

DATA PROCESSING IN GENOMICS: FROM THE SEQUENCERS TO THE RESULTS

GIGA Doctoral School for Health Sciences
Introduction to scientific computing

ARNAUD LAVERGNE

GIGA-Genomics

Bioinformatic team



GIGA - GENOMICS

- Platform
 - Sequencing services
 - Biological materials
 - DNA/RNA
 - Libraries
 - Sequencers
 - Data
 - High Throughput Sequencing (HTS) / Next Generation Sequencing (NGS)
 - Bioinformatic team
 - Data analysis
 - And more ...



Since the development of new next generation sequencing technologies, the field of genomics has had an immense boost.

Application of these new technologies have rapidly become common practice in many research and diagnostic fields.

Sequencing whole genomes of known and new species are now routine experiments.

The GIGA genomics platform offers a wide range of services in bulk or single cell DNA/RNA analysis. Technologies ranges from Sanger sequencing to high throughput genotyping, high throughput sequencing and to long read sequencing.

Beside generating data, the platform also offers services in bioinformatics analysis.

www.gigagenomics.uliege.be



EXPERIENCED STAFF



We help you in setting up the experimental design



We produce sequencing libraries for many applications



We sequence on short or long read sequencing platforms



We generate QC reports on the results



We provide support/advice in the analysis of your results



We analyze your data in depth in close collaboration with you

EQUIPMENT

ABI 3700 Sanger sequencer
48 fragments sequenced up to 1800 bp

2•Illumina MiSeq
25•10⁶ fragments sequenced up to 600 bp

2•Illumina NextSeq500
400•10⁶ fragments sequenced up to 300 bp

Illumina NovaSeq6000
20•10⁹ fragments sequenced up to 300 bp

Oxford Nanopore long read sequencer
10 to 20 Gbp of long reads >10kb

2•Illumina iScan + autoloader
Cost efficient array genotyping, up to 2000 samples/week

Chromium 10x Genomics
High throughput Single cell transcriptomics

Computer cluster
552 cores and 4.8T ram

Secured storage
1500T disk storage,
1500T tape storage

APPLICATIONS

Data generation

De novo genome sequencing

Whole genome re-sequencing

Bulk transcriptome analysis

Single cell transcriptome analysis

Long read DNA/RNA sequencing

Cohort genotyping

Metagenomics

Amplicon sequencing

TCR repertoire sequencing

Ribosome profiling

Data analysis

De novo assembly

Genome wide association analysis (GWAS)

Differential expression analysis

Single cell expression analysis

RNA velocity

Whole genome variant calling

GIGA PLATFORMS



Genomics



Cell Imaging



Flow Cytometry



CRC in vivo Imaging



CRC Preclinical Imaging



Immunohistology



Proteomics



Viral Vectors



Mouse Facility



Zebrafish Facility

CONTACTS

GIGA-Genomics
Wouter Coppieters
Platform manager
wouter.coppieters@uliege.be
+32 4 366 41 59

For academics Carine Bebrone
GIGA-Technology Platforms manager
carine.bebrone@uliege.be
+32 4 366 98 32

For business Caroline Thielen
Bridge2Health
caroline.thielen@b2h.be
+32 4 242 77 60

SEQUENCERS



MiSeq
540 Mb -15 Gb
4 – 56 hours



HiSeq
105 Gb - 1,5 Tb
1 – 3,5 days



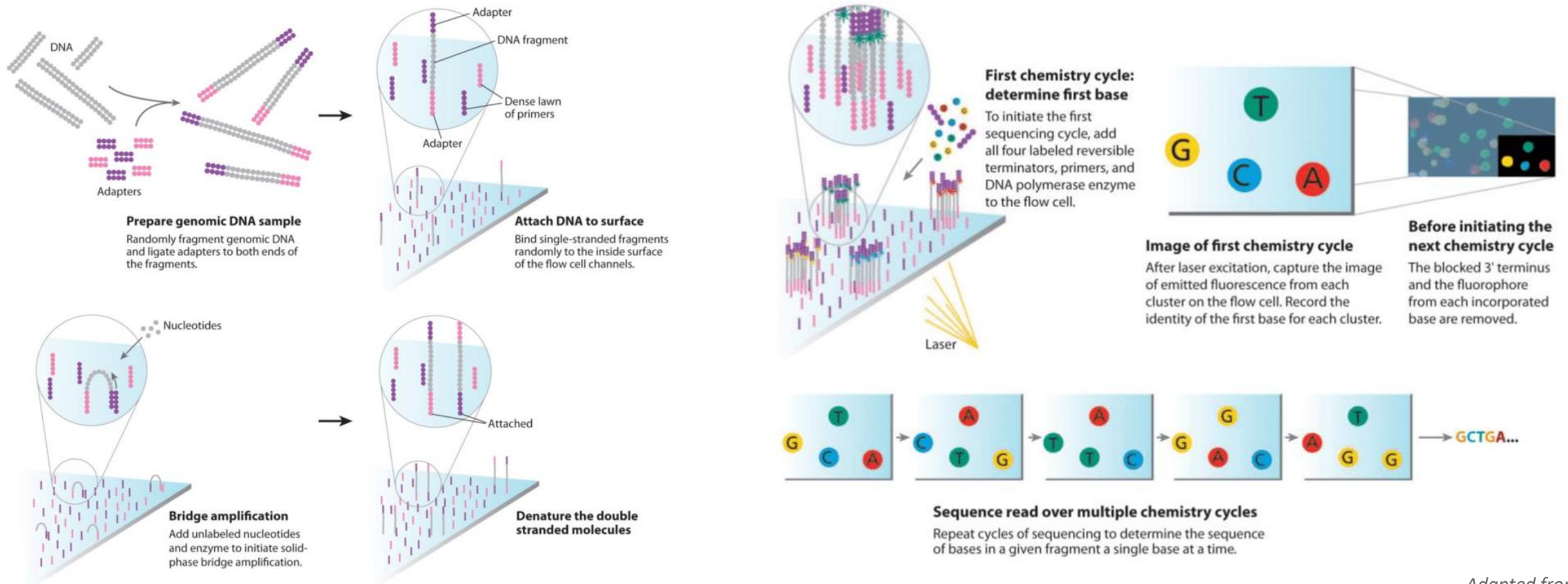
NextSeq
16,25 Gb - 120 Gb
11 – 29 hours



NovaSeq
65 Gb – 3 Tb
13 – 44 hours

Adapted from Illumina

SEQUENCING (HTS/NGS)



