université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

IFSBM « Big Data »

RNA-seq
Differential expression analysis pipeline

Daniel Gautheret

V.2023

Notre jeu de données « EMT »

EMT= Epithelial-Mesenchymal transition

Source:

AMERICAN SOCIETY FOR MICROBIOLOGY | Molecular and Cellular Biology

Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition

Yueqin Yang,[a,b] Juw Won Park,[c,d,e] Thomas W. Bebee,[a,b] Claude C. Warzecha,[a,b*] Yang Guo,[c,f] Xuequn Shang,[f] Yi Xing,[c] Russ P. Carstens[a,b]

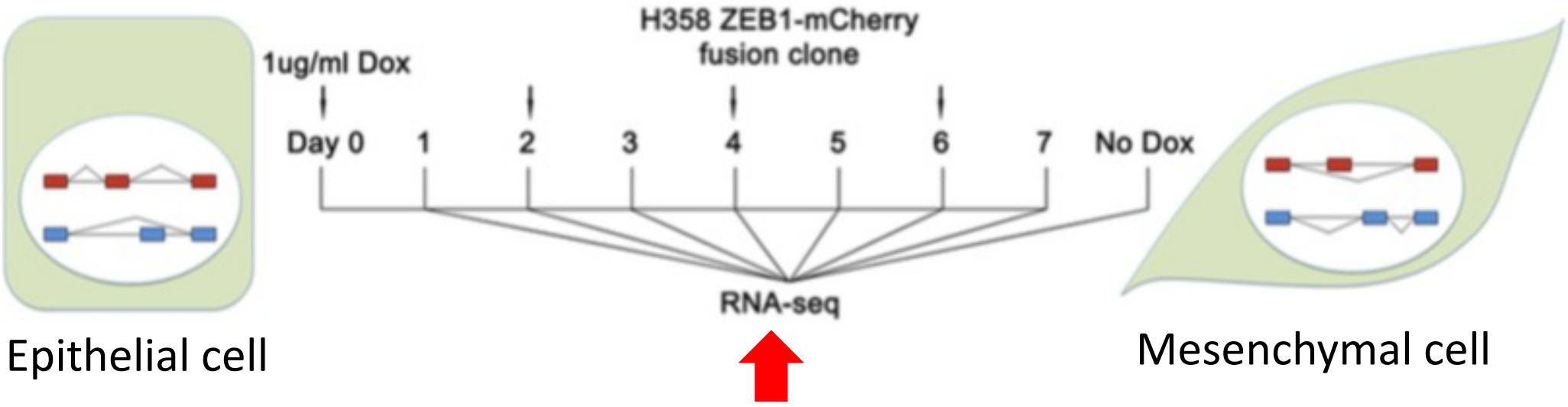
Departments of Genetics[a] and Medicine,[b] Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, Los Angeles, California, USA[c]; Department of Computer Engineering and Computer Science[d] and KBRIN Bioinformatics Core,[e] University of Louisville, Louisville, Kentucky, USA; School of Computer Science, Northwestern Polytechnical University, Xi'an, China[f]

Expérience: (sur lignée non-small cell lung cancer H358)

E

Facteur de transcription:
Induit l'EMT

M



Epithelial cell

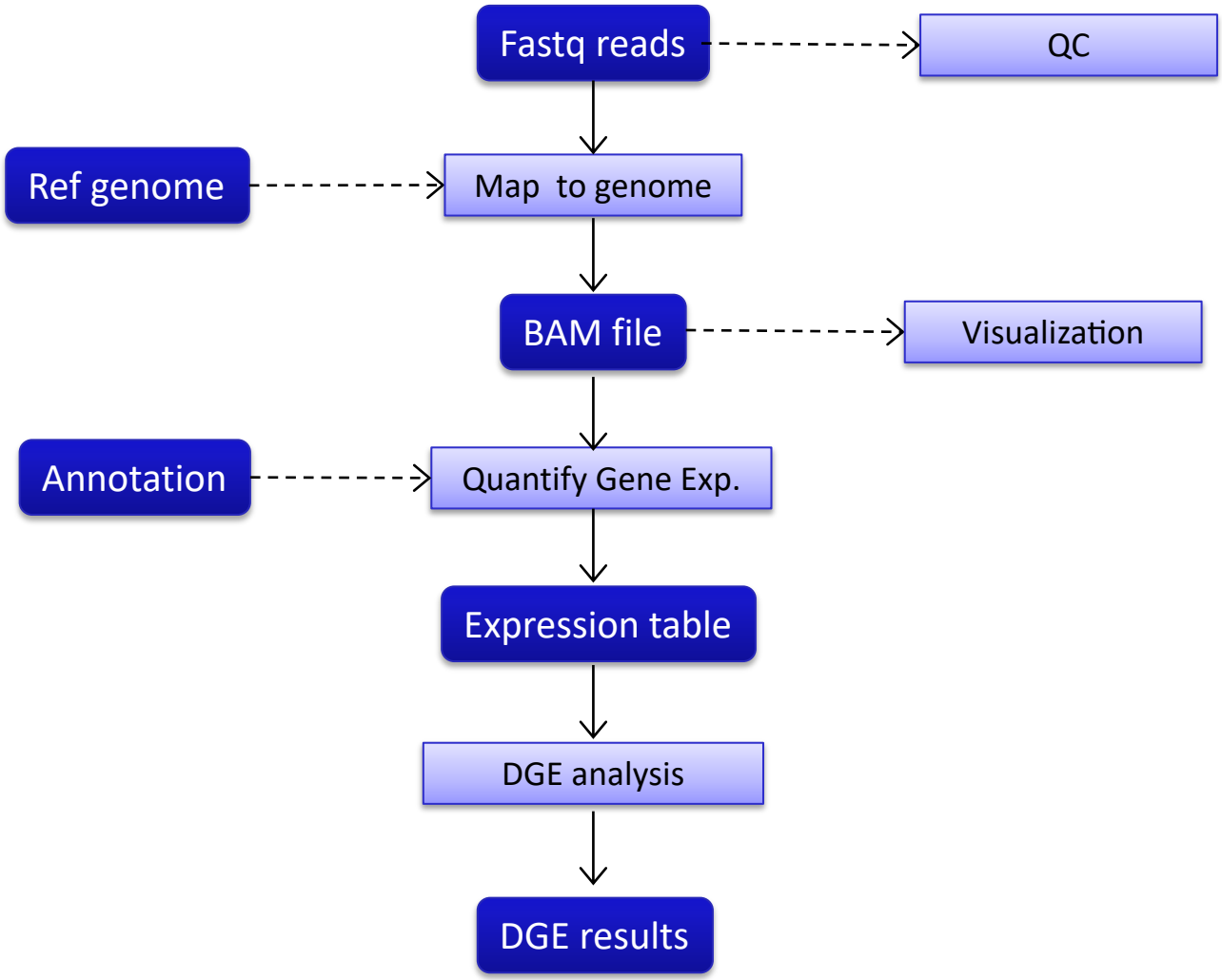
Mesenchymal cell

Triplicat à chaque point

Question

- Quels sont les gènes différentiellement exprimés au cours de l'EMT?

