RNAseq analyses workflow to find differentially expressed genes

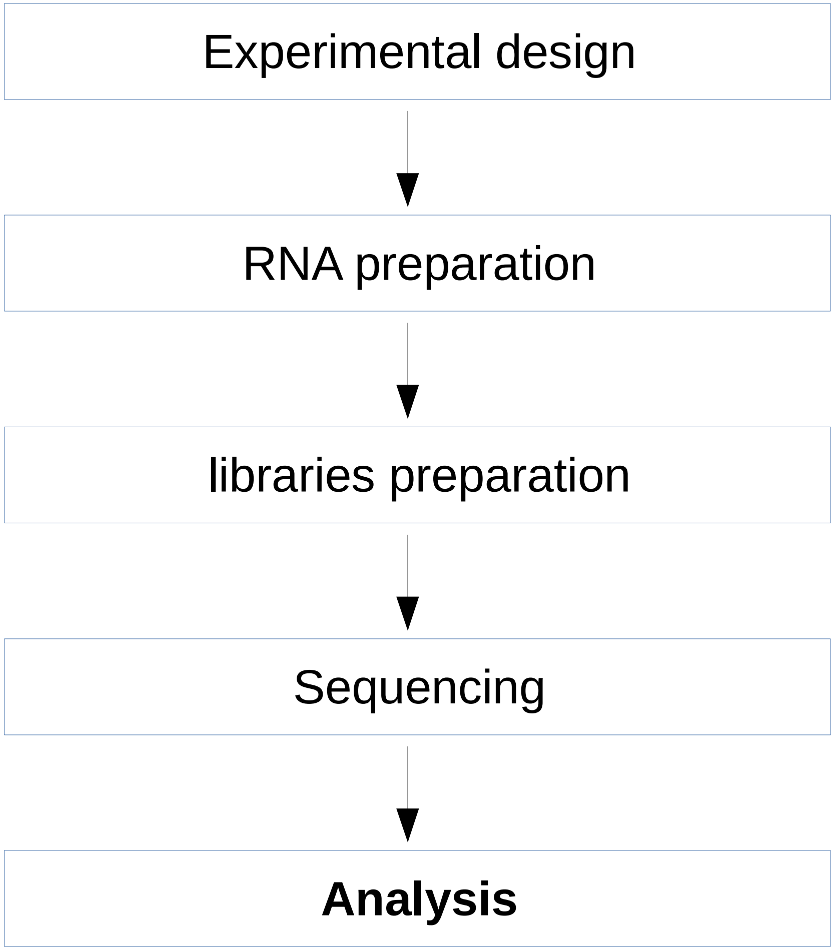
Club bioinfo
03/10/2019

Flora Borne

Aims of RNAseq

- **Measure relative gene expression**
- Discover and annotate complete transcripts
- Characterize alternative splicing and polyadenylation

RNAseq experiment



Experimental design

RNA preparation

libraries preparation

Sequencing

Analysis

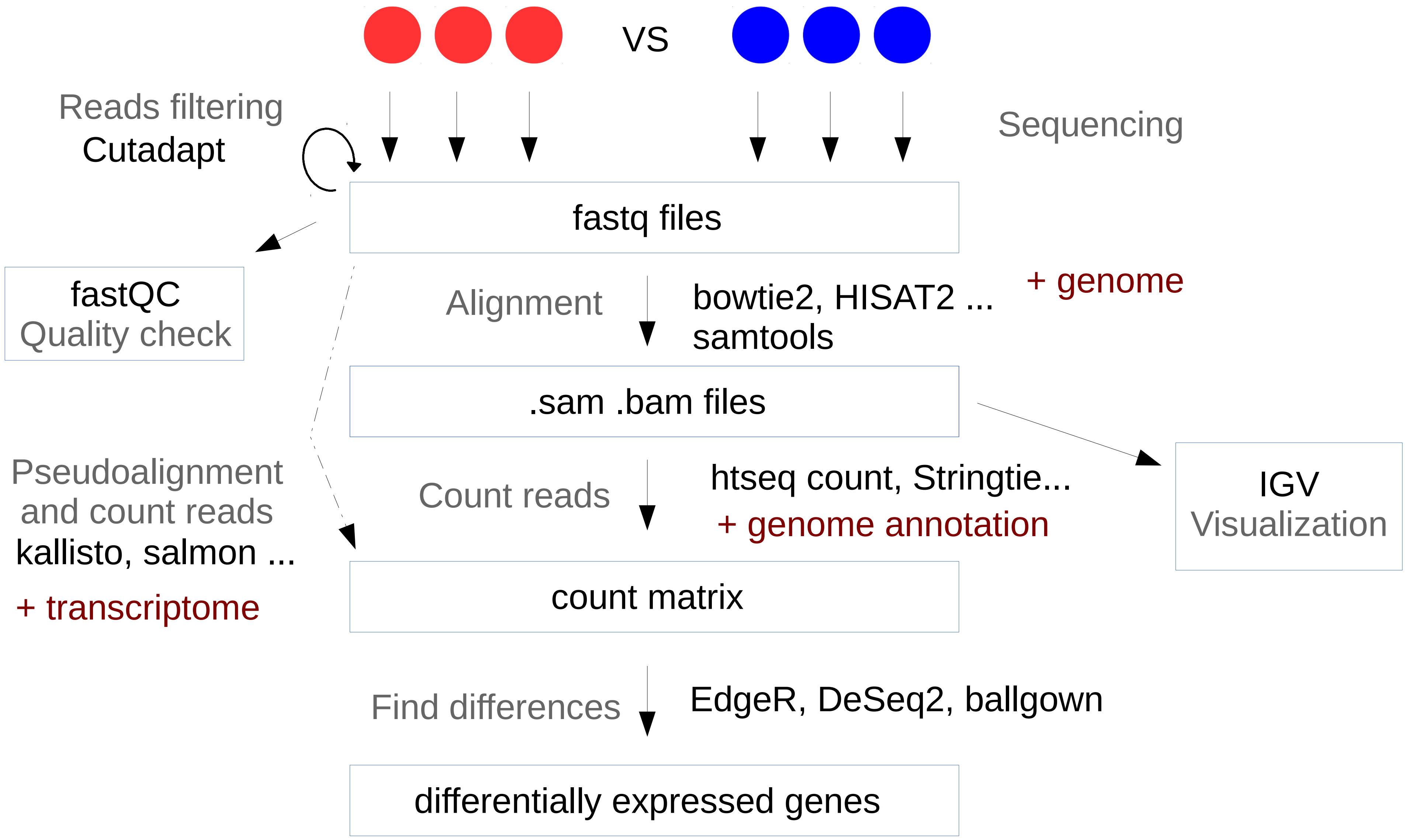
conditions?
replicates?
...

RNA extraction
and purification

cDNA preparation
add adapters

paired-end?
stranded?

RNAseq Analysis pipeline with reference genome



● ● ● VS ● ● ●

Sequencing

Reads filtering
Cutadapt

fastq files

fastQC
Quality check

Alignment

bowtie2, HISAT2 ...
samtools

+ genome

.sam .bam files

IGV
Visualization

Pseudoalignment
and count reads
kallisto, salmon ...

+ transcriptome

Count reads

htseq count, Stringtie...

+ genome annotation

count matrix

Find differences

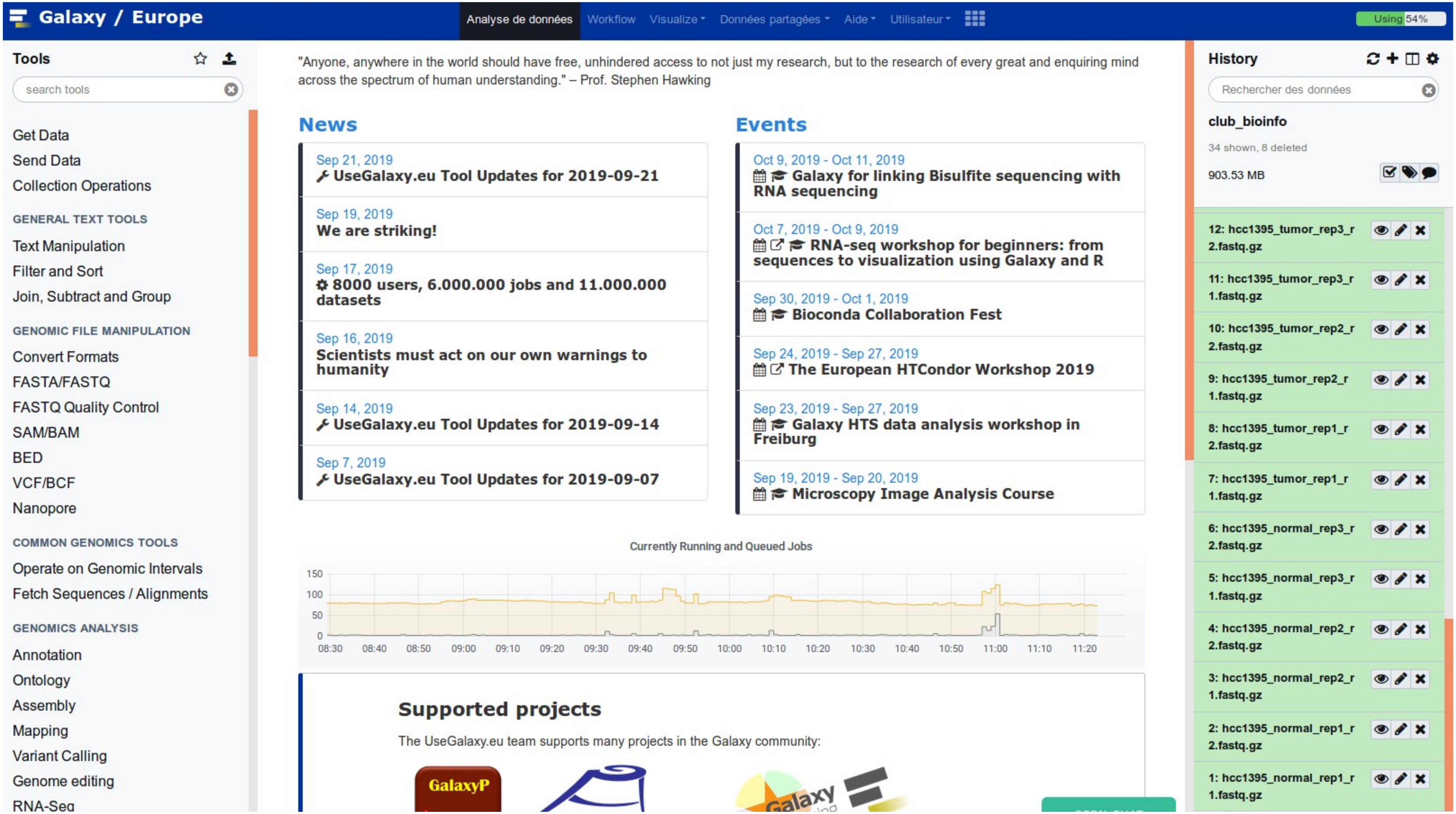
EdgeR, DeSeq2, ballgown

differentially expressed genes

Use Galaxy to perform RNAseq analysis

https://usegalaxy.eu/

https://galaxyproject.github.io/trainingmaterial/topics/introduction/slides/introduction.html#1



