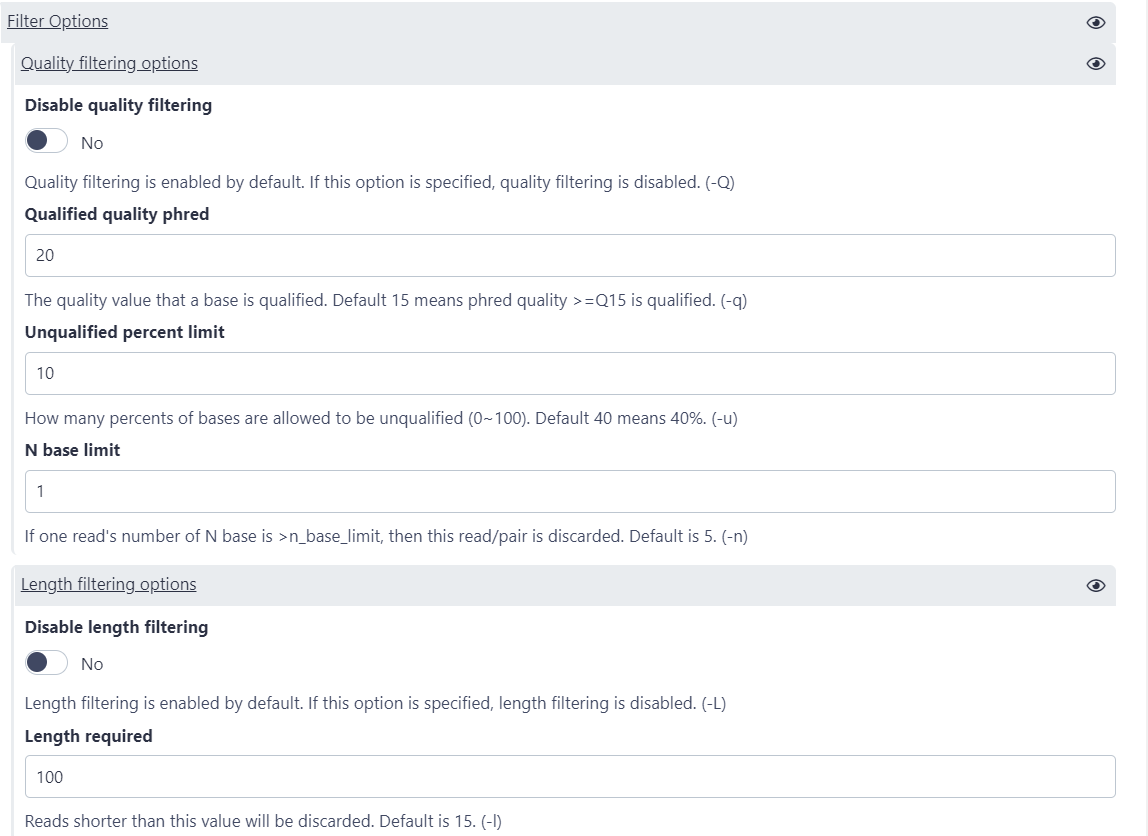
By Wei Guo, 4/14/2021, libinglin

**One. STR analysis**

1. Extract genome DNA from a cell line.
2. PCR Amplify the targeted STRs loci listed in step 8 for Target NGS.
3. Once NGS data is back, run ***FastQC*** to check fastq data quality.
4. Run ***fastp*** to trim or remove adaptors and low quality reads.