Transcriptomics protocole



Data

- Sequence libraries are polyA+, pair-end 2x100nt, each in biological triplicate.
- Sequencing is performed on a Illumina HiSeq 2500.
- Fastq files were obtained here: http://www.ncbi.nlm.nih.gov/sra?term=SRP066794

Subsampling

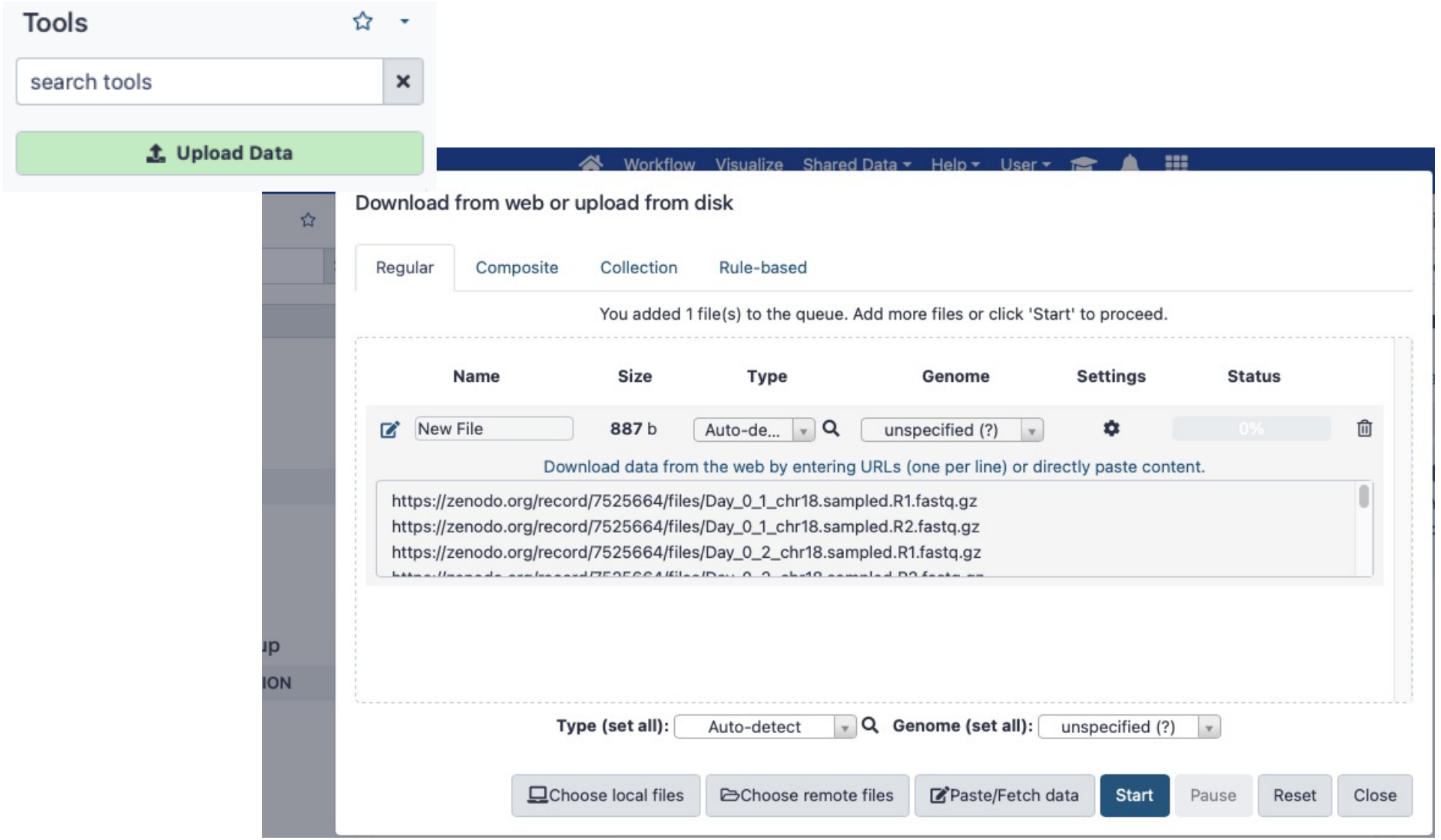
- Initial fastq files: 72Mx2 reads
- Subsampling
  - Reads mapped to HG38 with STAR
  - Select Chr18 : 685,000 x2 reads
  - Sampled by a factor of 0.5 (Samtools) : 343,000 x2 reads

This represents 0.5% of total reads, thus actual runtimes and space requirement would be up to 200 times higher than in our exercices.

Liens pour fastq

https://zenodo.org/record/7525664/files/Day_0_1_chr18.sampled.R1.fastq.gz
https://zenodo.org/record/7525664/files/Day_0_1_chr18.sampled.R2.fastq.gz
https://zenodo.org/record/7525664/files/Day_0_2_chr18.sampled.R1.fastq.gz
https://zenodo.org/record/7525664/files/Day_0_2_chr18.sampled.R2.fastq.gz
https://zenodo.org/record/7525664/files/Day_0_3_chr18.sampled.R1.fastq.gz
https://zenodo.org/record/7525664/files/Day_0_3_chr18.sampled.R2.fastq.gz
https://zenodo.org/record/7525664/files/Day_7_1_chr18.sampled.R1.fastq.gz
https://zenodo.org/record/7525664/files/Day_7_1_chr18.sampled.R2.fastq.gz
https://zenodo.org/record/7525664/files/Day_7_2_chr18.sampled.R1.fastq.gz
https://zenodo.org/record/7525664/files/Day_7_2_chr18.sampled.R2.fastq.gz
https://zenodo.org/record/7525664/files/Day_7_3_chr18.sampled.R1.fastq.gz
https://zenodo.org/record/7525664/files/Day_7_3_chr18.sampled.R2.fastq.gz

Download fastq files



Protocol: quality control

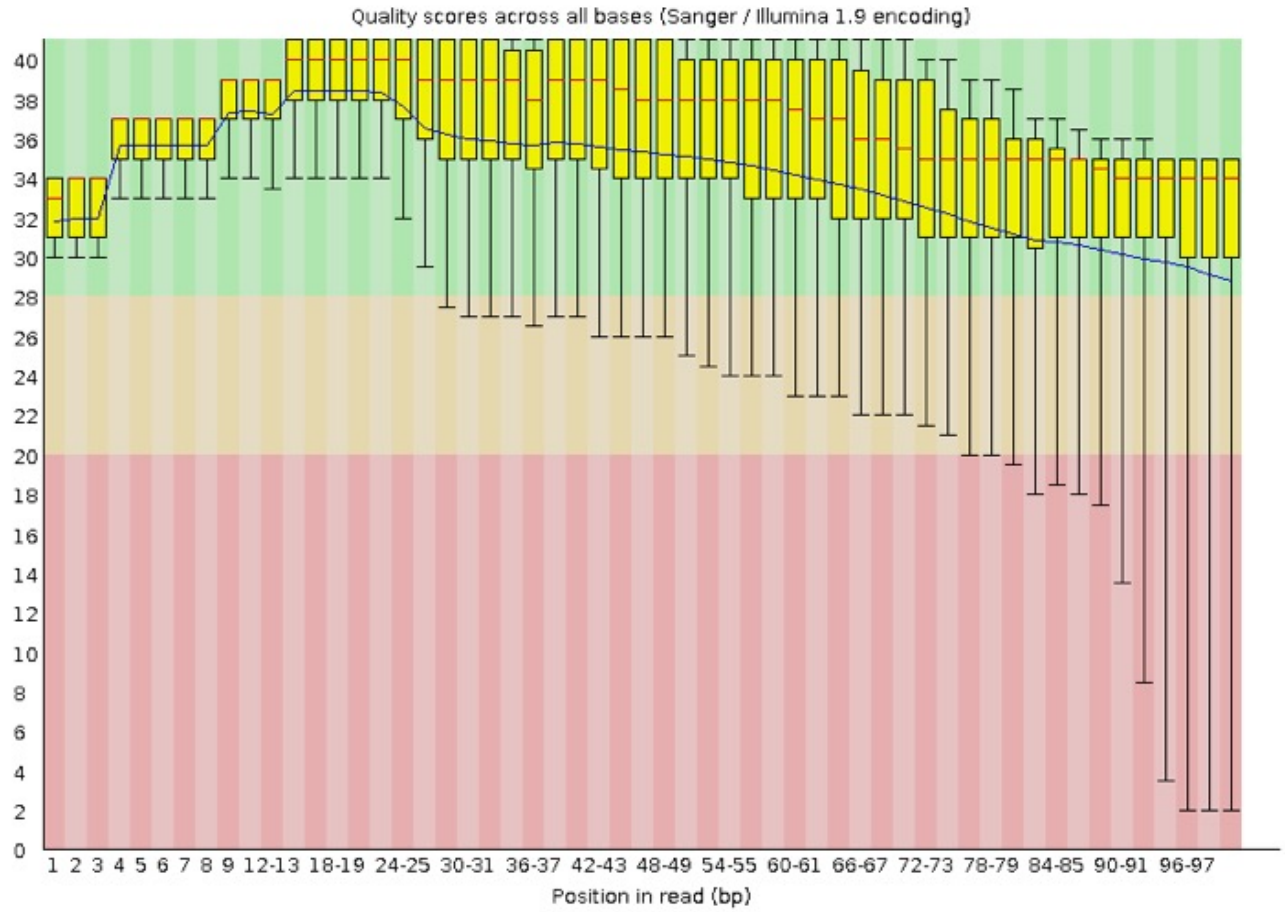
- Fastqc on all samples
  - Check one HTML report
- Multiqc on raw
  - Just check HTML report.