Galaxy / Europe

Analyse de données | Workflow | Visualize | Données partagées | Aide | Utilisateur

Using 54%

Tools

search tools

Get Data
Send Data
Collection Operations

GENERAL TEXT TOOLS

Text Manipulation
Filter and Sort
Join, Subtract and Group

GENOMIC FILE MANIPULATION

Convert Formats
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore

COMMON GENOMICS TOOLS

Operate on Genomic Intervals
Fetch Sequences / Alignments

GENOMICS ANALYSIS

Annotation
Ontology
Assembly
Mapping
Variant Calling
Genome editing
RNA-Seq

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

News

Sep 21, 2019
UseGalaxy.eu Tool Updates for 2019-09-21

Sep 19, 2019
We are striking!

Sep 17, 2019
8000 users, 6.000.000 jobs and 11.000.000 datasets

Sep 16, 2019
Scientists must act on our own warnings to humanity

Sep 14, 2019
UseGalaxy.eu Tool Updates for 2019-09-14

Sep 7, 2019
UseGalaxy.eu Tool Updates for 2019-09-07

Events

Oct 9, 2019 - Oct 11, 2019
Galaxy for linking Bisulfite sequencing with RNA sequencing

Oct 7, 2019 - Oct 9, 2019
RNA-seq workshop for beginners: from sequences to visualization using Galaxy and R

Sep 30, 2019 - Oct 1, 2019
Bioconda Collaboration Fest

Sep 24, 2019 - Sep 27, 2019
The European HTCondor Workshop 2019

Sep 23, 2019 - Sep 27, 2019
Galaxy HTS data analysis workshop in Freiburg

Sep 19, 2019 - Sep 20, 2019
Microscopy Image Analysis Course

Currently Running and Queued Jobs

Supported projects

The UseGalaxy.eu team supports many projects in the Galaxy community:

History

Rechercher des données

club_bioinfo

34 shown, 8 deleted

903.53 MB

12: hcc1395_tumor_rep3_r2.fastq.gz
11: hcc1395_tumor_rep3_r1.fastq.gz
10: hcc1395_tumor_rep2_r2.fastq.gz
9: hcc1395_tumor_rep2_r1.fastq.gz
8: hcc1395_tumor_rep1_r2.fastq.gz
7: hcc1395_tumor_rep1_r1.fastq.gz
6: hcc1395_normal_rep3_r2.fastq.gz
5: hcc1395_normal_rep3_r1.fastq.gz
4: hcc1395_normal_rep2_r2.fastq.gz
3: hcc1395_normal_rep2_r1.fastq.gz
2: hcc1395_normal_rep1_r2.fastq.gz
1: hcc1395_normal_rep1_r1.fastq.gz

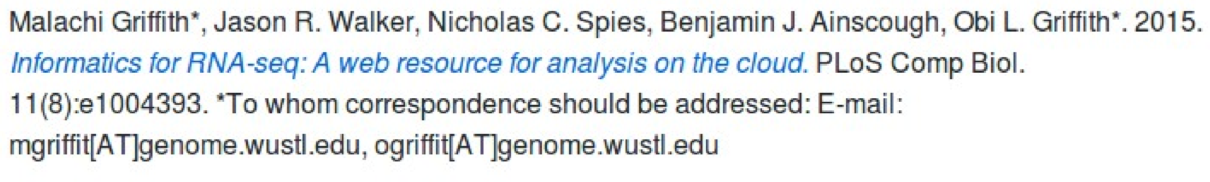
Use Galaxy to perform RNAseq analysis

https://usegalaxy.eu/

- **Create an account**
- **Import history: https://usegalaxy.eu:/u/fborne/h/clubbioinfo**

Data source

https://github.com/griffithlab/rnaseq_tutorial/wiki

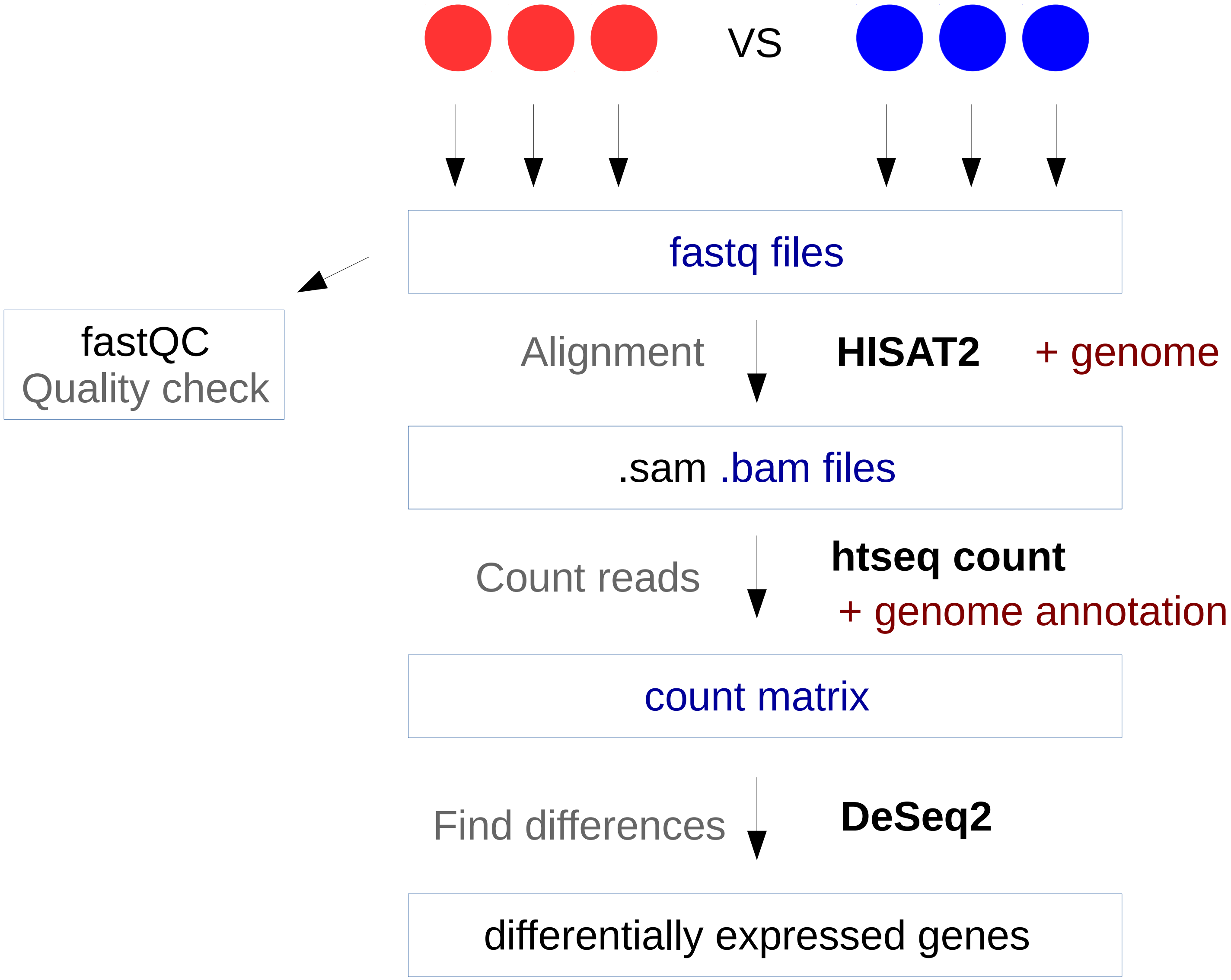


Malachi Griffith*, Jason R. Walker, Nicholas C. Spies, Benjamin J. Ainscough, Obi L. Griffith*. 2015. *Informatics for RNA-seq: A web resource for analysis on the cloud.* PLoS Comp Biol. 11(8):e1004393. *To whom correspondence should be addressed: E-mail: mgriffit[AT]genome.wustl.edu, ogriffit[AT]genome.wustl.edu

What you need

Files:

- 12 fastq files breast cancer cell line VS lymphoblastoid line (tumor vs normal)
- genome file chr22_with_ERCC92.fa
- annotation file chr22_with_ERCC92.gtf