Adapted from Illumina

COMPUTING

- DATA MANAGEMENT

- DATA PROCESSING
- STORAGE
- RESULTS

- REPRODUCIBILITY

- PIPELINES
- CONTAINERS



DATA MANAGEMENT



LAPTOP
250Gb – 1Tb
4-8 Go RAM
4-8 CPUs



LAB COMPUTER
1 - 5 Tb
32 - 128 Go RAM
8-16 CPUs



CLUSTER/STORAGE
1.5 Pb
256 Go RAM
32 CPUs

DATA MANAGEMENT

August 28, 2020



- 1.2 Tb
- 12.8 Billions of reads
- 612 samples
 - Multiple experiments
 - Unique combinations of indexes
- « Run »
- No storage on device



DATA PROCESSING

- DEMULTIPLEXING

- Reads → Samples
- ~ 20M / sample
- « Fastq »
 - Identifier
 - Sequence
 - Separator
 - Quality score
 - Ascii +33

- + 1.2 Tb (2.4 Tb)



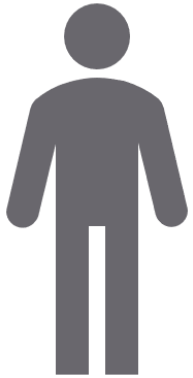
```
@A00801:49:H2VGKDSXY:2:1101:1624:1016 1:N:0:NGCTTAAG+TCGTGACC
CTTCTGGAGAGGAGTTCTCTGATATGAATTAAGGTTTTCCCTCTGTGCATGACCAGAAGAAGGTTTTATCTGTGCCACACTACTTTTCATTTTCTGTTGCCAGTTGGTCCAATA
AATCAAAGATGNTTCAAACCTGGTCCAATAACAAGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFF
@A00801:49:H2VGKDSXY:2:1101:2022:1016 1:N:0:NGCTTAAG+TCGTGACC
TGACAAAAGATACCTCATTATGGGAAATTGAGGAAGATACATATACAAGCACCCCAACCCATATTTAACATATTTGGCAATAACTCCCTCCATTCTCCCCCTCAATT
TCAAATAGTAGNTTTTAAAAAATTAAGACATGTC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
:FF,FF
@A00801:49:H2VGKDSXY:2:1101:4137:1016 1:N:0:NGCTTAAG+TCGTGACC
TTTTTTGCCCTTTCAAGTGTTATTTTATACATTTTTTGTATTAAGAAAGAAAAGCATAATTACCACAAATTACAAAGGACTAAAGCAGGACTAGAATAATGAATGAATCAC
TTCAGCTGGAANGCAGATACTCTCAATAATTAAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFF
```

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	6	54	21
"	34	1	7	55	22
#	35	2	8	56	23
\$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	;	59	26
'	39	6	<	60	27
(40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	30
+	43	10	@	64	31
,	44	11	A	65	32
-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	40
5	53	20			

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90,0000%
20	1 in 100	99,0000%
30	1 in 1000	99,9000%
40	1 in 10,000	99,9900%
50	1 in 100,000	99,9990%
60	1 in 1,000,000	99,9999%

RESEARCHERS





- 12 samples
- Fastq
- 52 Gb
- RNA-Seq

Transcriptomics
« Gene Expression »