FastQC

Basic Statistics

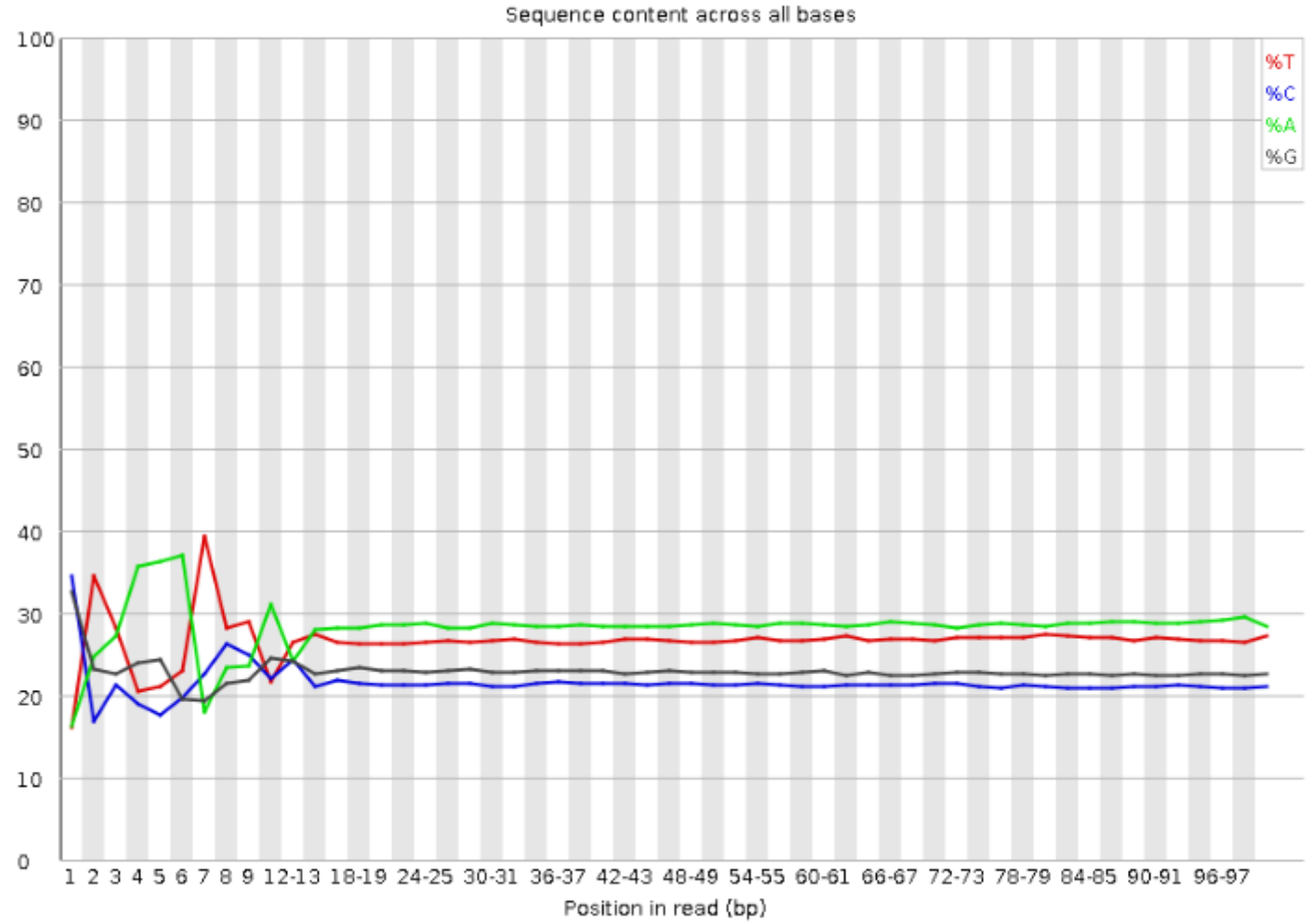
Measure	Value
Filename	Day_7_1_chr18_sampled_R2_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	235214
Sequences flagged as poor quality	0
Sequence length	100
%GC	44