VS

fastq files

fastQC
Quality check

Alignment

HISAT2

+ genome

.sam .bam files

Count reads

htseq count

+ genome annotation

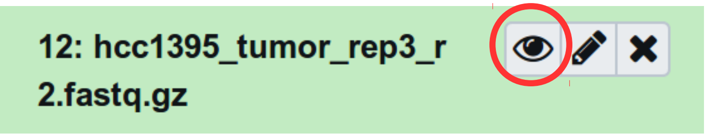
count matrix

Find differences

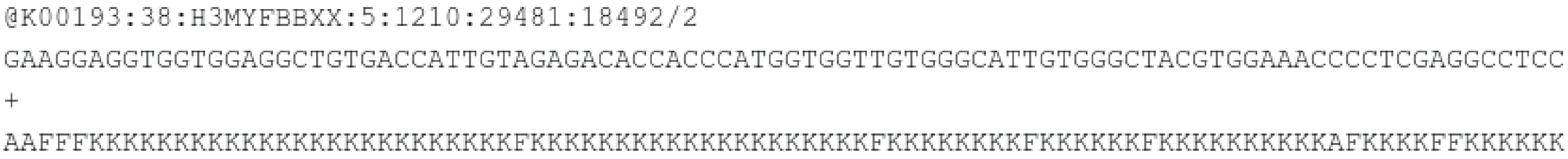
DeSeq2

differentially expressed genes

fastq files



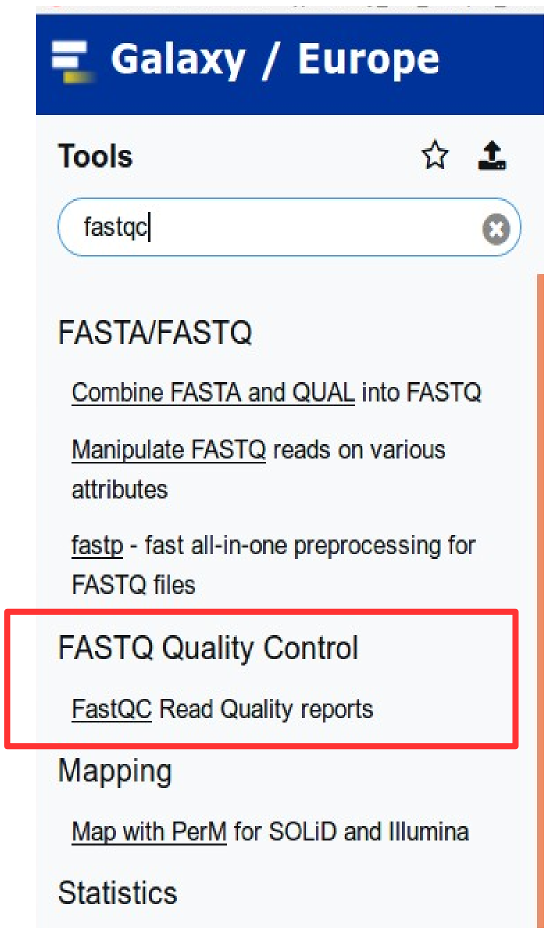
12: hcc1395_tumor_rep3_r2.fastq.gz



```
1  @K00193:38:H3MYFBBXX:5:1210:29481:18492/2
2  GAAGGAGGTGGTGGAGGCTGTGACCATTGTAGAGACACCACCCATGGTGGTTGTGGGCATTGTGGGCTACGTGGAAACCCCTCGAGGCCTCC
3  +
4  AAFFFKKKKKKKKKKKKKKKKKKKKKKKKKFKKKKKKKKKKKKKKKKKKKFKKKKKKKKFKKKKKKFKKKKKKKKKKAFKKKKFFKKKKKK
```

1 @ followed by name of the reads and sequencing information
2 Sequence of the read
3 + followed by additional information
4 Quality score of each base

Check quality with FastQC



Galaxy / Europe

Tools

fastqc

FASTA/FASTQ

Combine FASTA and QUAL into FASTQ

Manipulate FASTQ reads on various attributes

fastp - fast all-in-one preprocessing for FASTQ files

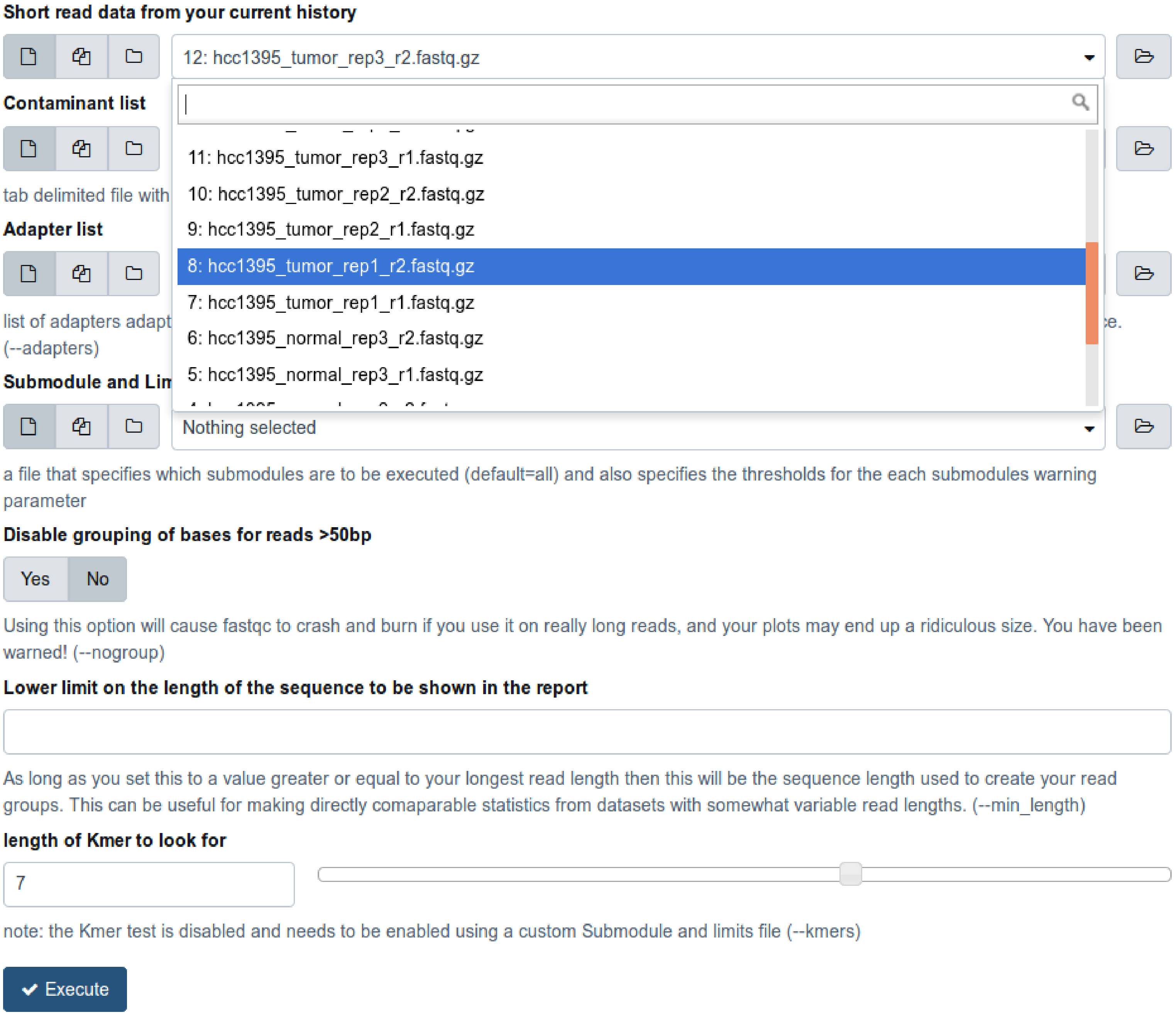
FASTQ Quality Control

FastQC Read Quality reports

Mapping

Map with PerM for SOLiD and Illumina

Statistics



Short read data from your current history

12: hcc1395_tumor_rep3_r2.fastq.gz

11: hcc1395_tumor_rep3_r1.fastq.gz
10: hcc1395_tumor_rep2_r2.fastq.gz
9: hcc1395_tumor_rep2_r1.fastq.gz
8: hcc1395_tumor_rep1_r2.fastq.gz
7: hcc1395_tumor_rep1_r1.fastq.gz
6: hcc1395_normal_rep3_r2.fastq.gz
5: hcc1395_normal_rep3_r1.fastq.gz

Contaminant list

tab delimited file with

Adapter list

list of adapters adapt ... e.
(--adapters)

Submodule and Lim

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Disable grouping of bases for reads >50bp

Yes No

Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)

Lower limit on the length of the sequence to be shown in the report

As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comaparable statistics from datasets with somewhat variable read lengths. (--min_length)

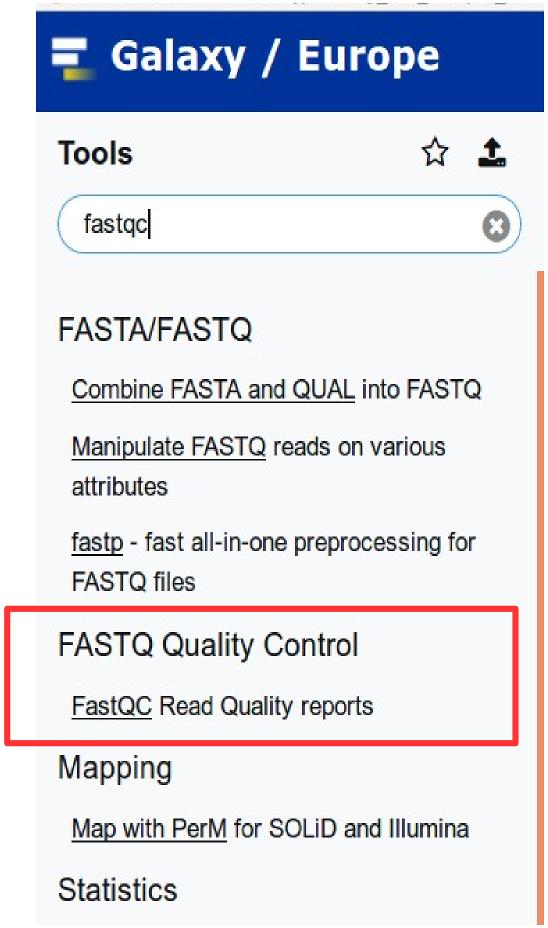
length of Kmer to look for

7

note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers)