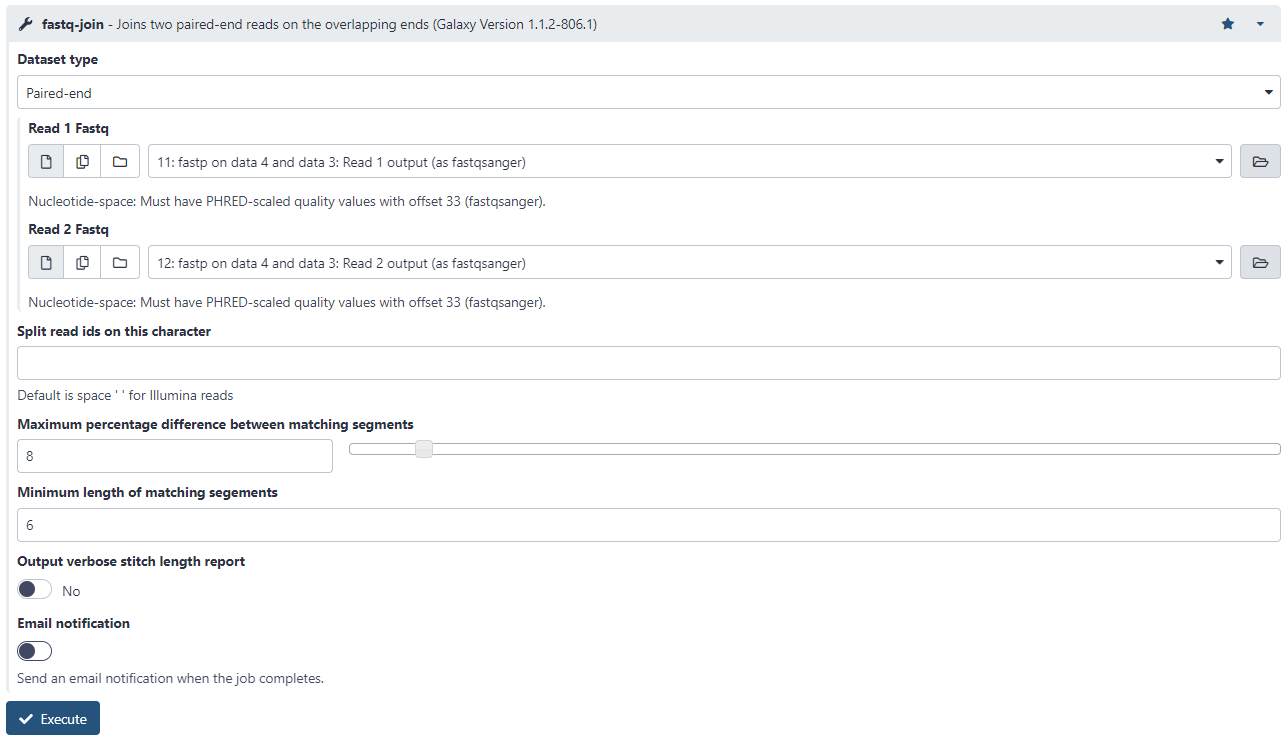