Per base sequence quality



FastQC

Per base sequence content



- Paramètres multiQC

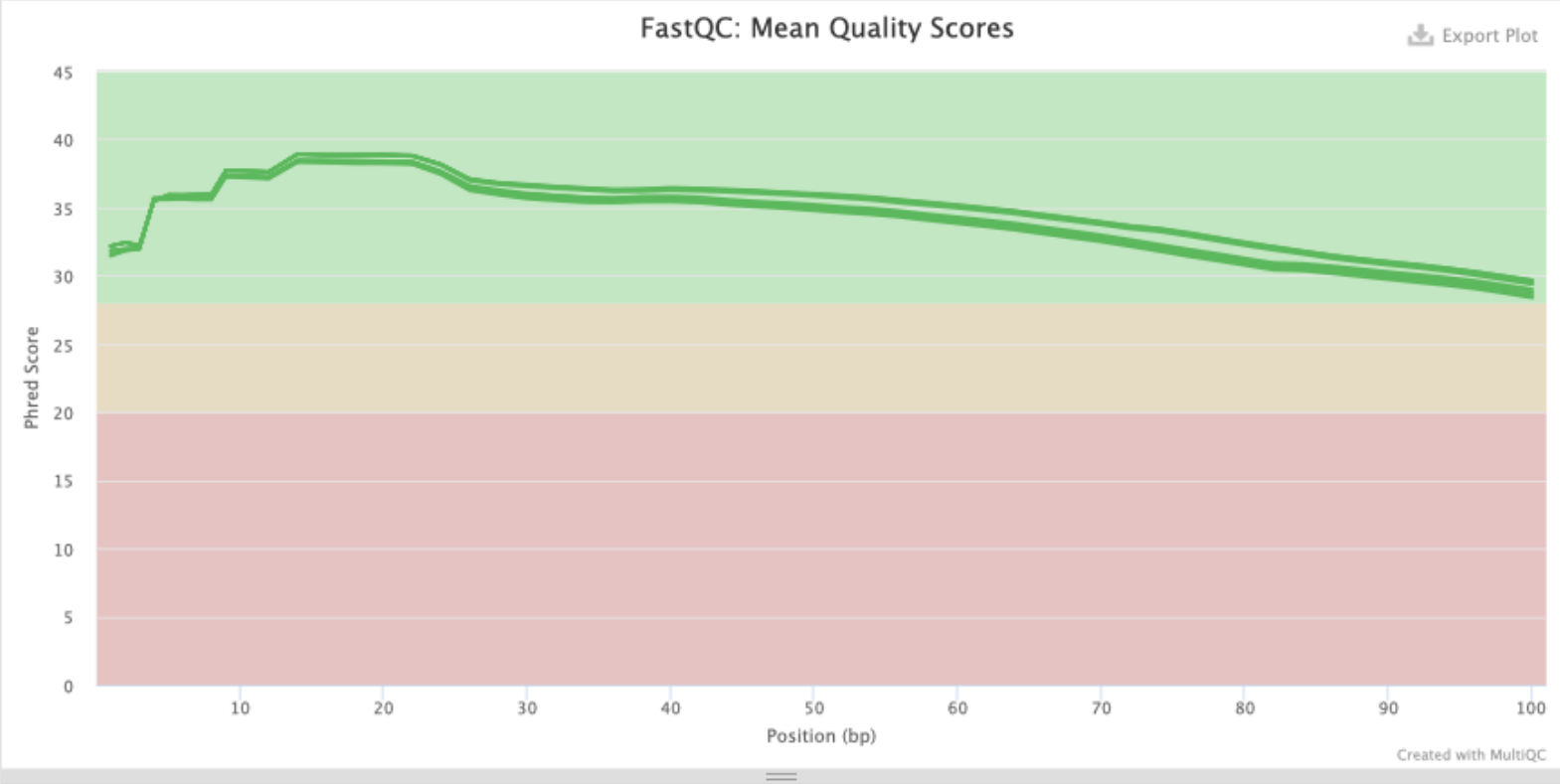
MultiQC

Sequence Quality Histograms 12

Help

The mean quality value across each base position in the read.

Y-Limits: on



Protocol: mapping

- STAR on 1 sample (R1+R2)
  - check log file
  - Visualize BAM file
- Download BAM and BAM index on your computer
- Visualize in IGV
- Data is stranded, R1 in reverse

STAR
(PE mode)

RNA STAR Gapped-read mapper for RNA-seq data (Galaxy Version 2.7.8a+galaxy1)

Single-end or paired-end reads

Paired-end (as individual datasets)

RNA-Seq FASTQ/FASTA file, forward reads

5: Day_0_1_chr18.sampled.R1.fastq.gz

RNA-Seq FASTQ/FASTA file, reverse reads

6: Day_0_1_chr18.sampled.R2.fastq.gz

Custom or built-in reference genome

Use a built-in index

Built-ins were indexed using default options

Reference genome with or without an annotation

use genome reference without builtin gene-model

Select the '... with builtin gene-model' option to select from the list of available indexes that were built with splice junction information. Select the '... without builtin gene-model' option to select from the list of available indexes without annotated splice junctions, and, optionally, provide your own splice-junction annonations.

Select reference genome

Human Dec. 2013 (GRCh38/hg38) (hg38)

If your genome of interest is not listed, contact the Galaxy team (--genomeDir)

Gene model (gff3,gtf) file for splice junctions

No gff3 or gtf dataset available.

Exon junction information for mapping splices (--sjdbGTFfile)

Length of the genomic sequence around annotated junctions

100

Used in constructing the splice junctions database. Ideal value is ReadLength-1 (--sjdbOverhang)

Use 2-pass mapping for more sensitive novel splice junction discovery

No

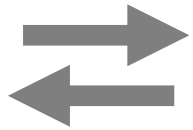
For a study with multiple samples, multisample 2-pass mapping is the most sensitive approach. It involves two separate runs of STAR for each sample, where, in the second run of each sample, the splice junctions found in any sample in the first runs are treated as additional known junctions. If you plan to use the mapping results as input for STAR-Fusion it is recommended that you use at least single-sample 2-pass mapping of all reads. (--twopassMode)

Per gene/transcript output

No per gene or transcript output

Fichier log de STAR

Started job on	Jan 11 17:32:50
Started mapping on	Jan 11 17:42:26
Finished on	Jan 11 17:42:48
Mapping speed, Million of reads per hour	56.10
Number of input reads	342819
Average input read length	200
UNIQUE READS:	
Uniquely mapped reads number	321996
Uniquely mapped reads %	93.93%
Average mapped length	197.43
Number of splices: Total	157996
Number of splices: Annotated (sjdb)	0
Number of splices: GT/AG	157155
Number of splices: GC/AG	678
Number of splices: AT/AC	9
Number of splices: Non-canonical	154
Mismatch rate per base, %	0.25%
Deletion rate per base	0.01%
Deletion average length	1.47
Insertion rate per base	0.01%
Insertion average length	1.28
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	8703
% of reads mapped to multiple loci	2.54%
Number of reads mapped to too many loci	40
% of reads mapped to too many loci	0.01%
UNMAPPED READS:	
Number of reads unmapped: too many mismatches	0
% of reads unmapped: too many mismatches	0.00%
Number of reads unmapped: too short	11857
% of reads unmapped: too short	3.46%
Number of reads unmapped: other	223
% of reads unmapped: other	0.07%
CHIMERIC READS:	
Number of chimeric reads	0
% of chimeric reads	0.00%

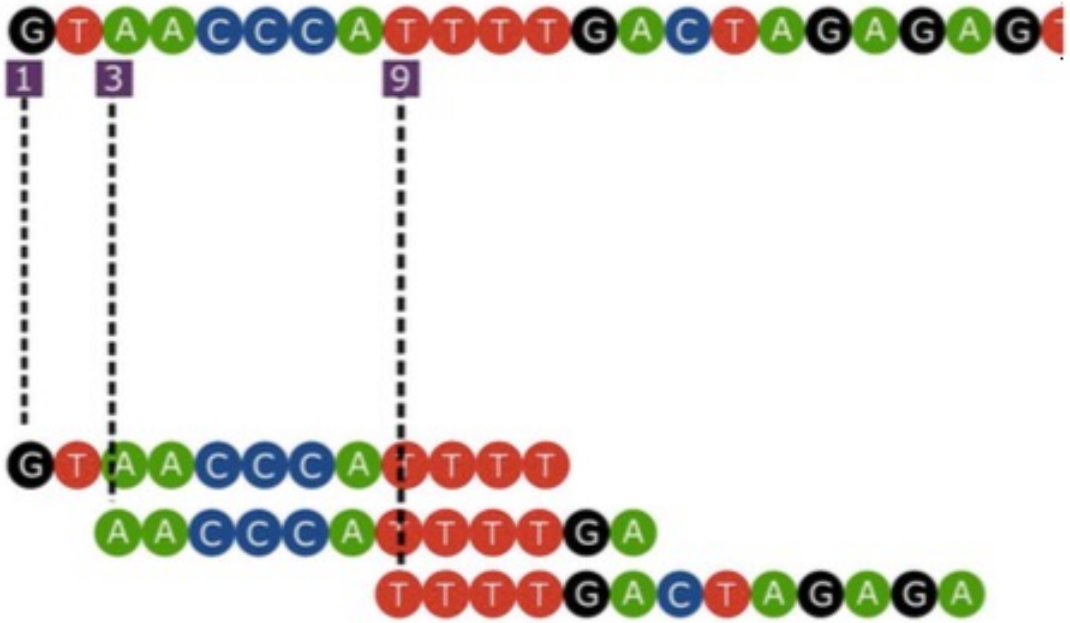
SAM  BAM

SAMtools

Fichier texte

Fichier binaire

Format SAM/BAM
(séquences alignées sur une référence)