✔ Execute

Check quality with FastQC



Galaxy / Europe

Tools

fastqc

FASTA/FASTQ

Combine FASTA and QUAL into FASTQ

Manipulate FASTQ reads on various attributes

fastp - fast all-in-one preprocessing for FASTQ files

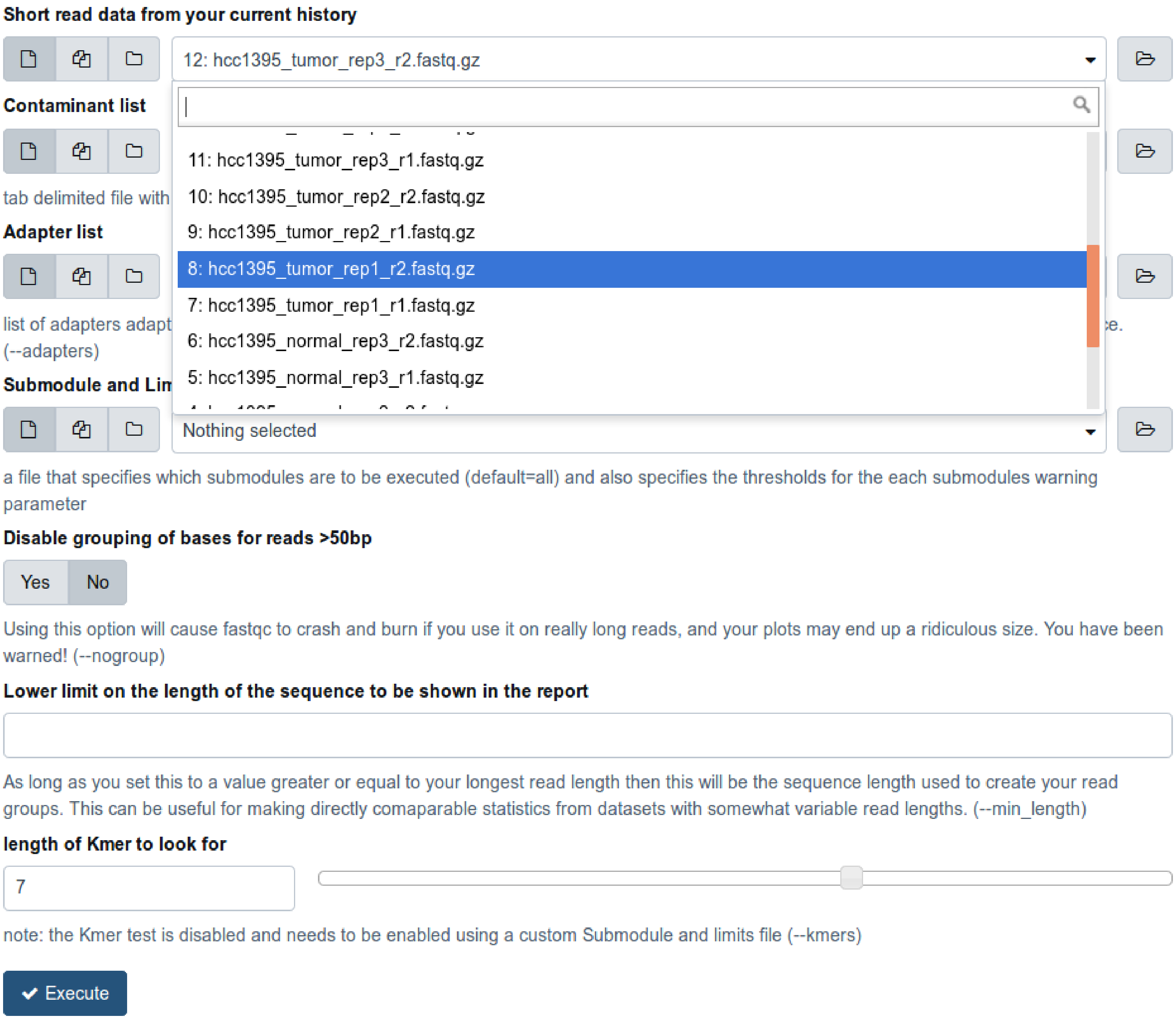
FASTQ Quality Control

FastQC Read Quality reports

Mapping

Map with PerM for SOLiD and Illumina

Statistics



Short read data from your current history

12: hcc1395_tumor_rep3_r2.fastq.gz

- 11: hcc1395_tumor_rep3_r1.fastq.gz
- 10: hcc1395_tumor_rep2_r2.fastq.gz
- 9: hcc1395_tumor_rep2_r1.fastq.gz
- 8: hcc1395_tumor_rep1_r2.fastq.gz
- 7: hcc1395_tumor_rep1_r1.fastq.gz
- 6: hcc1395_normal_rep3_r2.fastq.gz
- 5: hcc1395_normal_rep3_r1.fastq.gz

Contaminant list

tab delimited file with

Adapter list

list of adapters adapt ... e.
(--adapters)

Submodule and Lim

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Disable grouping of bases for reads >50bp

Yes No

Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)

Lower limit on the length of the sequence to be shown in the report

As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comaparable statistics from datasets with somewhat variable read lengths. (--min_length)

length of Kmer to look for

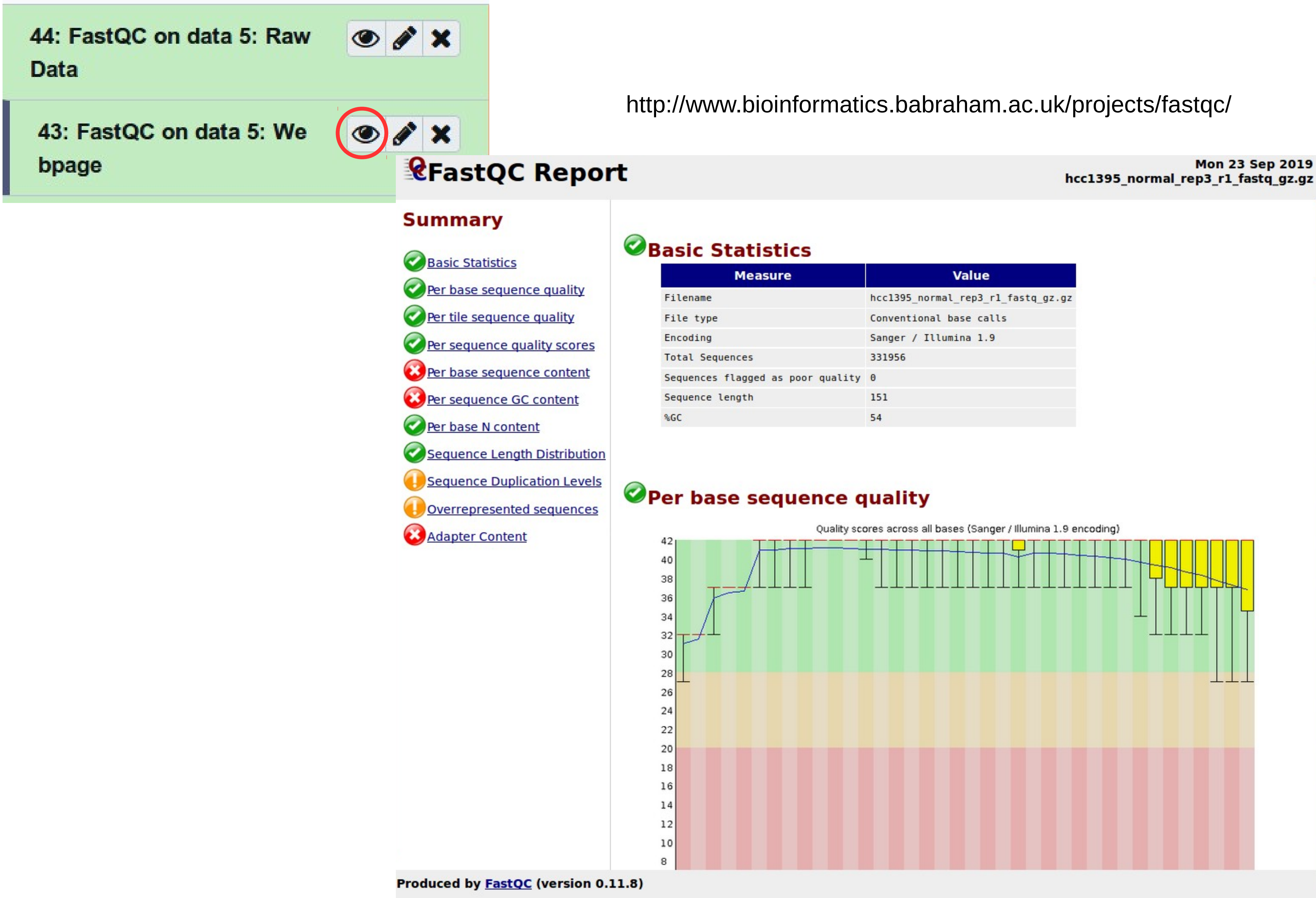
7

note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers)

✓ Execute

Check quality with FastQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



44: FastQC on data 5: Raw Data

43: FastQC on data 5: Webpage

FastQC Report

Mon 23 Sep 2019
hcc1395_normal_rep3_r1_fastq_gz.gz

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Adapter Content

✓ Basic Statistics

Measure	Value
Filename	hcc1395_normal_rep3_r1_fastq_gz.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	331956
Sequences flagged as poor quality	0
Sequence length	151
%GC	54

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
42
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8

Produced by FastQC (version 0.11.8)