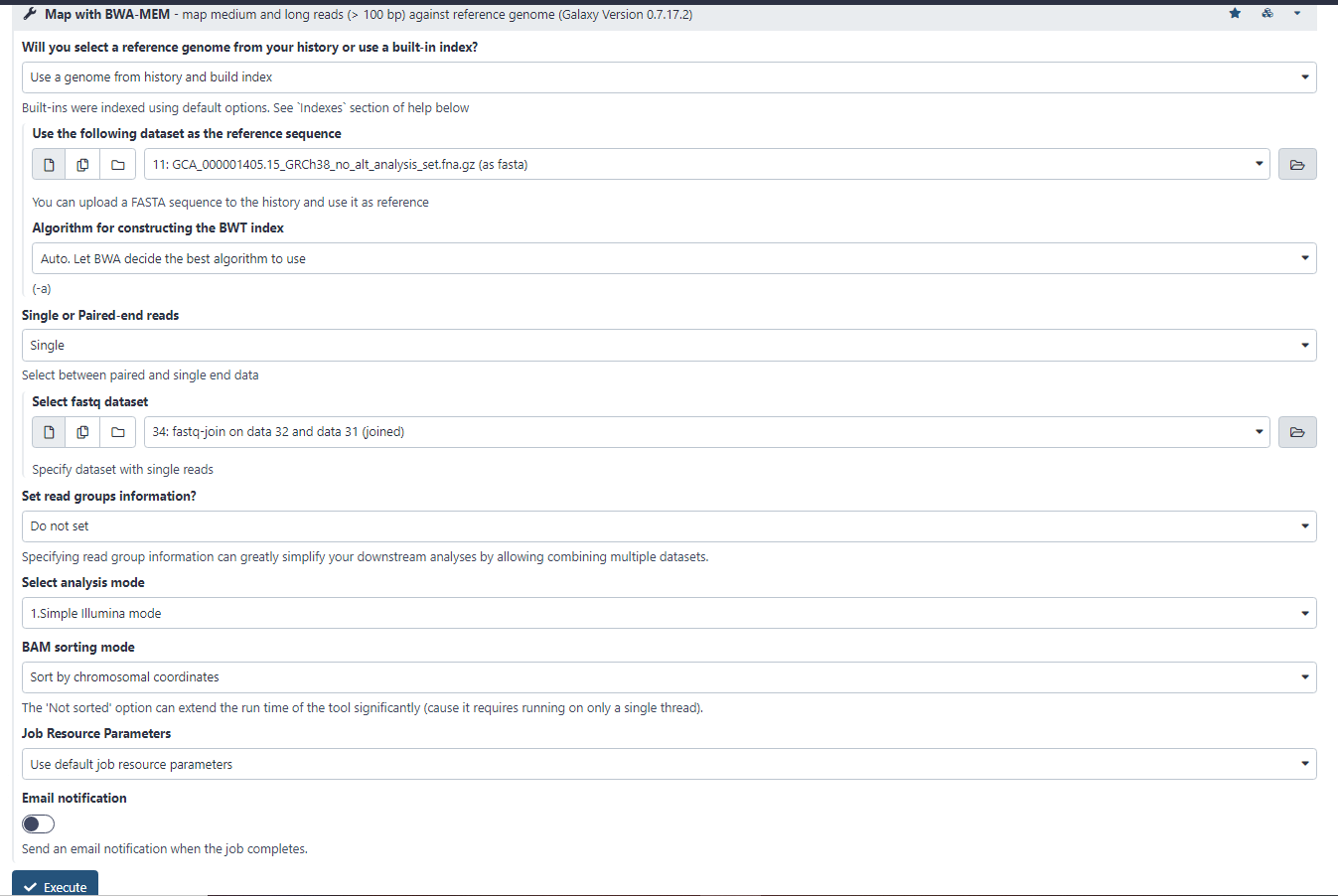


20 to 30, [25]