Information minimale:

```
chr7 1324324    ACGTGCGTTTGCGT
chr8 1724354    GCGTGATGCGTAAG
chr8 1424324    GTATGTTATATGTA
```

SAM/BAM file

@SQ SN:chrY_KI270740v1_random LN:37240					
@PG ID:STAR PN:STAR VN:2.7.8a CL:STAR --runThreadN 10 --genomeDir /data/db/data_managers/rnastar/2.7.4a/hg38/hg38/dataset_412f3413-7					
@CO user command line: STAR --runThreadN 10 --genomeLoad NoSharedMemory --genomeDir /data/db/data_managers/rnastar/2.7.4a/hg38/hg					
Day_0_1.50295732.1	83	chr1	10411512	60	9S41M1509N50M
Day_0_1.447494.1	163	chr1	10411513	60	40M1509N50M10S
Day_0_1.447494.1	83	chr1	10560745	60	15S85M
Day_0_1.42485259.1	355	chr1	44122139	3	2S98M
Day_0_1.12613904.1	99	chr1	44122147	60	100M
Day_0_1.33247661.1	99	chr1	44122155	60	100M
Day_0_1.2387938.1	99	chr1	44122157	60	100M
Day_0_1.4275681.1	99	chr1	44122157	60	100M
Day_0_1.7914668.1	99	chr1	44122157	60	100M
Day_0_1.10596929.1	99	chr1	44122157	60	100M
Day_0_1.21350327.1	99	chr1	44122157	60	100M
Day_0_1.70718417.1	99	chr1	44122157	60	100M
Day_0_1.440151.1	99	chr1	44122158	3	100M
Day_0_1.768078.1	99	chr1	44122158	3	100M
Day_0_1.7215815.1	99	chr1	44122158	3	100M
Day_0_1.13017033.1	99	chr1	44122158	3	100M
Day_0_1.26195406.1	99	chr1	44122158	60	100M
Day_0_1.28318160.1	99	chr1	44122158	60	100M

Format SAM/BAM

- A real SAM:

```
@RG     ID:group1       SM:1425_CD34    PL:ILLUMINA     LB:lib1 PU:unit1
@PG     ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa mem -M -t 2 -A 2 -E 1 -R @RG\tID:group1\tSM:1425_CD34\tPL:ILLUMINA\tLB:lib1\tPU:unit1 /root/myd
ERR166338.13782800      83      chr13   32890449        60      101M    =       32890343        -207    GGGACTGAATTAGAATTCAAACAAATTTTCCAGCGCTT
ERR166338.13782800      163     chr13   32890343        60      75M     =       32890449        207     CACTAGCCACGTTTCGAGTGCTTAATGTGGCTAGTGGC
ERR166338.26716588      99      chr13   32890406        60      101M    =       32890553        222     AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAG
ERR166338.26716588      147     chr13   32890553        60      75M     =       32890406        -222    TTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTA
ERR166338.27259961      99      chr13   32890496        60      101M    =       32890558        137     ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.27259961      147     chr13   32890558        60      75M     =       32890496        -137    GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA
ERR166338.63037998      99      chr13   32890496        60      101M    =       32890558        137     ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.63037998      147     chr13   32890558        60      75M     =       32890496        -137    GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA
```

read ID

flag

position

mapping qual.

CIGAR

mate info

Les Flags

Example:

- Decimal Flag Value 83
- Binary Flag Value

2048	1024	512	256	128	64	32	16	8	4	2	1
0	0	0	0	0	1	0	1	0	0	1	0

12 bits

- To each bit corresponds a meaning

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Lexique:

- Segment= a continuous aligned part of a read
- SEQ=the sequence of a segment
- Template= a DNA strand being sequenced (le fragment entier)

Le champ CIGAR

Example: 52M36890N45M3S

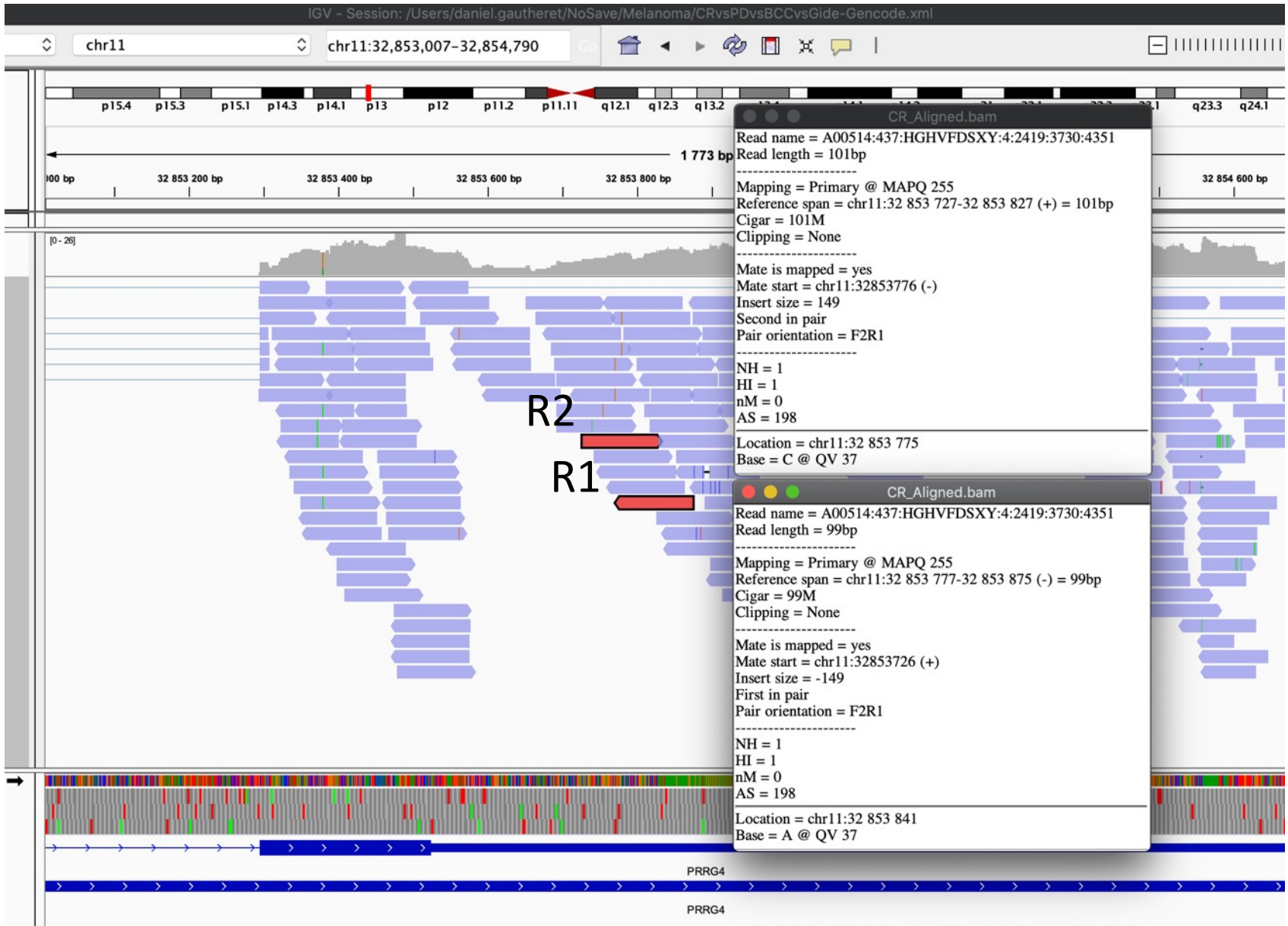
All Cigar operations

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

IGV

- Online version:
  - Tracks + local file + selectionner les 2 fichiers (BAM et BAI) avec la souris