Quality and filtering reads

https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html

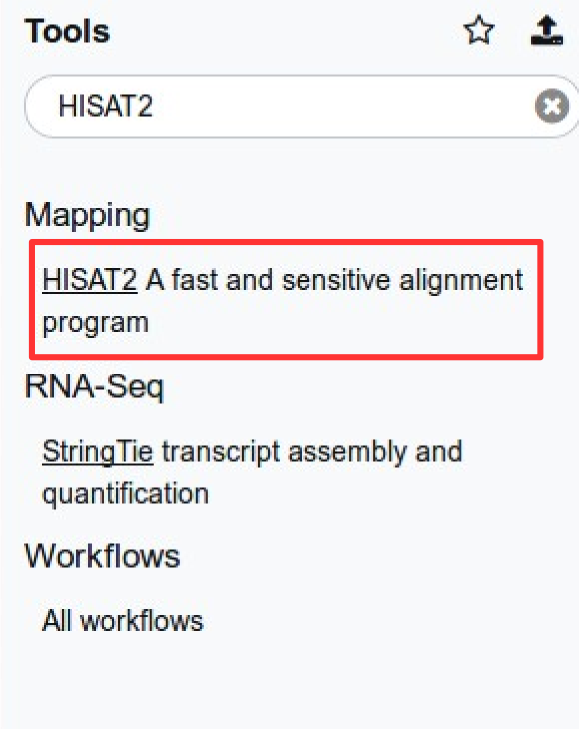
Align reads using HISAT2

Use a genome from history
chr22_with_ERCC92.fa

Paired-end

#1
hcc1395_tumor_rep1_r1.fastq.gz
#2
hcc1395_tumor_rep1_r2.fastq.gz

Summary Options: Print alignment summary to a file -> Yes


Tools

HISAT2

Mapping

HISAT2 A fast and sensitive alignment program

RNA-Seq

StringTie transcript assembly and quantification

Workflows

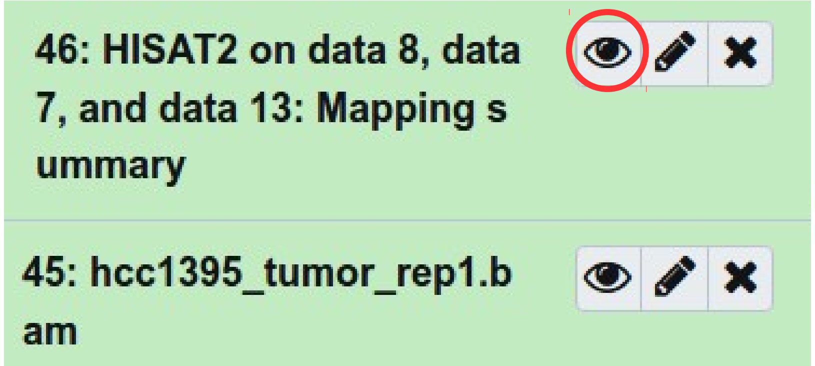
All workflows

Execute

Rename BAM file: hcc1395_tumor_rep1.bam

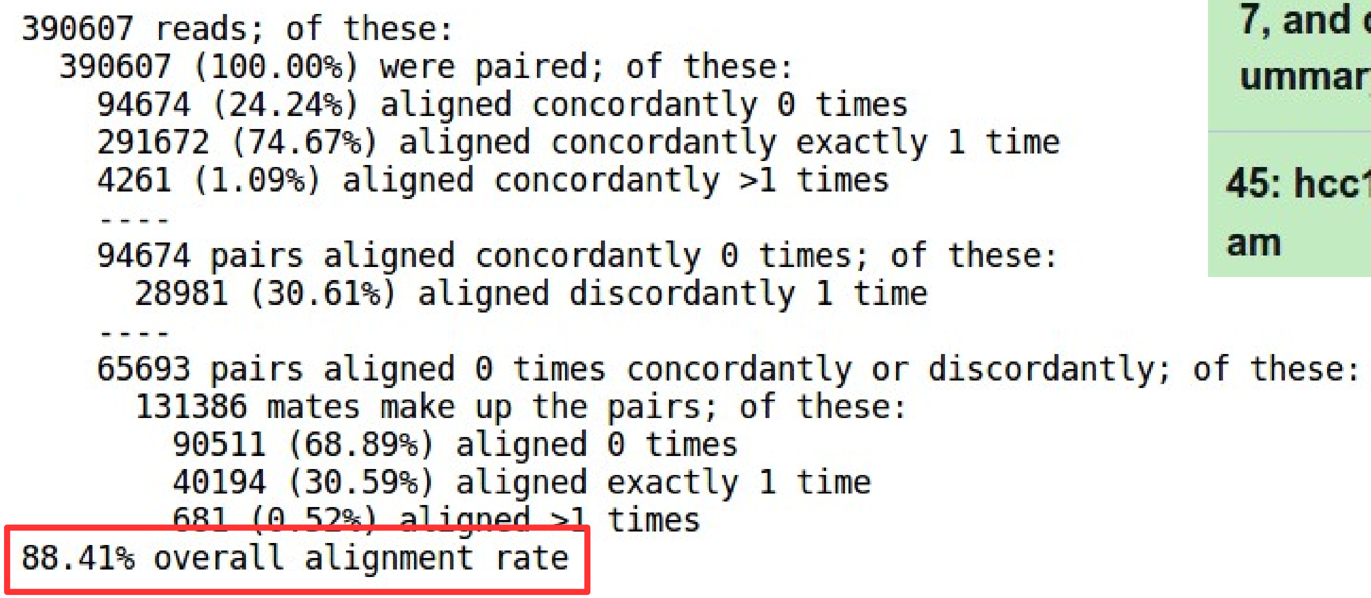
http://ccb.jhu.edu/software/hisat2/manual.shtml

Align reads using HISAT2



46: HISAT2 on data 8, data 7, and data 13: Mapping summary

45: hcc1395_tumor_rep1.bam



```
390607 reads; of these:
  390607 (100.00%) were paired; of these:
    94674 (24.24%) aligned concordantly 0 times
    291672 (74.67%) aligned concordantly exactly 1 time
    4261 (1.09%) aligned concordantly >1 times
    ----
    94674 pairs aligned concordantly 0 times; of these:
      28981 (30.61%) aligned discordantly 1 time
    ----
    65693 pairs aligned 0 times concordantly or discordantly; of these:
      131386 mates make up the pairs; of these:
        90511 (68.89%) aligned 0 times
        40194 (30.59%) aligned exactly 1 time
        681 (0.52%) aligned >1 times
88.41% overall alignment rate
```