WORKFLOW

- QC Sequencing
- Mapping
- Quantification
- QC Mapping/Quantification

- Downstream Analysis
 - Clustering
 - Differential Expression

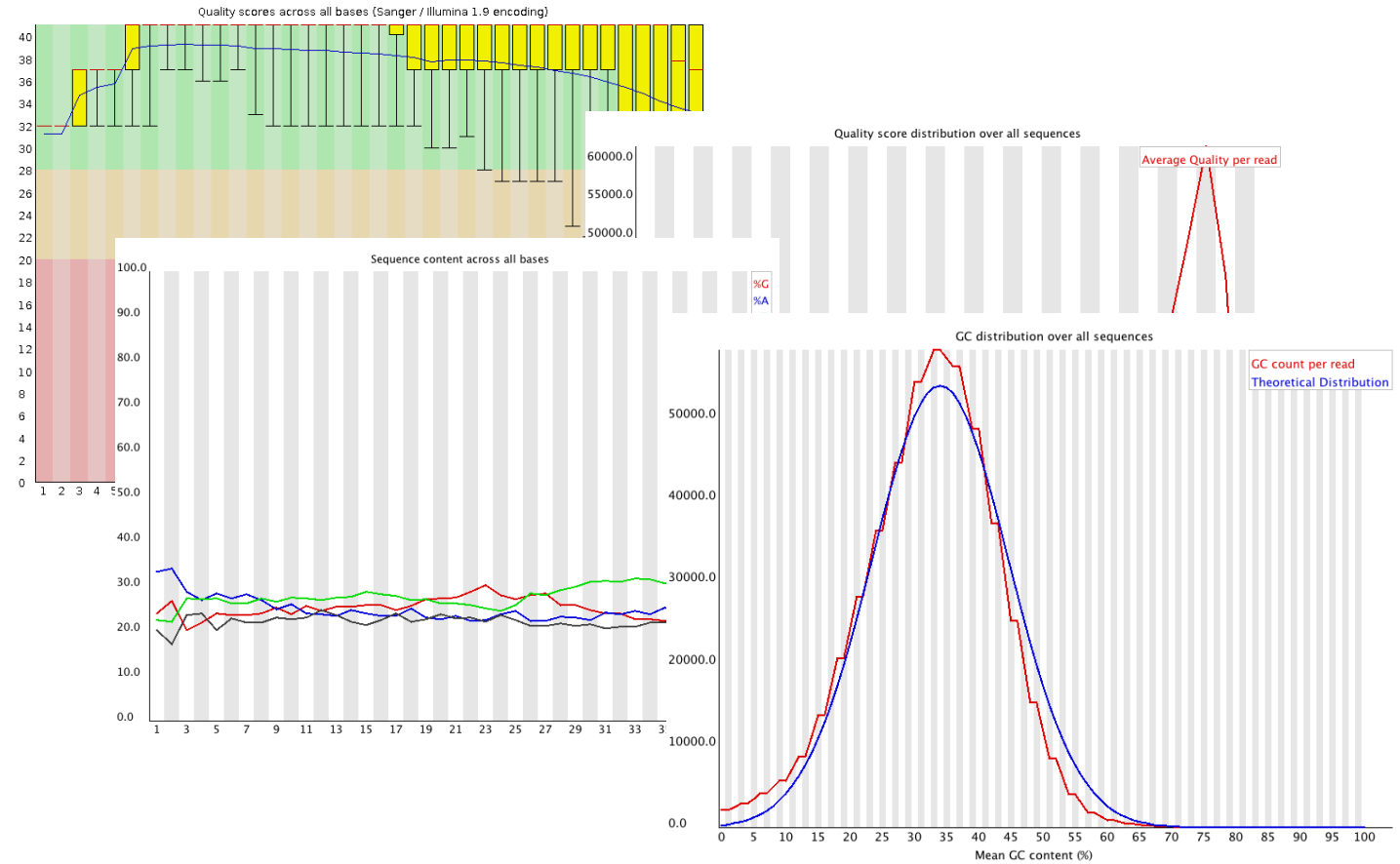
QC SEQUENCING

Dr GIGA
52 Gb



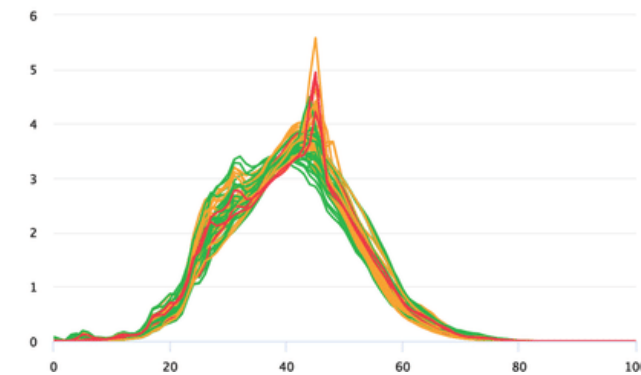
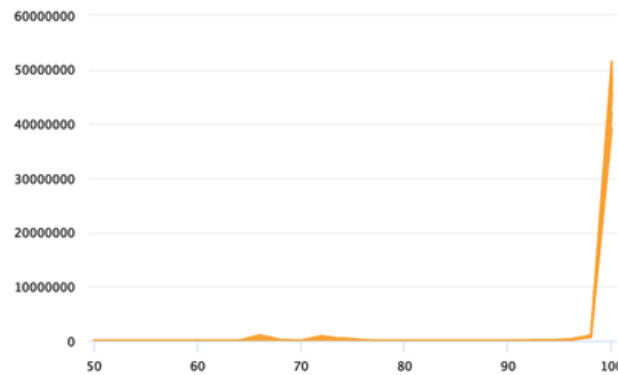
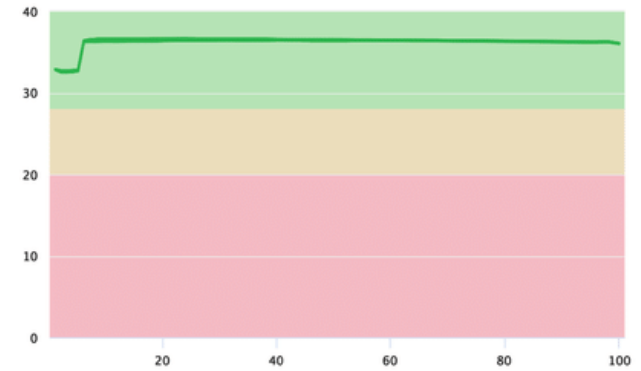
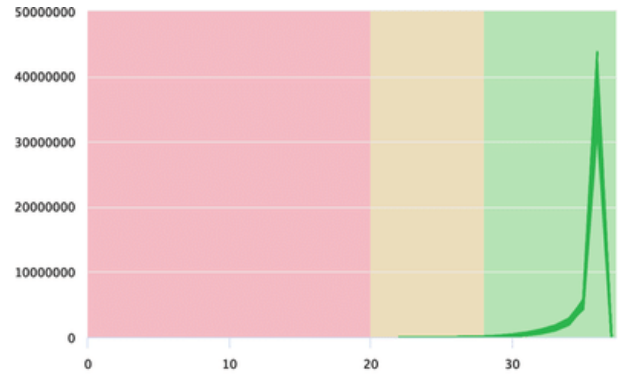
- Number of reads
- Base calling quality
- Sequence quality
- GC content
- Sequence length
- Duplication levels
- Adapter content
- Overrepresented sequences
- ...

< 1 Mb



QC SEQUENCING

Dr GIGA
52 Gb



MAPPING

Dr GIGA
52 Gb

- Alignment
- Origin of reads
- Reference
 - Genome sequence
 - Gene set
- Database
 - Ensembl, UCSC, ...



REFERENCE

Genome (FASTA)

```
>1 dna:chromosome chromosome:GRCh38:1:1:248956422:1 REF
CCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
CCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
CCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
AACCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
GACCTGAGGAGAACTGTGCTCCGCCCTCAGAGTACCACCGAAATCTGTGACAGAGGACA
ACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGC
CGTTGCAAAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCGC
AGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGC
CGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCGC
CGGCGCAGGCGCAGACACATGCTAGCGCGTCCGGGTGGAGGCGTGGCGCAGGCGCA
GAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGT
GGAGGCGTGGCGCAGGCGCAGAGAGGCGCACCGCGCCGGCGCAGGCGCAGAGACA
CATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAGACGCAAGCCTACG
GGCGGGGGTTGGGGGGGGCGTGTGTTGCAGGAGCAAAGTTCGCACGGCGCCGGGCTG
GGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAACTCACGTACGGTGGCGCGG
CGCAGAGACGGGTAGAACCTCAGTAATCCGAAAAGCCGGATCGACCGCCCTTGCTT
GCAGCCGGGCACTACAGGACCCGCTTGCTCACGGTGTGCTGTGC
```