Comprendre alignement et orientation des reads avec IGV



Loci intéressants (voir avec IGV)

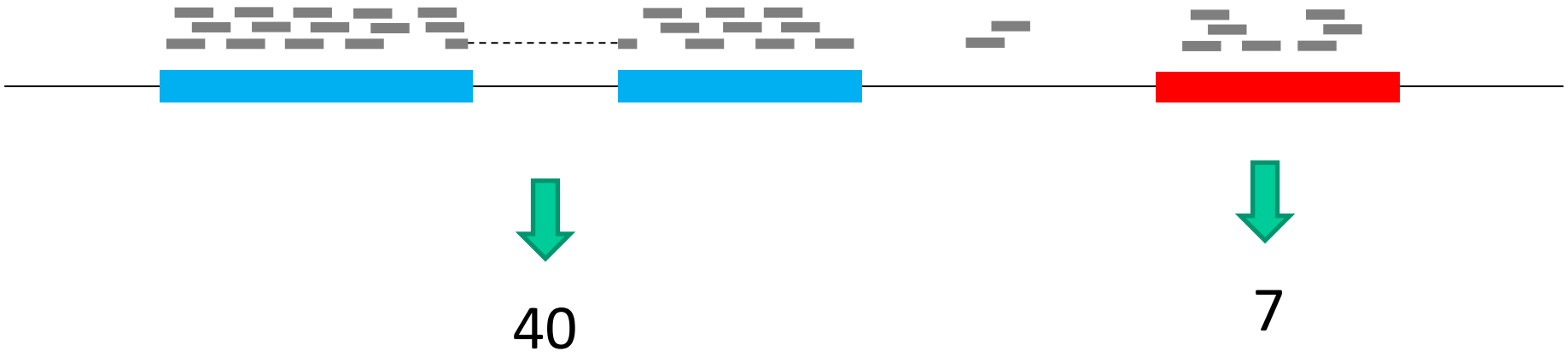
- chr18:2,740,603-2,740,845 (SNP)
- chr18:9,256,014-9,256,135 (indels)
- Cas d'épissage alternatif dans les gènes C18orf21 (saut exon 2), SLC39A6 (saut exon 2)
- De nouveaux gènes non annotés : chr18:32,852,613-32,868,170, chr18:29,139,651-29,170,765

Protocol: quantification

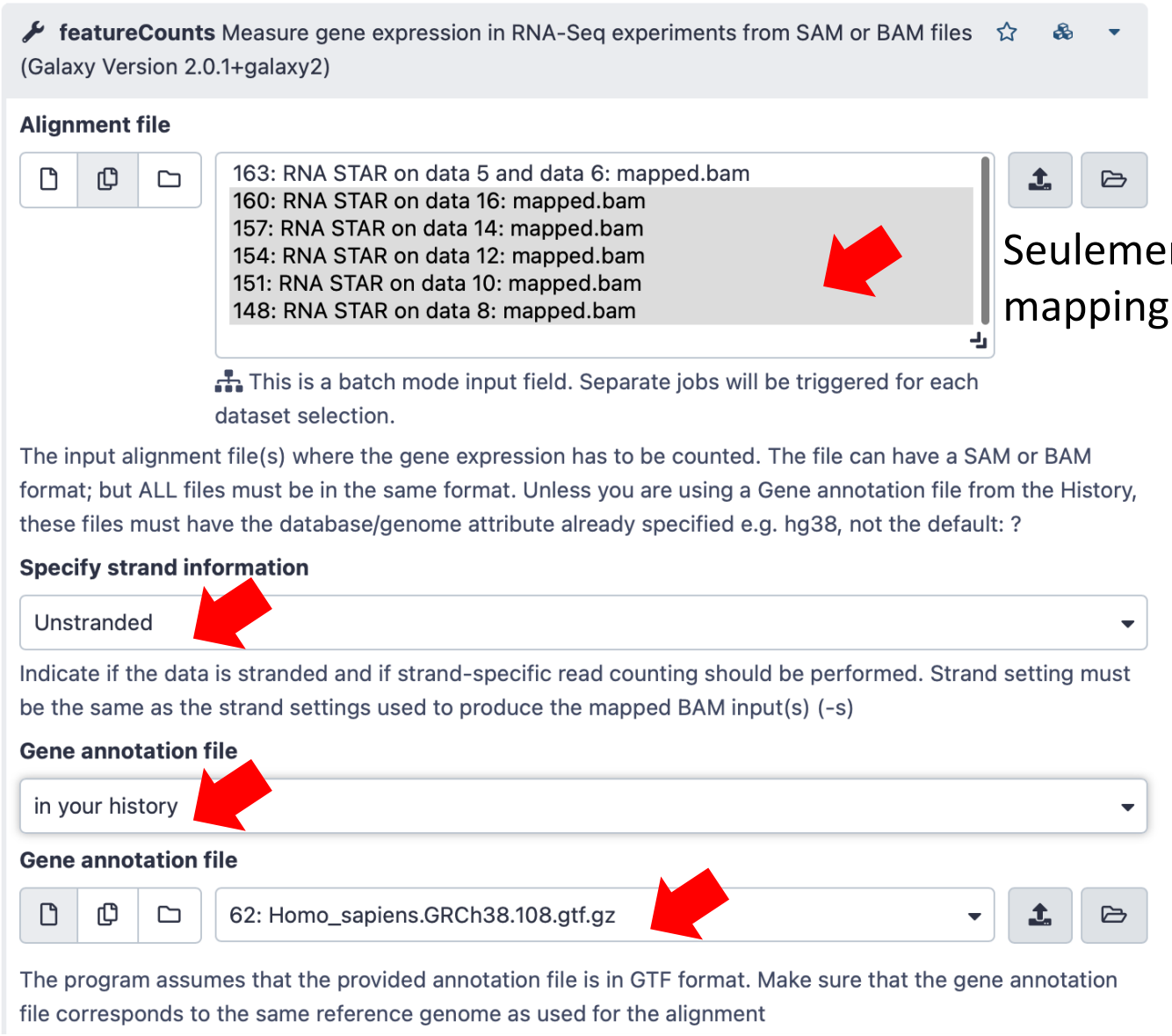
- STAR on all samples (just use R2 for speed)
  - Check alignment results (see log files)
- Upload Annotation (gtf) file for human
  - https://ftp.ensembl.org/pub/release-108/gtf/homo_sapiens/Homo_sapiens.GRCh38.108.gtf.gz
- Feature count on 6 BAM
- MultiQC on FeatureCount summary (6 files)
- Column join > Table with 7 columns

STAR (SE mode + multifiles)