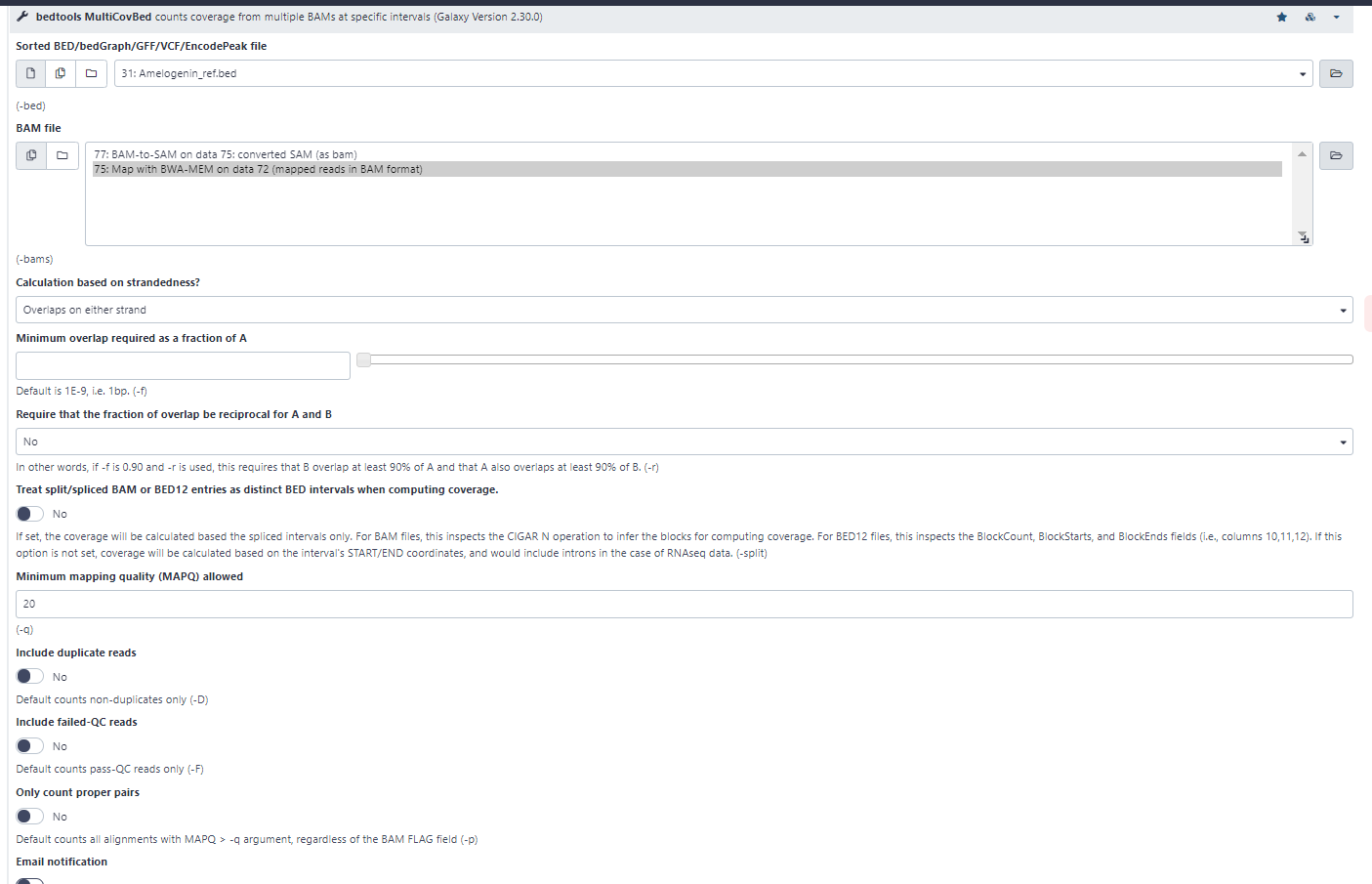
1. Run **fastq-join or SeqPrep** to merge paired-end reads



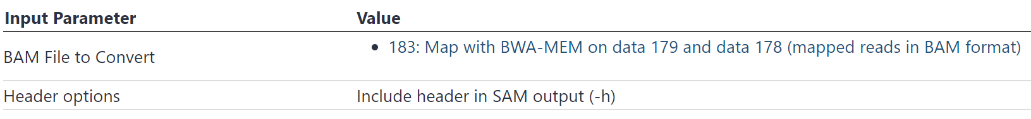
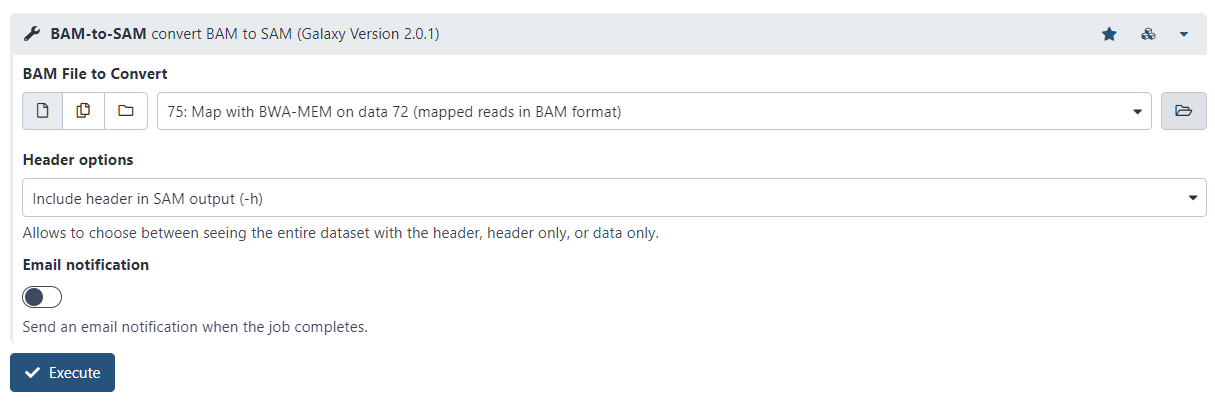
1. Run ***Map with BWA-MEM*** to map reads to human reference genome (default hg38 from UCSC for human).