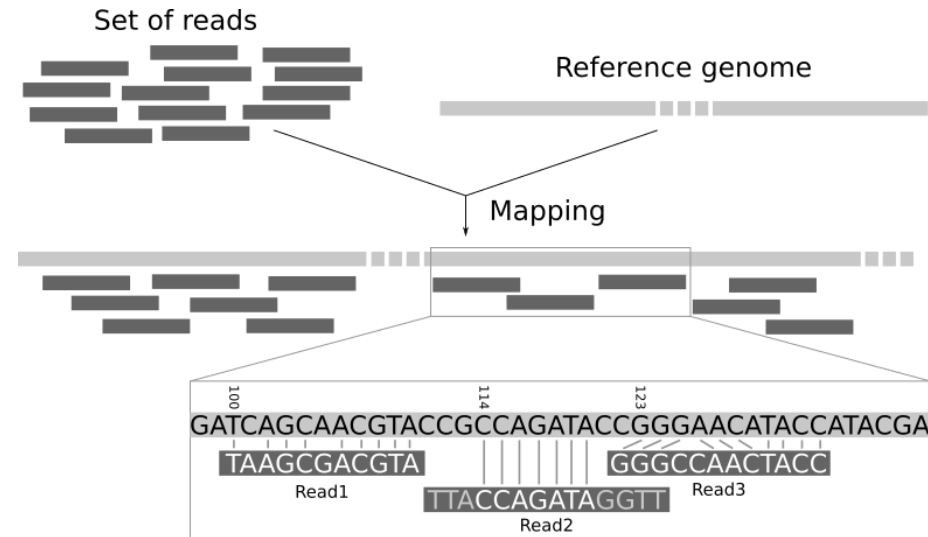
Gene Set (GTF)

```
#!genome-build GRCh38.p12
#!genome-version GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.27
#!genebuild-last-updated 2019-03
1      havana  gene      11869   14409   .      +      .      gene_id "ENSG00000223972"; gene_version "5"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";1
      havana  transcript 11869   14409   .      +      .      gene_id "ENSG00000223972"; gene_version "5"; transcript_id
      "ENST00000456328"; transcript_version "2"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
      transcript_name "DDX11L1-202"; transcript_source "havana"; transcript_biotype "lncRNA"; tag "basic"; transcript_support_level "1";
1      havana  exon      11869   12227   .      +      .      gene_id "ENSG00000223972"; gene_version "5"; transcript_id
      "ENST00000456328"; transcript_version "2"; exon_number "1"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype
      "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-202"; transcript_source "havana"; transcript_biotype "lncRNA"; exon_id
      "ENSE00002234944"; exon_version "1"; tag "basic"; transcript_support_level "1";
1      havana  exon      12613   12721   .      +      .      gene_id "ENSG00000223972"; gene_version "5"; transcript_id
      "ENST00000456328"; transcript_version "2"; exon_number "2"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype
      "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-202"; transcript_source "havana"; transcript_biotype "lncRNA"; exon_id
      "ENSE00003582793"; exon_version "1"; tag "basic"; transcript_support_level "1";
1      havana  exon      13221   14409   .      +      .      gene_id "ENSG00000223972"; gene_version "5"; transcript_id
      "ENST00000456328"; transcript_version "2"; exon_number "3"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype
      "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-202"; transcript_source "havana"; transcript_biotype "lncRNA"; exon_id
      "ENSE00002312635"; exon_version "1"; tag "basic"; transcript_support_level "1";
1      havana  transcript 12010   13670   .      +      .      gene_id "ENSG00000223972"; gene_version "5"; transcript_id
      "ENST00000450305"; transcript_version "2"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
      transcript_name "DDX11L1-201"; transcript_source "havana"; transcript_biotype "transcribed_unprocessed_pseudogene"; tag "basic"; transcript_support_level
      "NA";
1      havana  exon      12010   12057   .      +      .      gene_id "ENSG00000223972"; gene_version "5"; transcript_id
      "ENST00000450305"; transcript_version "2"; exon_number "1"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype
      "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-201"; transcript_source "havana"; transcript_biotype
      "transcribed_unprocessed_pseudogene"; exon_id "ENSE00001948541"; exon_version "1"; tag "basic"; transcript_support_level "NA";
```

MAPPING

Dr GIGA
56 Gb

- Homo Sapiens
 - Genome (3.2 Gb)
 - Gene Set (1.2 Gb)
- Softwares
 - STAR
 - HISAT
 - ...
- High RAM/CPU's



MAPPING

Dr GIGA
86 Gb

- Genome Indexing
 - Quick queries
 - 20M reads



- High RAM/CPU



- H.Sapiens ~30 Gb



MAPPING

- SAM/BAM files (~ 32 Gb)
 - FLAG - Information
 - RNAME - Chromosome
 - POS – Location of 1st base
 - MAPQ – Quality score
 - CIGAR - Operations

Flag	Description
1	read is mapped
2	read is mapped as part of a pair
4	read is unmapped
8	mate is unmapped
16	read reverse strand
32	mate reverse strand
64	first in pair
128	second in pair
256	not primary alignment
512	read fails platform/vendor quality checks
1024	read is PCR or optical duplicate

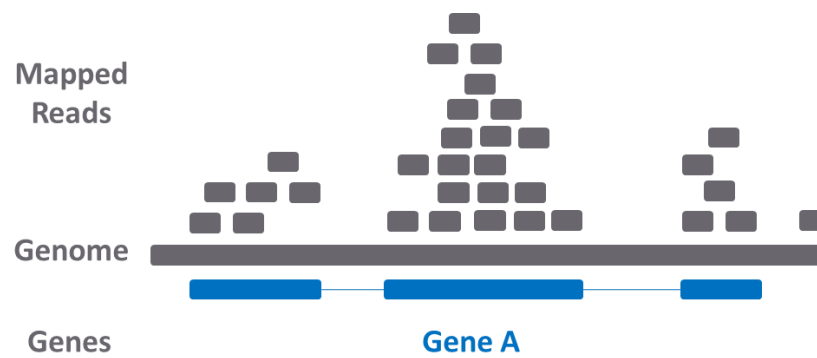
Paired-End

```

A00801:76:HGJCYDSXY:4:1544:20401:36699 99 1 3112677 255 150M = 3112770 244
CTAGGAGATAGTAGGGATTGGGAAGCAACTACTGAAAGGTCTGTGTCTTCTTTGTGGATGATAAAATATTCTGGAATTATATTGTATGCTAGGCGCACAACTTGTGACCATAGTACAGATATTCAACAGATAAATTTGTGTGCTATGA
F:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
NH:i:1 HI:i:1 AS:i:299 nM:i:0 RG:Z:SV2-CTRL2_NGS20-0393_AHGJCYDSXY_S241_L004_R1_001
  