FeatureCounts



FeatureCount parameters



Seulement les résultats du mapping x6

FeatureCount multiQC

General Statistics

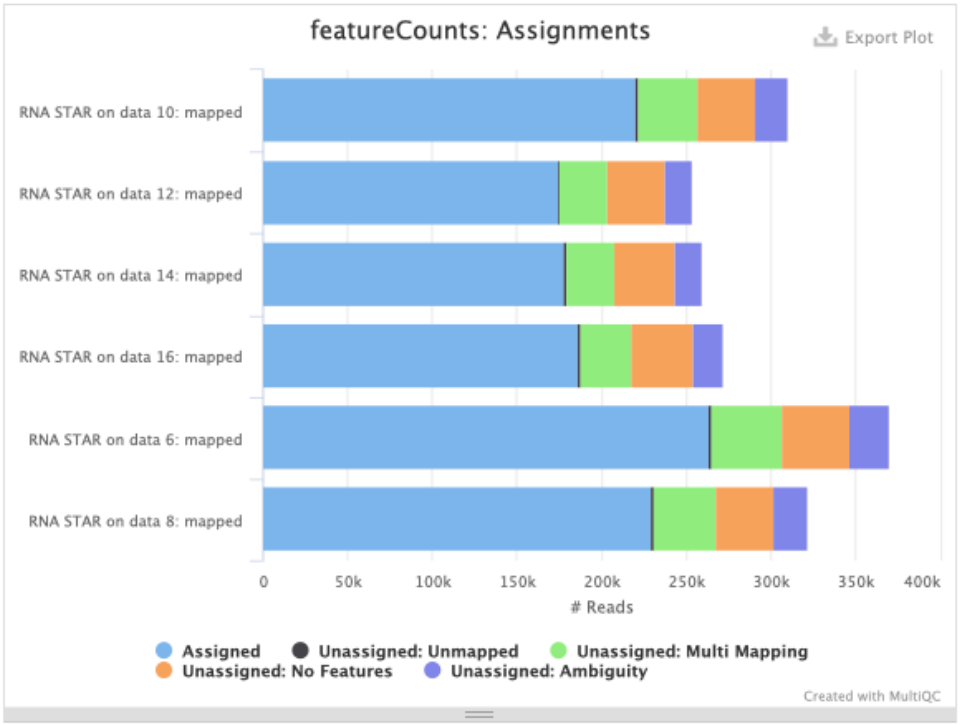
Copy table | Configure Columns | Plot | Showing 6/6 rows and 2/2 columns.

Sample Name	% Assigned	M Assigned
RNA STAR on data 10: mapped	71.1%	0.2
RNA STAR on data 12: mapped	68.7%	0.2
RNA STAR on data 14: mapped	68.6%	0.2
RNA STAR on data 16: mapped	68.7%	0.2
RNA STAR on data 6: mapped	71.3%	0.3
RNA STAR on data 8: mapped	71.3%	0.2

featureCounts

Subread featureCounts is a highly efficient general-purpose read summarization program that counts mapped reads for genomic features such as genes, exons, promoter, gene bodies, genomic bins and chromosomal locations.

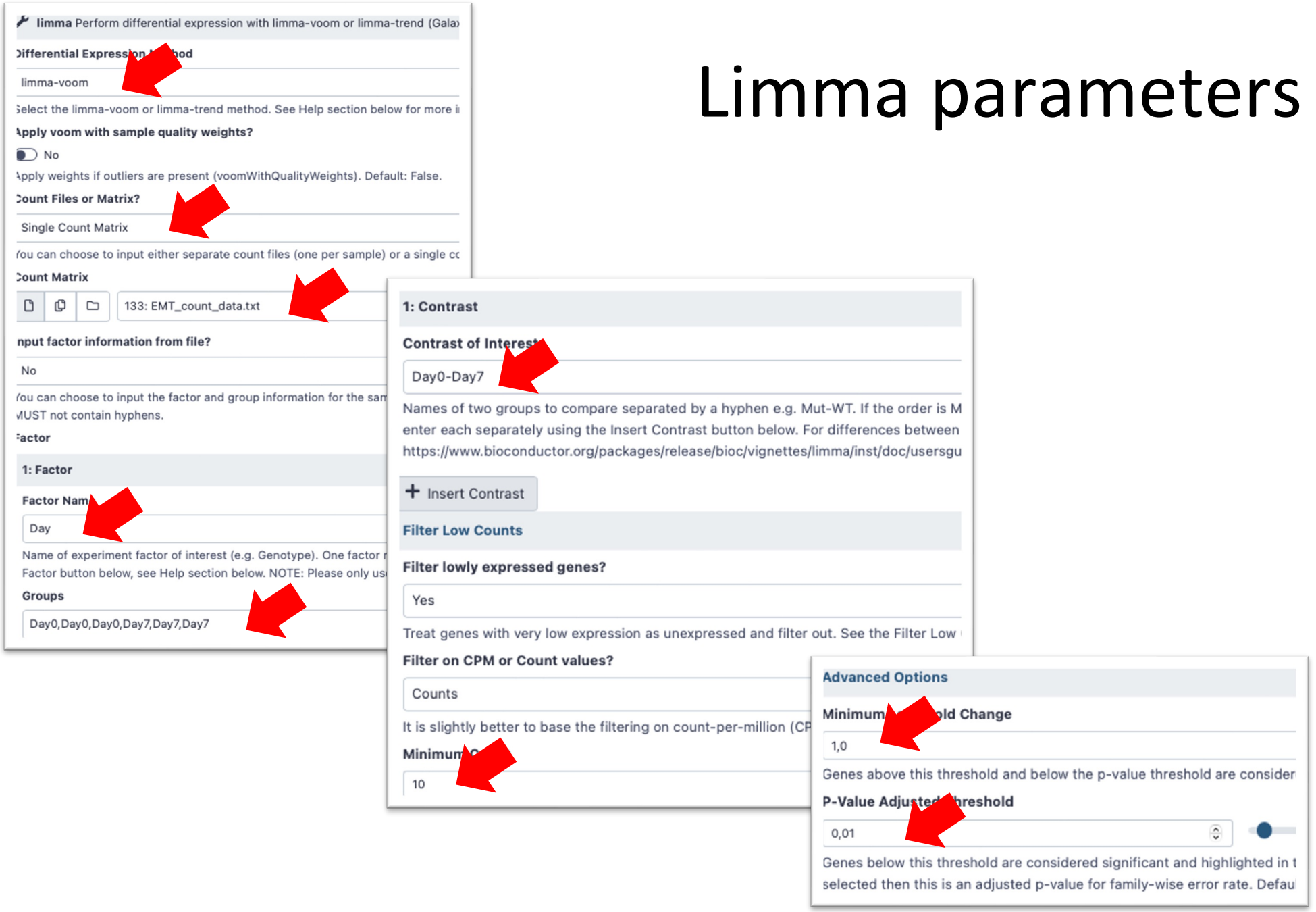
Number of Reads | Percentages



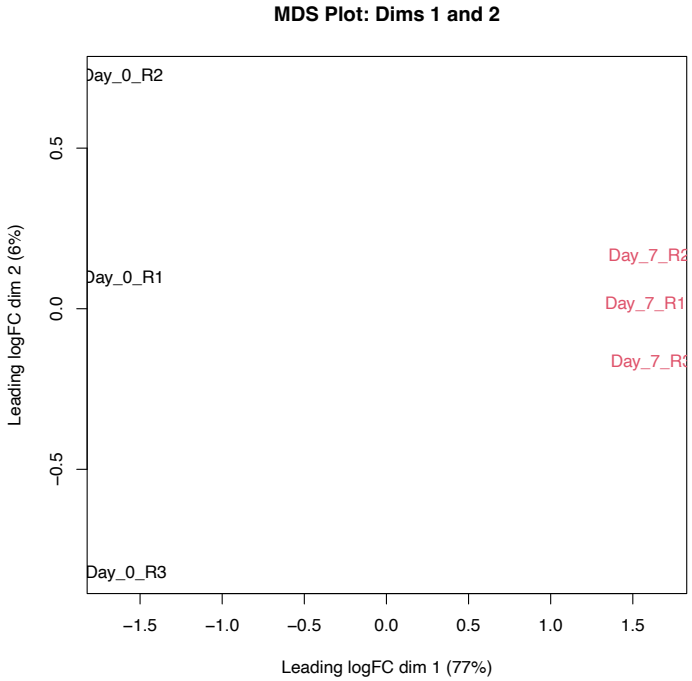
Protocol: differential expression

- Upload complete count file for EMT experiment (from https://github.com/gustaveroussy/IFSBM-bigdata)
  - Upload / tsv format: EMT_count_data.txt
  - Check it
- Limma
  - Use Voom mode (for RNA-seq)