1. Run ***bedtools MultiCovBed*** to check if the Amelogenin gene region (chrY:6869821-6870058, or chrX:11296780-11296999) in the BAM file from step 5 is covered by reads. Note：the Amelogenin\_ref.bed file is available in the reference genome shared folder.



1. Run ***BAM-to-SAM*** to convert the BAM file format of the mapped reads by ***BWA-MEM*** to SAM format.

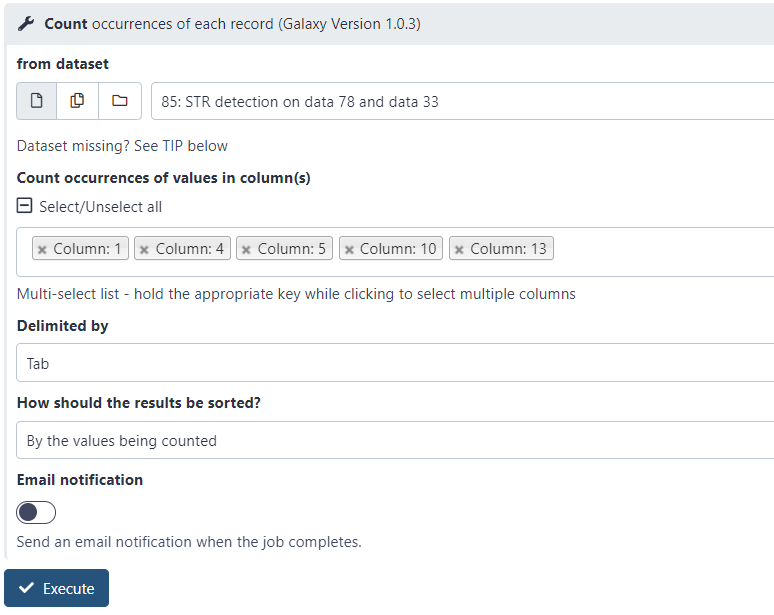


1. Run ***STR detection*** with supplied reference genome (download from UCSC hg38 for human or available in our shared restricted folder) to analyze STR profiles for targeted loci on a genome.