```

QUANTIFICATION (RNA)

- Gene Expression



	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

QUANTIFICATION (RNA)

Dr GIGA
118 Gb

- Gene Expression
- « Count matrix » (3.2 Mb)
- Major output



Each column is a sample

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666
AARSA	4451	2727	3281	3121	1326	2488	2074	1657

Each row is a gene

QC MAPPING/QUANTIFICATION

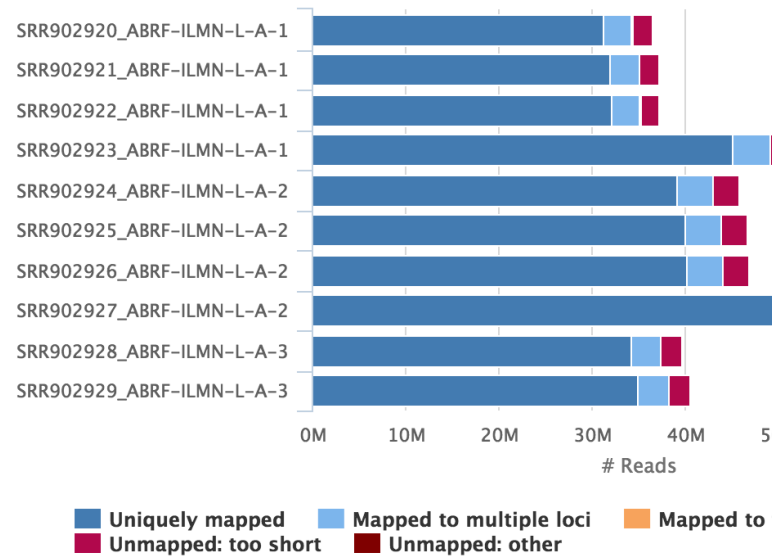


General Statistics

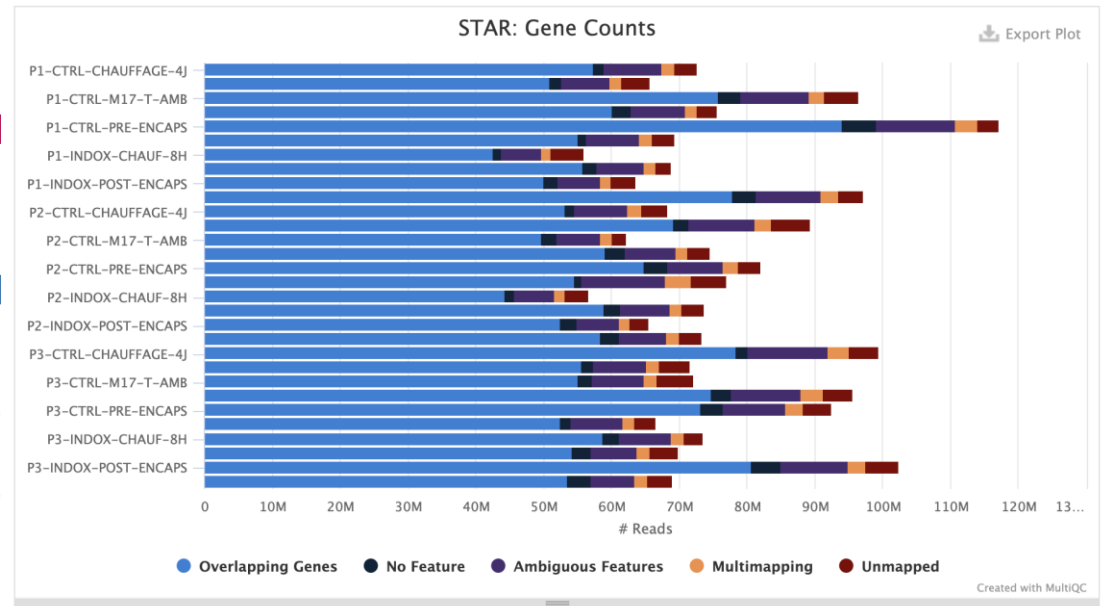
[Copy table](#)
[Configure Columns](#)
[Plot](#)
 Showing 8/8 rows and 8/10 columns.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
SRR902920	97.8%	104.4	4.0%	97.8	4.0%	78.9%	51%	104.4
SRR902921	87.1%	92.0	3.5%	87.1	3.5%	77.2%	49%	92.0

STAR Alignment Scores



STAR: Gene Counts



DOWNSTREAM ANALYSIS

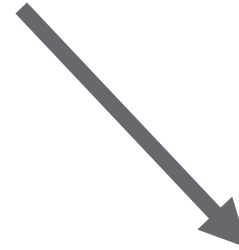
- Experimental Design

- R

- DESeq2
- EdgeR
- Voom
- ROTS
- Limma
- ...