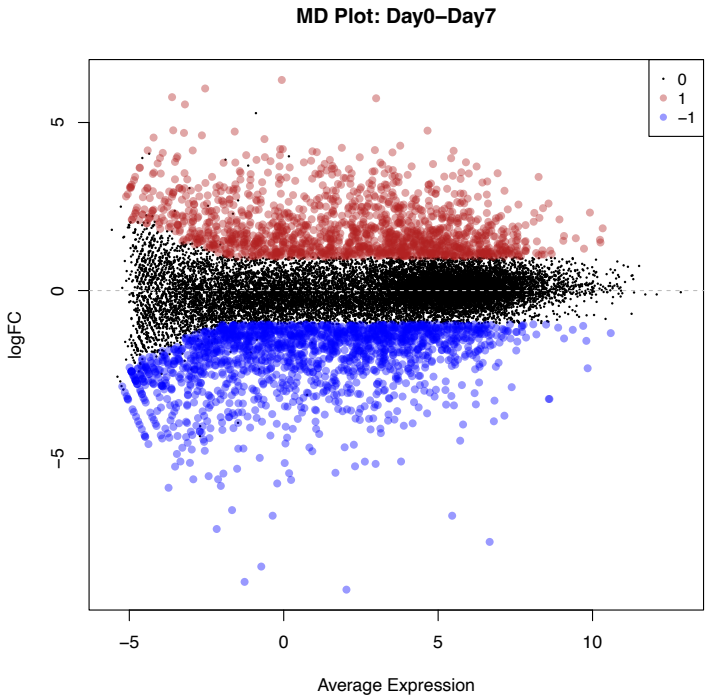
Limma parameters

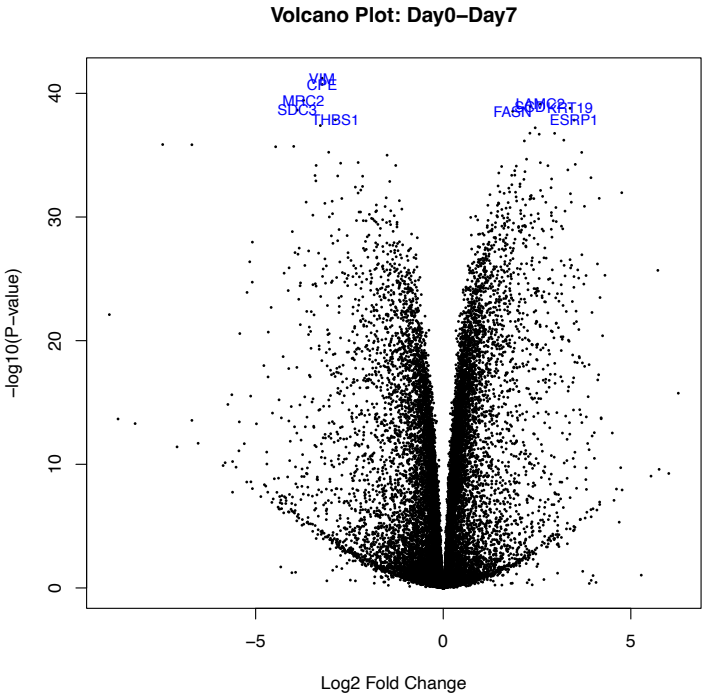


Graphiques importants



(AKA: MA plot)





Most significant DE genes



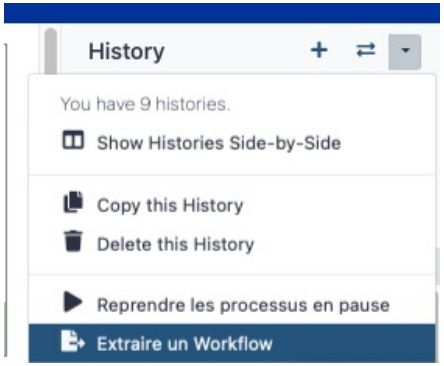
GeneID	logFC	AveExpr	t	P.Value	adj.P.Val	B
VIM	-3.22672225992452	8.60497356712831	-173.57300906047	6.00351582458693e-42	1.18653486757136e-37	86.1130654834889
CPE	-3.22604684433126	8.58396905143803	-166.144884638473	1.90592805759801e-41	1.88343810651835e-37	84.9989541685801
MRC2	-3.7232569936223	7.14577366972087	-148.533632276197	3.67431008825395e-40	2.4206354861417e-36	81.7683179289556
LAMC2	2.58910164196602	8.57498816681594	144.627105927079	7.426848156501e-40	3.66960567412715e-36	81.4401308541912
SCD	2.31993204798972	9.91537685616716	142.221696516008	1.15642308816339e-39	4.57110918289223e-36	81.0415829806911
KRT19	3.3807007962168	8.25334805402281	149.642698126606	1.6003670780297e-39	5.27160915502984e-36	80.6115827778653
SDC3	-3.88988783919027	6.80493584525354	-138.656477258642	2.26051798451957e-39	6.38241106372068e-36	79.8942128630122
FASN	1.85169242587574	10.3290877705531	137.459845739674	2.84174908931951e-39	7.02054112516385e-36	80.1263667415084
THBS1	-2.87433705706947	7.38242568917219	-129.427975821774	1.39257415088914e-38	2.89484259195975e-35	78.4465034457735
ESRP1	3.48988355556255	6.92599459703203	129.180587210833	1.46470481277057e-38	2.89484259195975e-35	78.2077678470495
FBN1	-3.27534780989352	6.82435809008211	-124.333592376246	4.0189453309793e-38	7.22094868377044e-35	77.2808392150983
MAL2	2.45117913366047	8.39739624714793	122.382218827569	6.10130944827492e-38	1.00488566613088e-34	77.073430813214
CDH1	2.30980069779057	8.11531935733231	117.794956575786	1.67206538422226e-37	2.41845066203e-34	76.0648975376207
ADGRF1	2.96853685894318	7.14854500355236	117.686693534202	1.71313040216657e-37	2.41845066203e-34	75.9504687444773
CDH3	2.5580605948251	7.60011090844146	116.963986001507	2.01548438176068e-37	2.65560222140787e-34	75.8546082092273
SPTB	3.21356922941936	6.43274874645418	112.148417410123	6.11118352968501e-37	7.54883945504341e-34	74.5341631210229
EPCAM	2.16226734250385	8.28058386385349	111.545445594787	7.04518810802464e-37	8.19065280982347e-34	74.630442323102
ZEB1	-7.47784145538577	6.6686385152746	-135.599651739795	1.37491553477488e-36	1.48533096410256e-33	71.7435383963473

Gènes différentiels à observer dans l'expérience EMT

- ZEB1: le gène induit dans l'expérience. Doit avoir monté à Day7.
- CDH1 (E-cadherin) : marqueur de cellules epithéliales. Devrait avoir baissé à Day7.
- VIM (vimentin): marqueur de cellules mesenchymateuses. Devrait avoir monté à Day7.
- ESRP1, ESRP2: Deux facteur d'epissage importants trouvés upregulated dans l'article à Day7.

Lien vers article: https://doi.org/10.1128/MCB.00019-16

Affichage du workflow



+ « create workflow
+ « edit »

Rearranger le workflow pour faire clairement apparaitre les étapes

- Nettoyer les étapes non essentielles
- Essayer fonction « auto-layout »