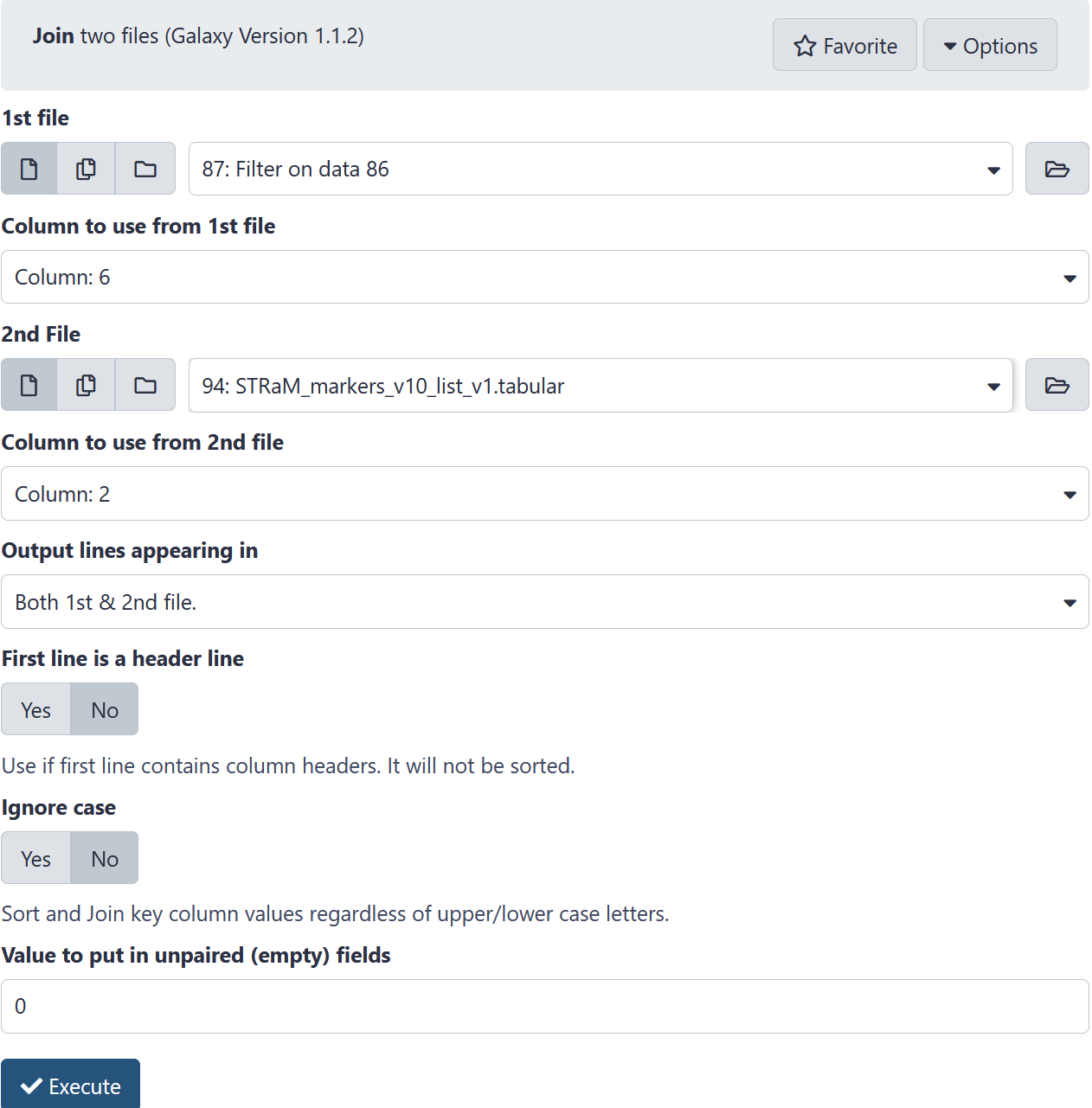
Note： Hamming distance (hamming threshold of microsatellites) needs to be adjusted according to the STR markers