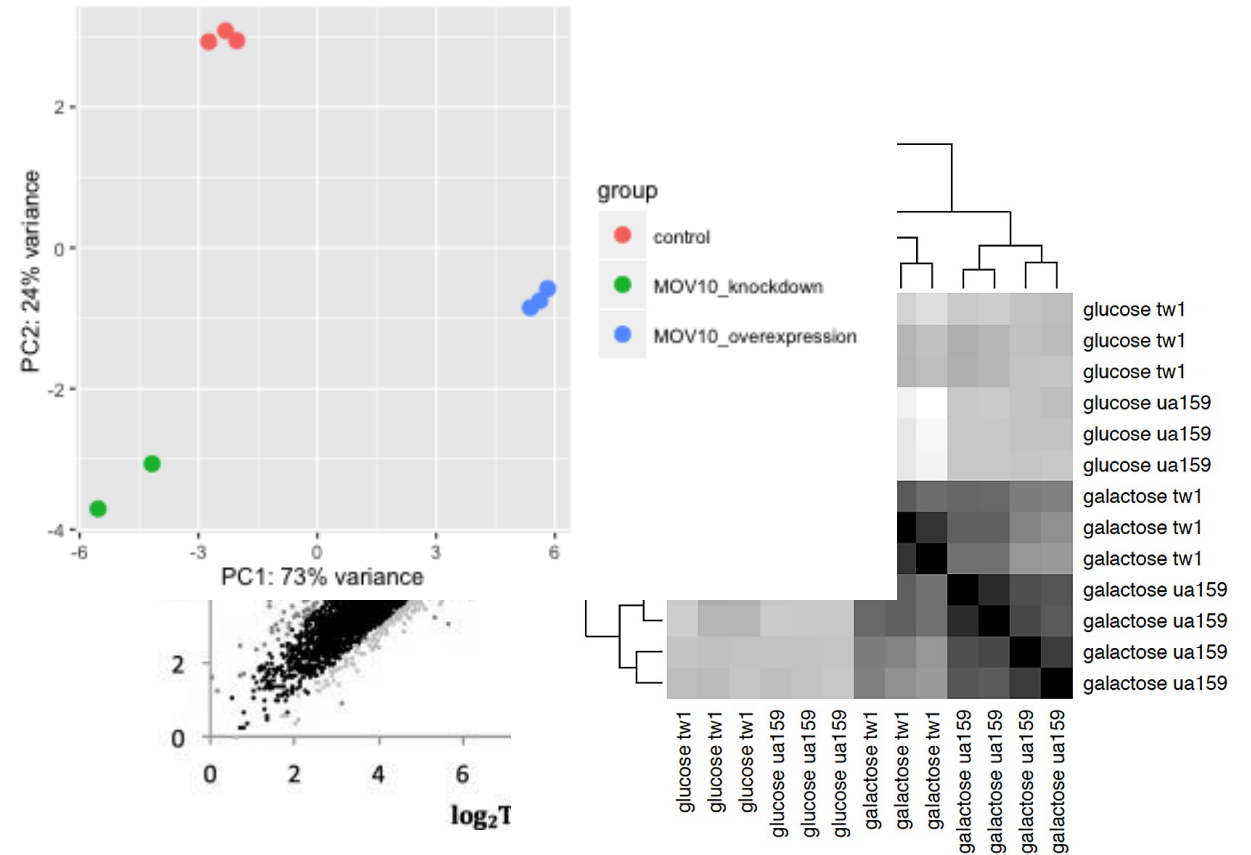


Counts



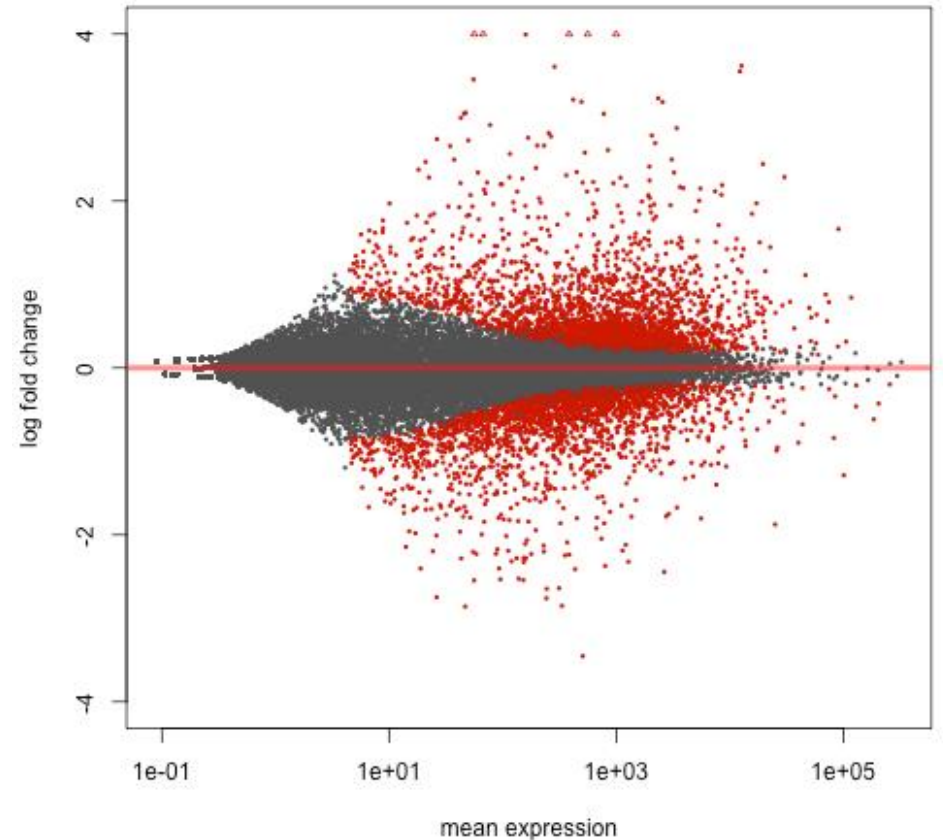
DOWNSTREAM ANALYSIS

- Experimental Design
- Clustering
 - Sample correlation
 - Euclidian distance
 - Principal Component Analysis



DOWNSTREAM ANALYSIS

- Experimental Design
- Clustering
 - Sample correlation
 - Euclidian distance
 - Principal Component Analysis
- Differential Expression Analysis
 - Pairwise comparisons
 - MA plot



DOWNSTREAM ANALYSIS

- Experimental Design
- Clustering
 - Sample correlation
 - Euclidian distance
 - Principal Component Analysis
- Differential Expression Analysis
 - Pairwise comparisons
 - MA plot
 - DE genes

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
PAX5	1531,91362	6,1280938	0.14505696	28,091789	1,23E-173	1,59E-169
SOX9	348,04912	3,5537120	0.15748166	20,475861	3,53E-93	2,27E-89
PDX1	830,75570	-1,8973788	0.12438094	-15,203018	3,38E-52	1,45E-48
ISL1	655,25202	-1,9729198	0.13344796	-14,715372	5,14E-49	1,65E-45
ARX	526,74210	2,2554297	0.15754888	14,227414	6,19E-46	1,59E-42