Count reads per transcript using htseq-count

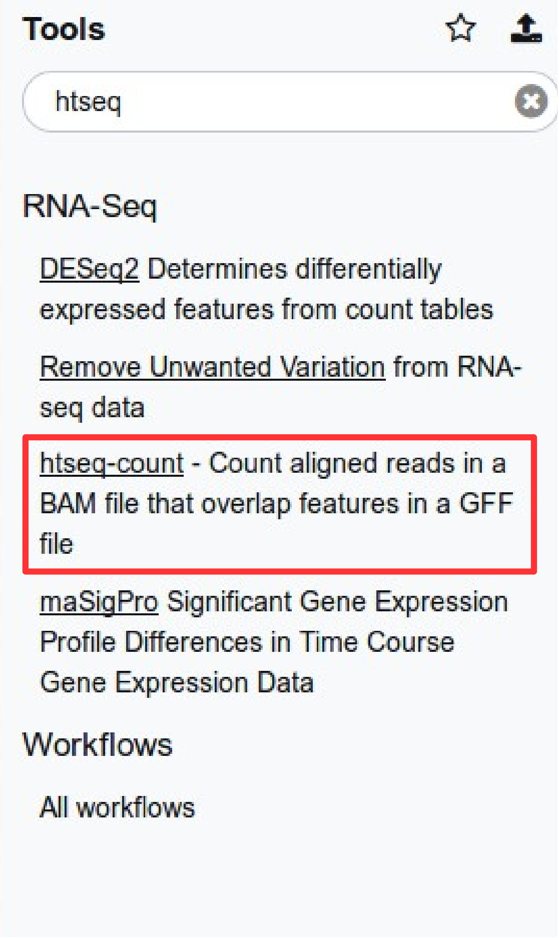
BAM file
hcc1395_tumor_rep1.bam

GFF File
chr22_with_ERCC92.gtf

Stranded
No

ID Attribute
gene_id

Execute



Tools

htseq

RNA-Seq

DESeq2 Determines differentially expressed features from count tables

Remove Unwanted Variation from RNA-seq data

htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file

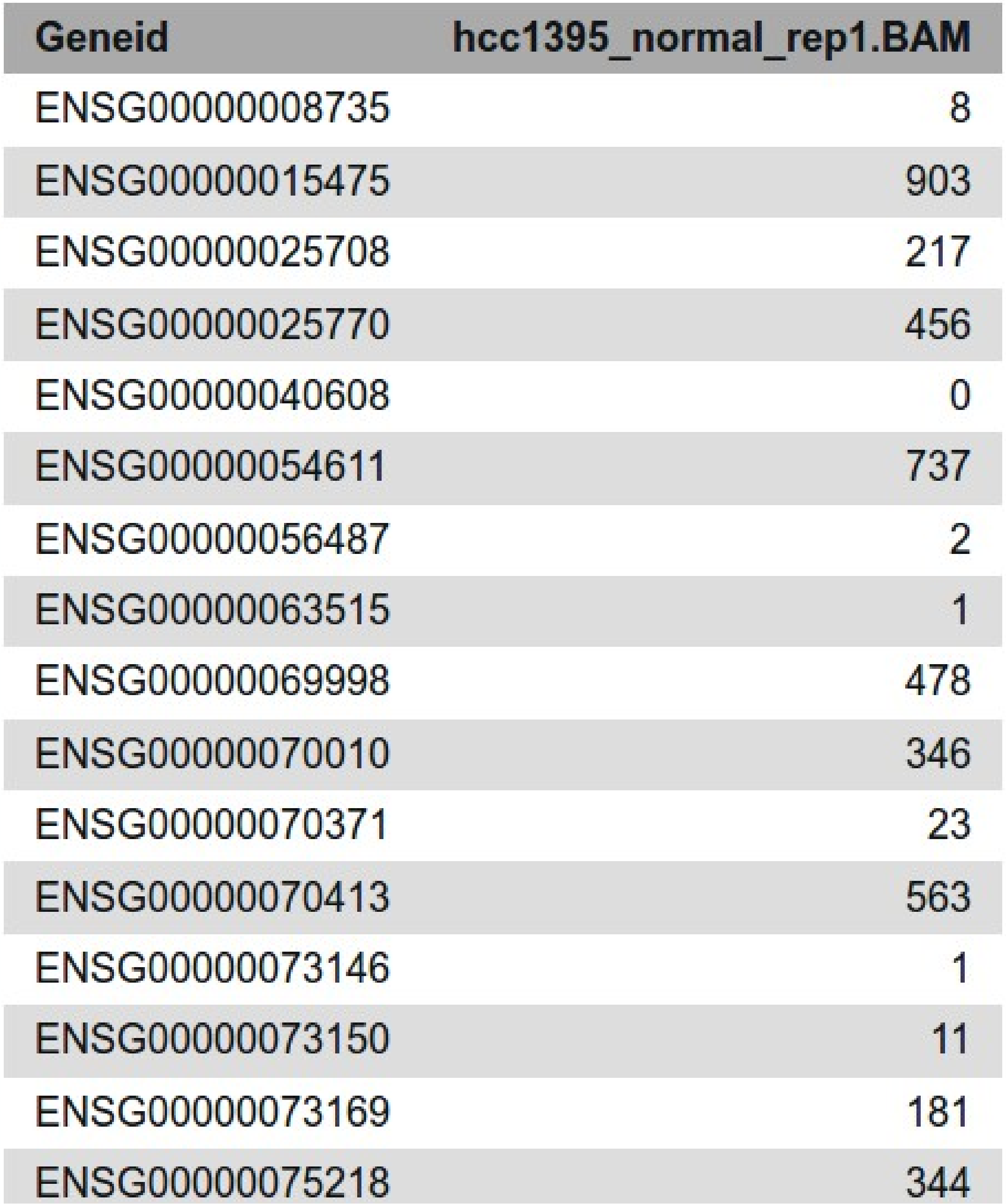
maSigPro Significant Gene Expression Profile Differences in Time Course Gene Expression Data

Workflows

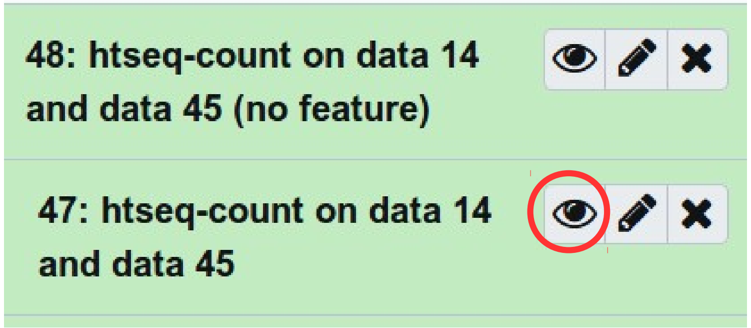
All workflows

https://htseq.readthedocs.io/en/release_0.11.1/

Count reads per transcript using htseq-count



Geneid	hcc1395_normal_rep1.BAM
ENSG00000008735	8
ENSG00000015475	903
ENSG00000025708	217
ENSG00000025770	456
ENSG00000040608	0
ENSG00000054611	737
ENSG00000056487	2
ENSG00000063515	1
ENSG00000069998	478
ENSG00000070010	346
ENSG00000070371	23
ENSG00000070413	563
ENSG00000073146	1
ENSG00000073150	11
ENSG00000073169	181
ENSG00000075218	344

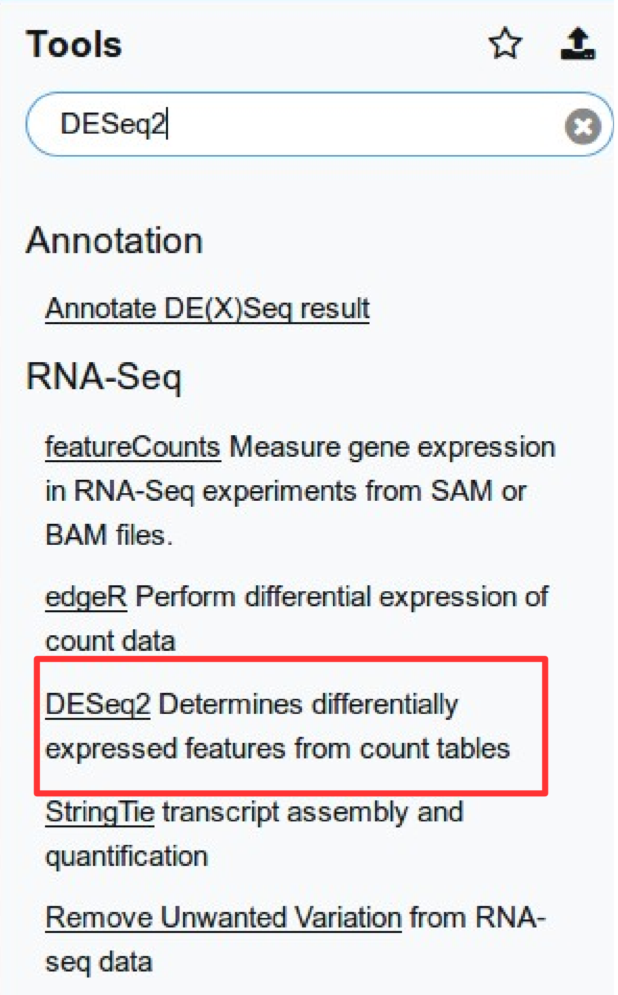


48: htseq-count on data 14 and data 45 (no feature)

47: htseq-count on data 14 and data 45

Find differentially expressed genes using DESeq2

https://www.bioconductor.org/packages/release/bioc/manuals/DESeq2/man/DESeq2.pdf



Tools

DESeq2

Annotation

Annotate DE(X)Seq result

RNA-Seq

featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files.

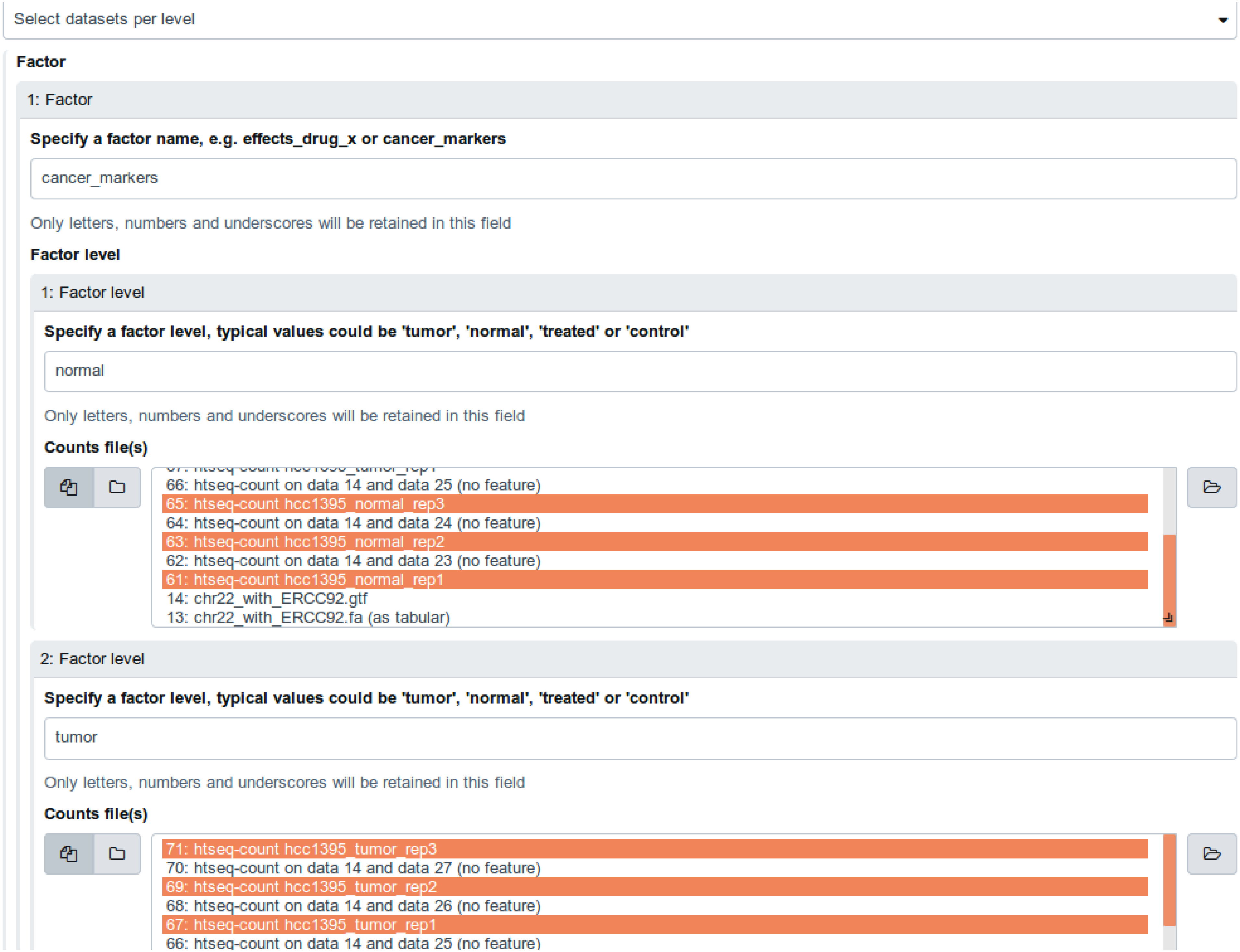
edgeR Perform differential expression of count data

DESeq2 Determines differentially expressed features from count tables

StringTie transcript assembly and quantification

Remove Unwanted Variation from RNA-seq data

Find differentially expressed genes using DESeq2



Select datasets per level

Factor

1: Factor

Specify a factor name, e.g. effects_drug_x or cancer_markers

cancer_markers

Only letters, numbers and underscores will be retained in this field

Factor level

1: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

normal

Only letters, numbers and underscores will be retained in this field

Counts file(s)