1. Run ***Count*** *occurrences of each record* to count reads on individual STR.



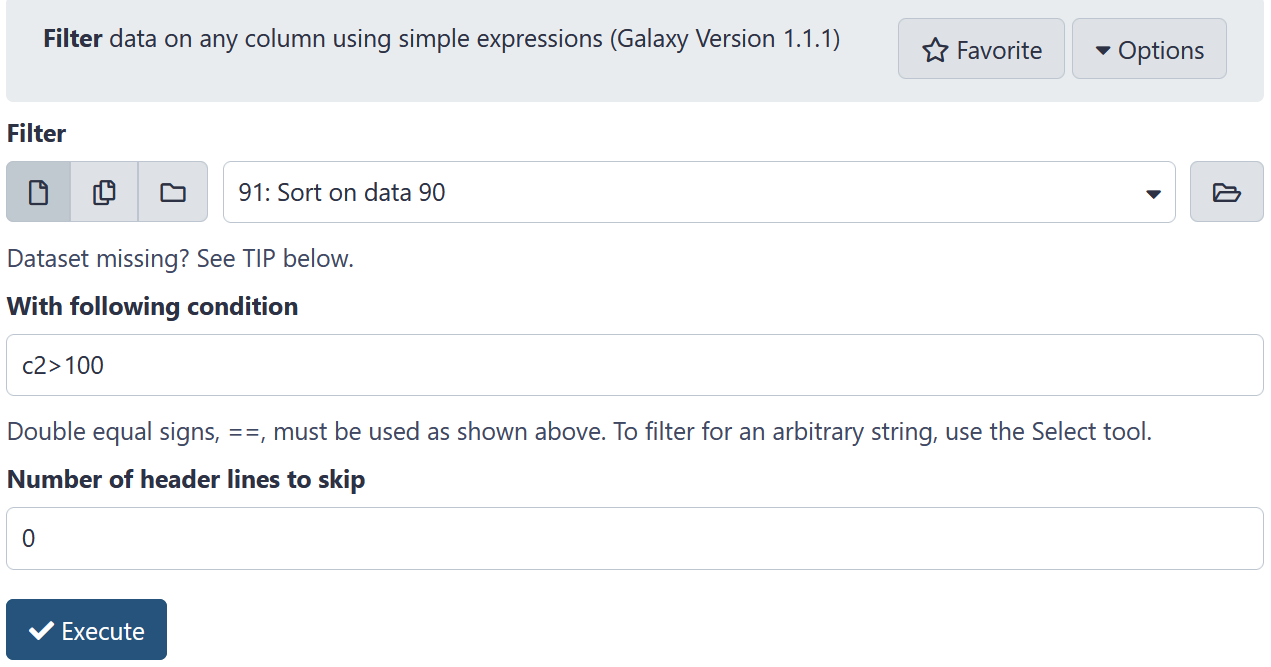
1. Identify the STRaM markers by ***join***ing (comparing) the Marker list file with the count file from step 10. The program is located in the Galaxy **Text Manipulation** section. The marker list file is available in our Lab restricted shared:\Bioinformatics\reference genome\STR.



1. Run **Sort** data in ascending or descending order (for 3 columns) to list data with chromosomes and read counts. The program is located in the Galaxy **Text Manipulation** section.

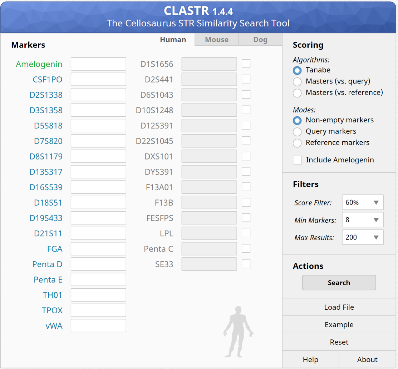
|  |
| --- |
|  |
|  |
|  |

1. Run ***Filter*** *data on any column using simple expressions* to exclude counts <100 (depending on the NGS data). The program is located in the Galaxy **Filter and Sort** section



1. Compare the STR data from STR analysis from step 13 to those obtained from the flank analysis (a separate protocol). **If The STR nucleotide counts of a marker between STR analysis and flank analysis are not equal or missing, the sequences of STR or flanking ends should be validated manually and STR repeat counts should be corrected.**

|  |  |
| --- | --- |
| **STR analysis** | **Flank analysis** |
|  |  |
| STR nucleotide counts in column 3. No. of STR repeats = (STR nucleotide counts)/4.  Read counts for each STR in column 2. | STR nucleotide counts in column 4. No. of STR repeats = (STR nucleotide counts)/4.  Read counts for each STR in column 1. |

1. Assemble STR profile table for a cell line.
2. Identify the cell lines by searching the online STR database on Cellosaurus (<https://web.expasy.org/cellosaurus-str-search/>) or STRaM database (web site is under construction).

STR file format:

* Column 1 = length of STR (bp)
* Column 2 = length of left flanking region (bp)
* Column 3 = length of right flanking region (bp)
* Column 4 = repeat motif (bp)
* Column 5 = hamming distance
* Column 6 = read name
* Column 7 = read sequence with soft masking of STR
* Column 8 = read quality (the same Phred score scale as input)
* Column 9 = read name (The same as column 6)
* Column 10 = chromosome
* Column 11 = left flanking region start
* Column 12 = left flanking region stop
* Column 13 = STR start as infer from pair-end
* Column 14 = STR stop as infer from pair-end
* Column 15 = right flanking region start
* Column 16 = right flanking region stop
* Column 17 = STR length in reference
* Column 18 = STR sequence in reference