Summary

out of 21769 with nonzero total read count

adjusted p-value < 0.1

LFC > 0 (up) : 985, 4.5%

LFC < 0 (down) : 929, 4.3%

outliers [1] : 0, 0%

low counts [2] : 8914, 41%

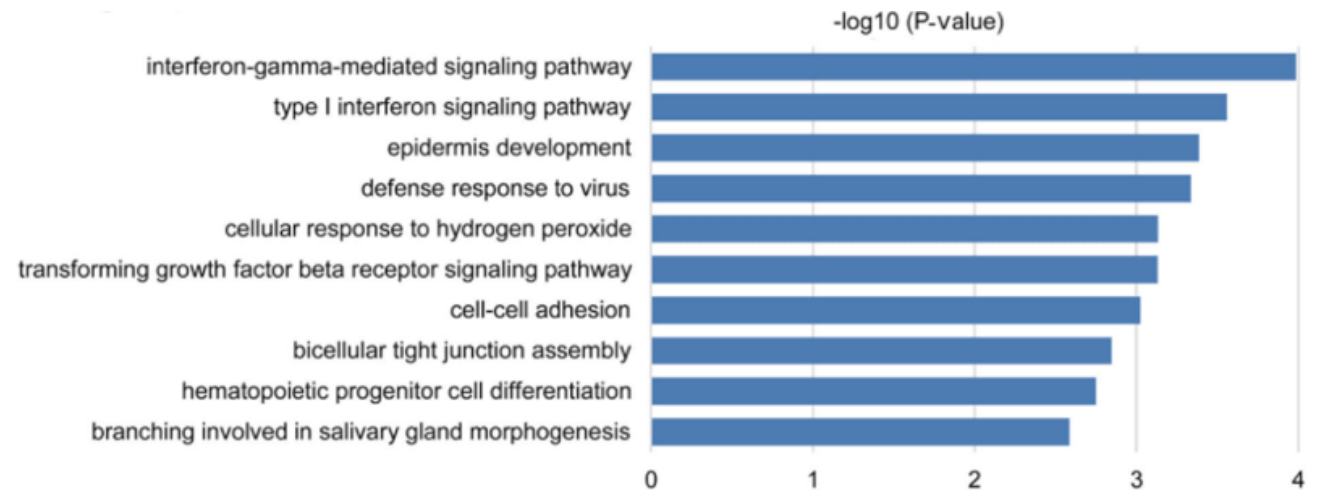
(mean count < 8)

[1] see 'cooksCutoff' argument of ?results

[2] see 'independentFiltering' argument of ?results

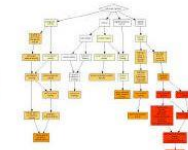
DOWNSTREAM ANALYSIS

- Biological meaning
- Gene ontology / Gene Set Enrichment Analysis
 - GSEA
 - Enrichr
 - GOrilla
 - PANTHER
 - ...

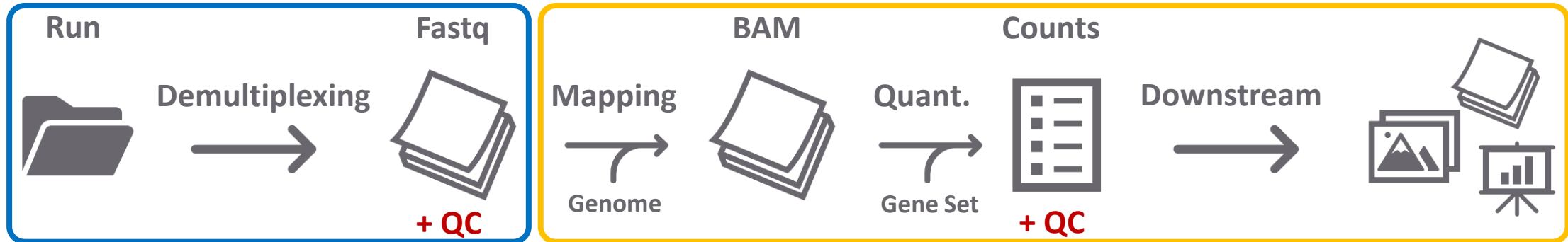


GORILLA

Gene Ontology enRIchment anaLysis and visualiZation tool



SUMMARY



REPRODUCIBILITY

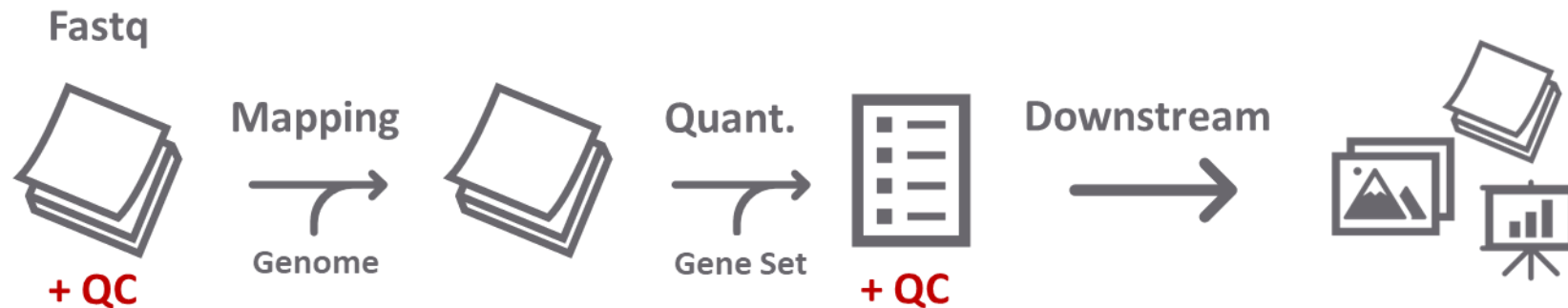
PIPELINES & CONTAINERS

REPRODUCIBILITY

- Pipelines
 - Set of successive actions
 - Softwares
 - Parameters
 - References