67: htseq-count hcc1395_tumor_rep1
66: htseq-count on data 14 and data 25 (no feature)
65: htseq-count hcc1395_normal_rep3
64: htseq-count on data 14 and data 24 (no feature)
63: htseq-count hcc1395_normal_rep2
62: htseq-count on data 14 and data 23 (no feature)
61: htseq-count hcc1395_normal_rep1
14: chr22_with_ERCC92.gtf
13: chr22_with_ERCC92.fa (as tabular)

2: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

tumor

Only letters, numbers and underscores will be retained in this field

Counts file(s)

71: htseq-count hcc1395_tumor_rep3
70: htseq-count on data 14 and data 27 (no feature)
69: htseq-count hcc1395_tumor_rep2
68: htseq-count on data 14 and data 26 (no feature)
67: htseq-count hcc1395_tumor_rep1
66: htseq-count on data 14 and data 25 (no feature)

Find differentially expressed genes using DESeq2

- Normalize counts for (estimate size factor)

sequencing depth

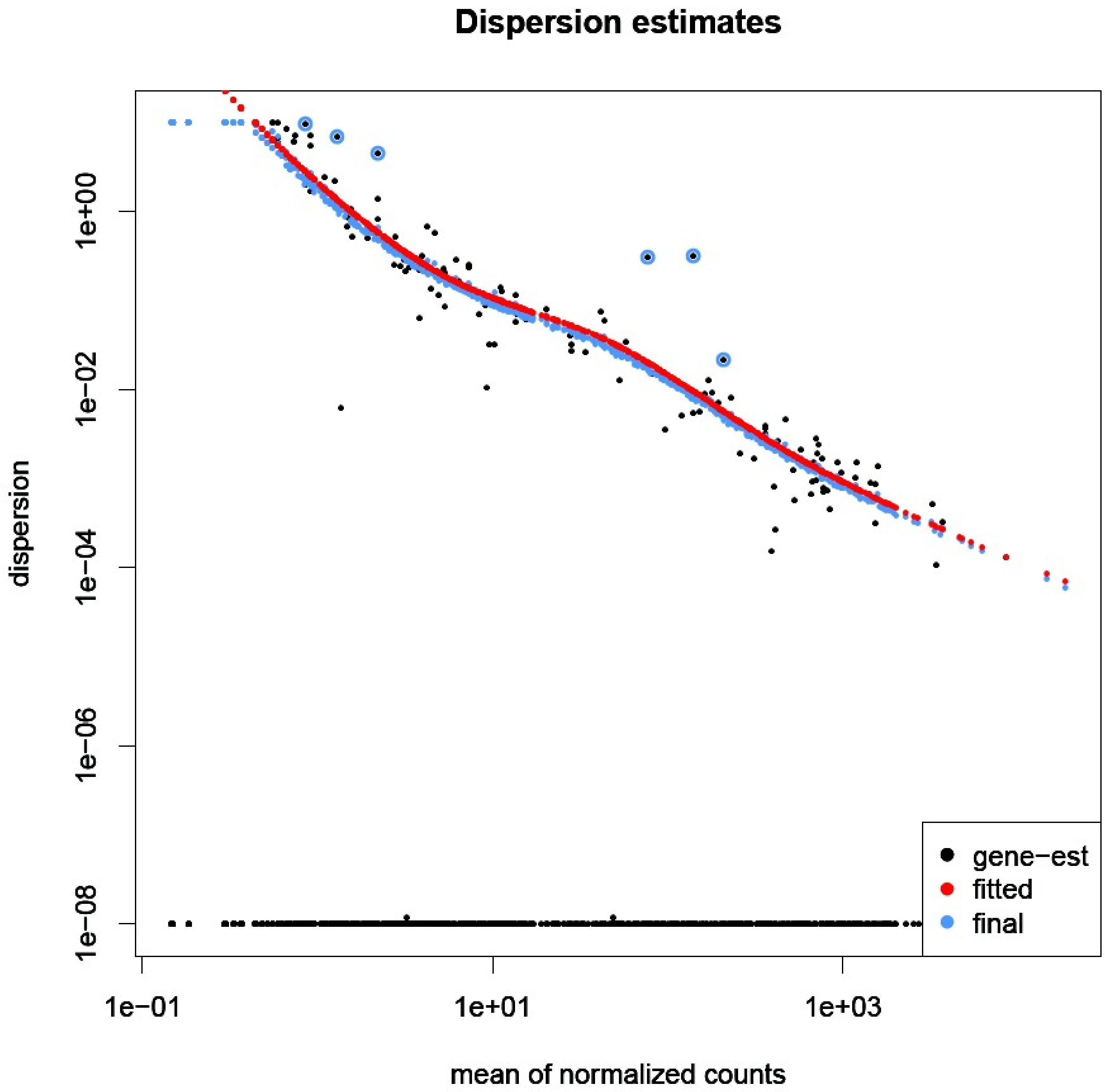
gene_ID	**Sample1**	**Sample2**
geneA	4	8
geneB	105	210
geneC	86	172
geneD	205	410
total reads	**400**	**800**

library composition

gene_ID	**Sample1**	**Sample2**
geneA	4	16
geneB	105	430
geneC	86	354
geneD	605	0
total reads	**800**	**800**

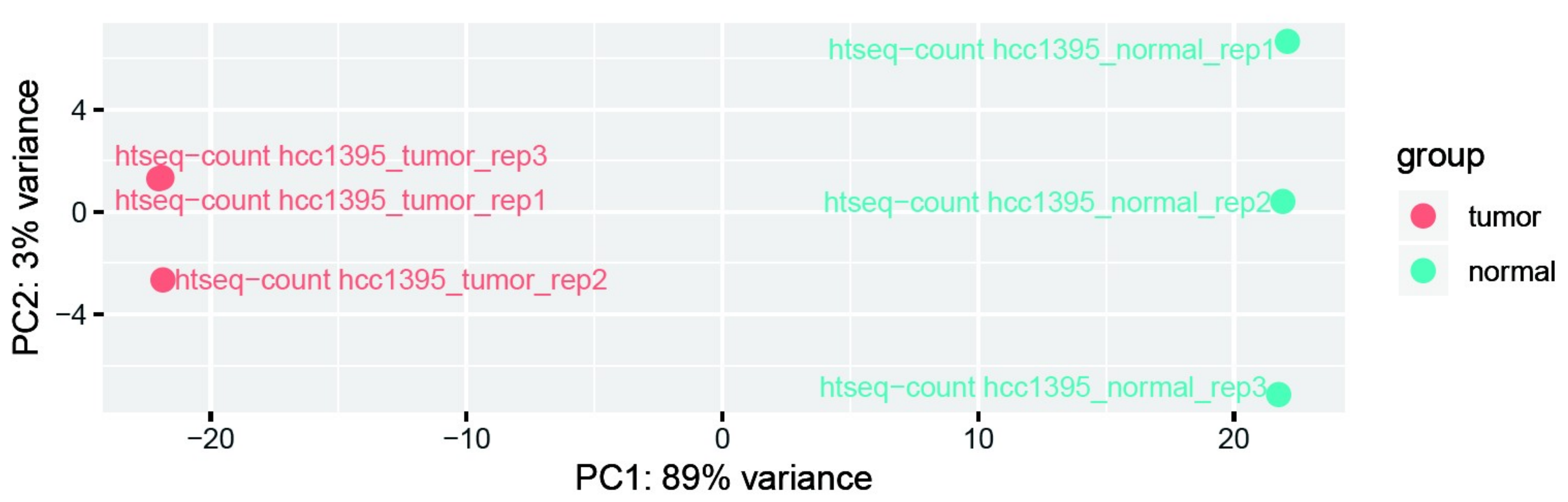
Find differentially expressed genes using DESeq2

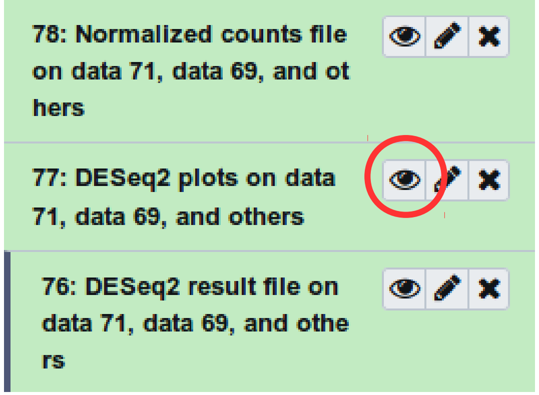
- Estimation of dispersion
- Modelize data with Negative Binomial
- Wald statistics



DESeq2 outputs

Principal component analysis on normalized counts





78: Normalized counts file on data 71, data 69, and others

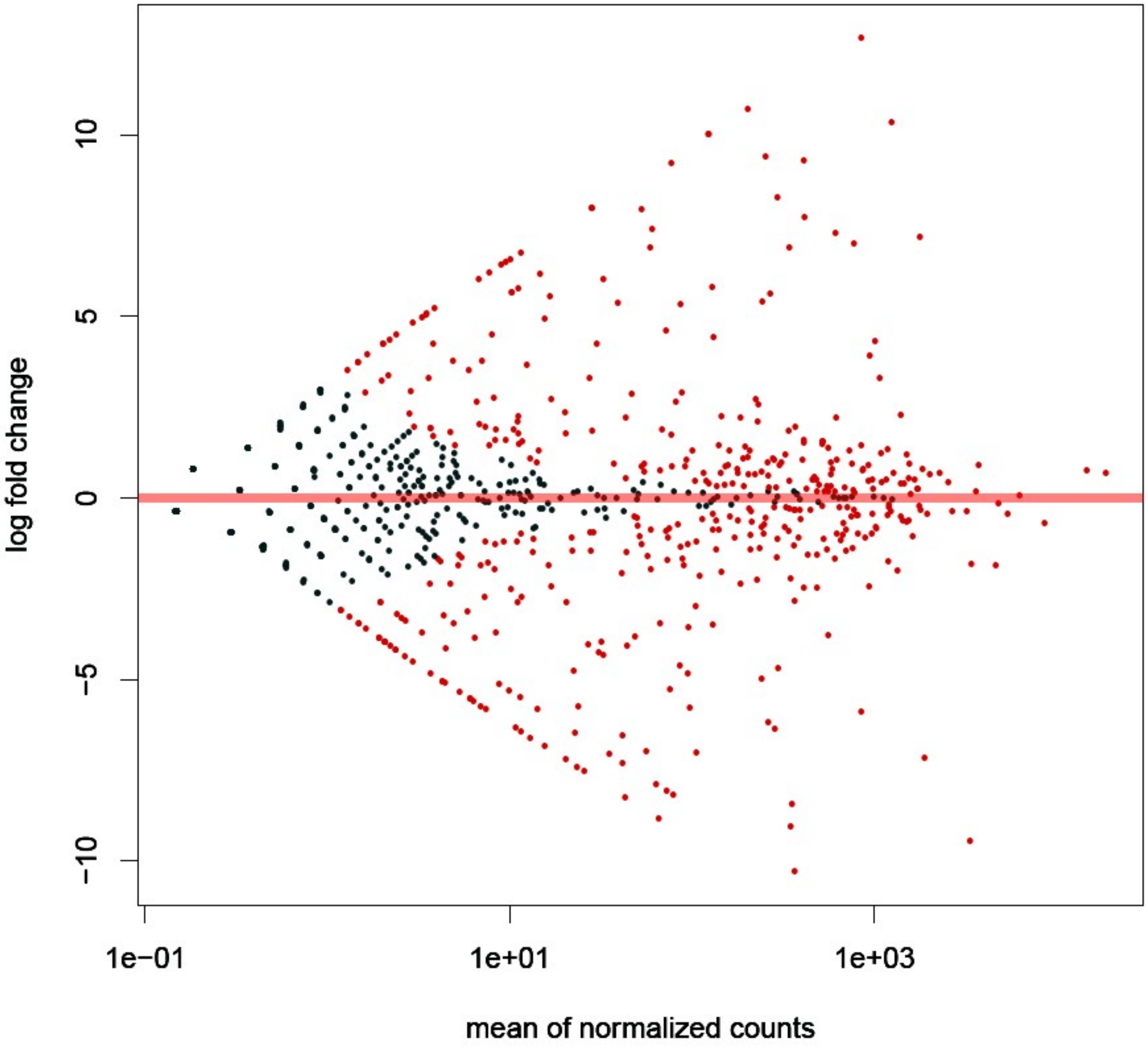
77: DESeq2 plots on data 71, data 69, and others

76: DESeq2 result file on data 71, data 69, and others

DESeq2 outputs

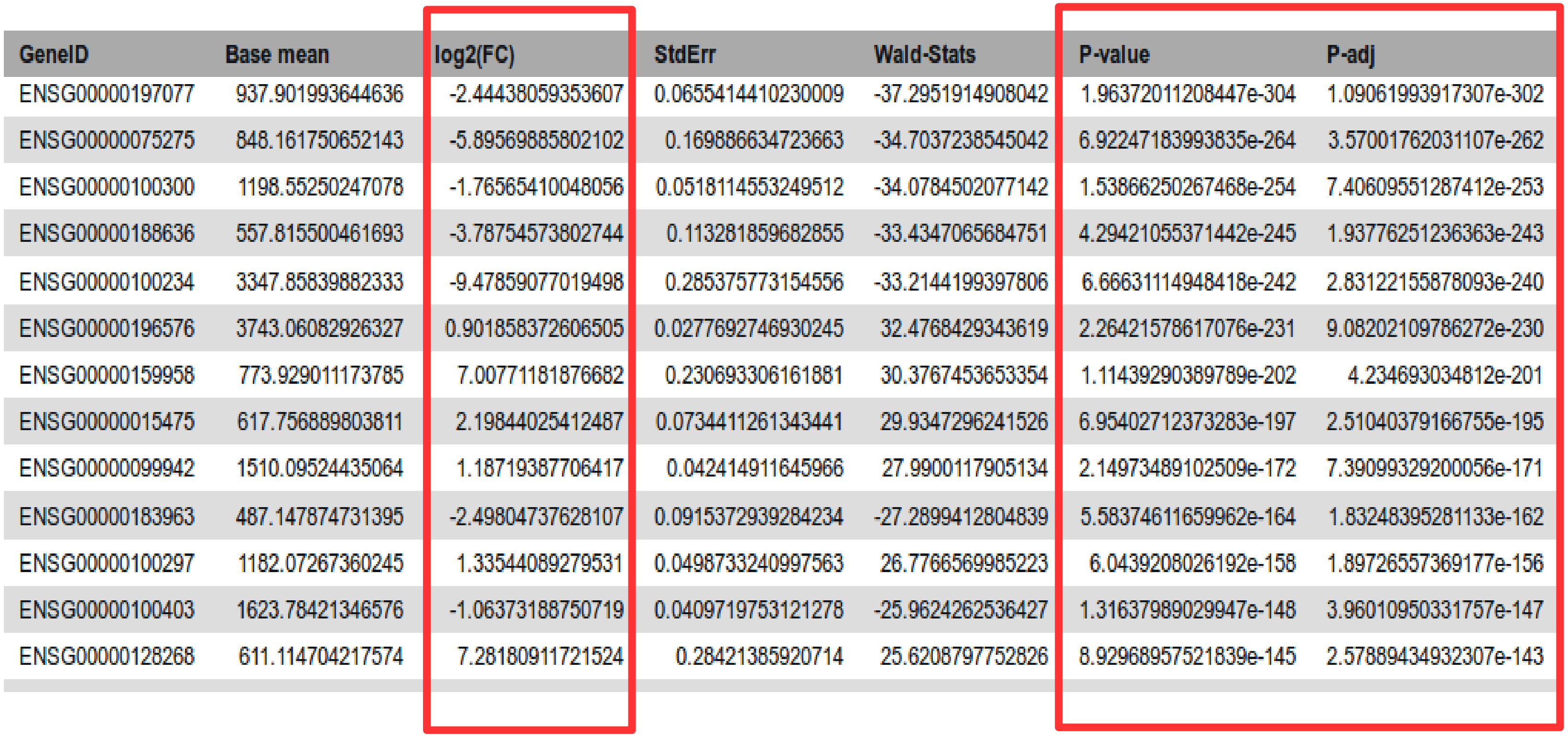
MA plot

log2(FC) = log2(normalize_counts_**normal** / normalized_count_**tumor**)
log2(FC) > 0 up in normal
log2(FC) < 0 up in tumor



p<0.1

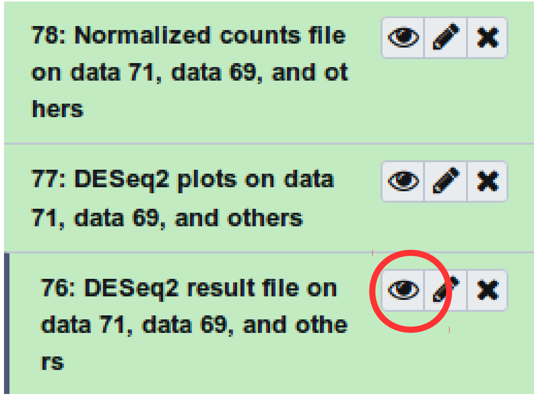
DeSeq2 results



GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
ENSG00000197077	937.901993644636	-2.44438059353607	0.0655414410230009	-37.2951914908042	1.96372011208447e-304	1.09061993917307e-302
ENSG00000075275	848.161750652143	-5.89569885802102	0.169886634723663	-34.7037238545042	6.92247183993835e-264	3.57001762031107e-262
ENSG00000100300	1198.55250247078	-1.76565410048056	0.0518114553249512	-34.0784502077142	1.53866250267468e-254	7.40609551287412e-253
ENSG00000188636	557.815500461693	-3.78754573802744	0.113281859682855	-33.4347065684751	4.29421055371442e-245	1.93776251236363e-243
ENSG00000100234	3347.85839882333	-9.47859077019498	0.285375773154556	-33.2144199397806	6.66631114948418e-242	2.83122155878093e-240
ENSG00000196576	3743.06082926327	0.901858372606505	0.0277692746930245	32.4768429343619	2.26421578617076e-231	9.08202109786272e-230
ENSG00000159958	773.929011173785	7.00771181876682	0.230693306161881	30.3767453653354	1.11439290389789e-202	4.234693034812e-201
ENSG00000015475	617.756889803811	2.19844025412487	0.0734411261343441	29.9347296241526	6.95402712373283e-197	2.51040379166755e-195
ENSG00000099942	1510.09524435064	1.18719387706417	0.042414911645966	27.9900117905134	2.14973489102509e-172	7.39099329200056e-171
ENSG00000183963	487.147874731395	-2.49804737628107	0.0915372939284234	-27.2899412804839	5.58374611659962e-164	1.83248395281133e-162
ENSG00000100297	1182.07267360245	1.33544089279531	0.0498733240997563	26.7766569985223	6.0439208026192e-158	1.89726557369177e-156
ENSG00000100403	1623.78421346576	-1.06373188750719	0.0409719753121278	-25.9624262536427	1.31637989029947e-148	3.96010950331757e-147
ENSG00000128268	611.114704217574	7.28180911721524	0.28421385920714	25.6208797752826	8.92968957521839e-145	2.57889434932307e-143

differentially
expressed?

significant?



78: Normalized counts file on data 71, data 69, and others

77: DESeq2 plots on data 71, data 69, and others

76: DESeq2 result file on data 71, data 69, and others

Documentation

Tutorial on galaxy
https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html

Tutorial about DESeq2
https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html
https://hbctraining.github.io/DGE_workshop/lessons/05_DGE_DESeq2_analysis2.html

Thank you!