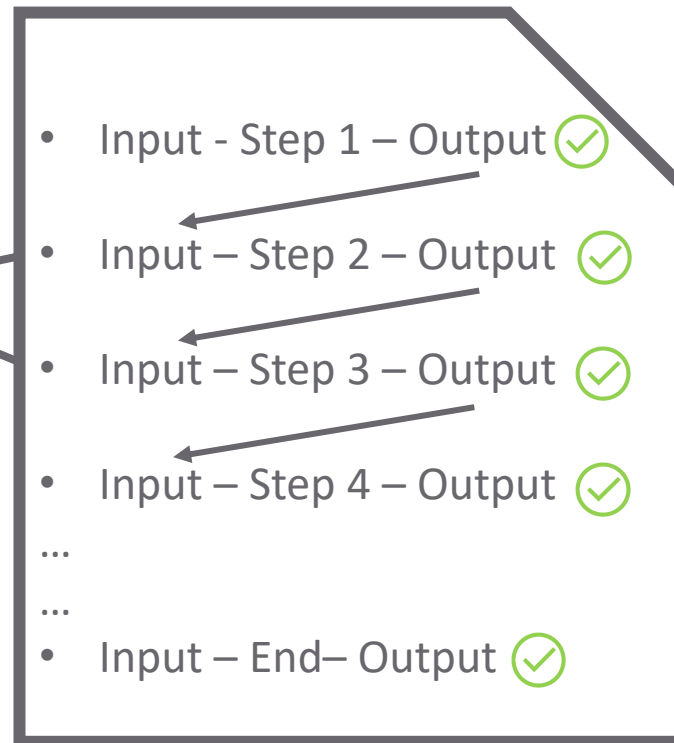
nextflow

Snakemake



REPRODUCIBILITY

- Pipelines
 - Scripts

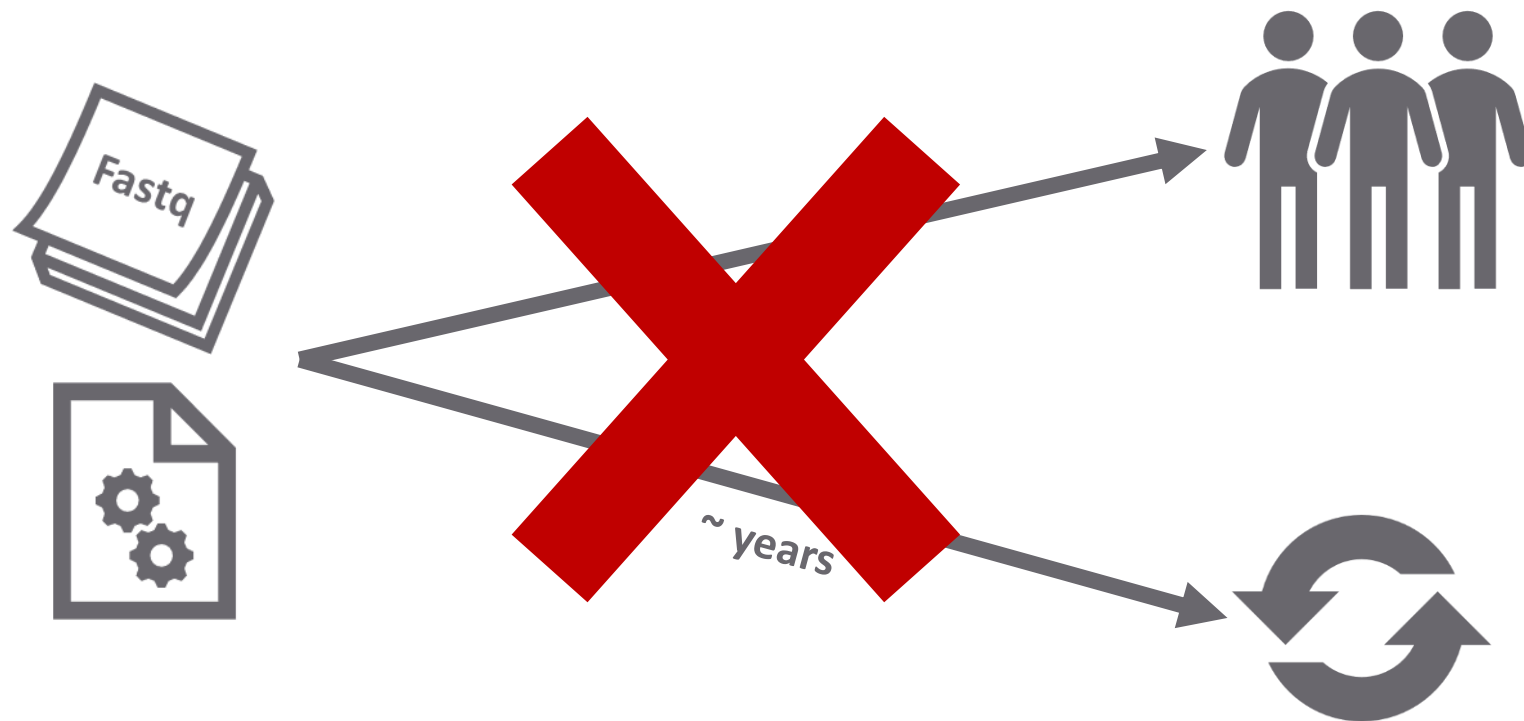


nextflow

Snakemake

nf-core 

REPRODUCIBILITY



REPRODUCIBILITY

- Variability

- Updates / Versioning
 - Softwares
 - References

STAR 2.7.5b - 2020/08/01
STAR 2.7.5c - 2020/08/16
STAR 2.7.6a - 2020/09/19

- Compatibility
- Format
- Knowledge

List of currently available archives

- [Ensembl GRCh37](#): Full Feb 2014 archive with BLAST, VEP and BioMart
- [Ensembl 101: Aug 2020](#) - this site
- [Ensembl 100: Apr 2020](#)
- [Ensembl 99: Jan 2020](#)
- [Ensembl 98: Sep 2019](#)
- [Ensembl 97: Jul 2019](#)
- [Ensembl 96: Apr 2019](#)
- [Ensembl 95: Jan 2019](#)
- [Ensembl 94: Oct 2018](#)
- [Ensembl 93: Jul 2018](#)
- [Ensembl 92: Apr 2018](#)
- [Ensembl 91: Dec 2017](#)

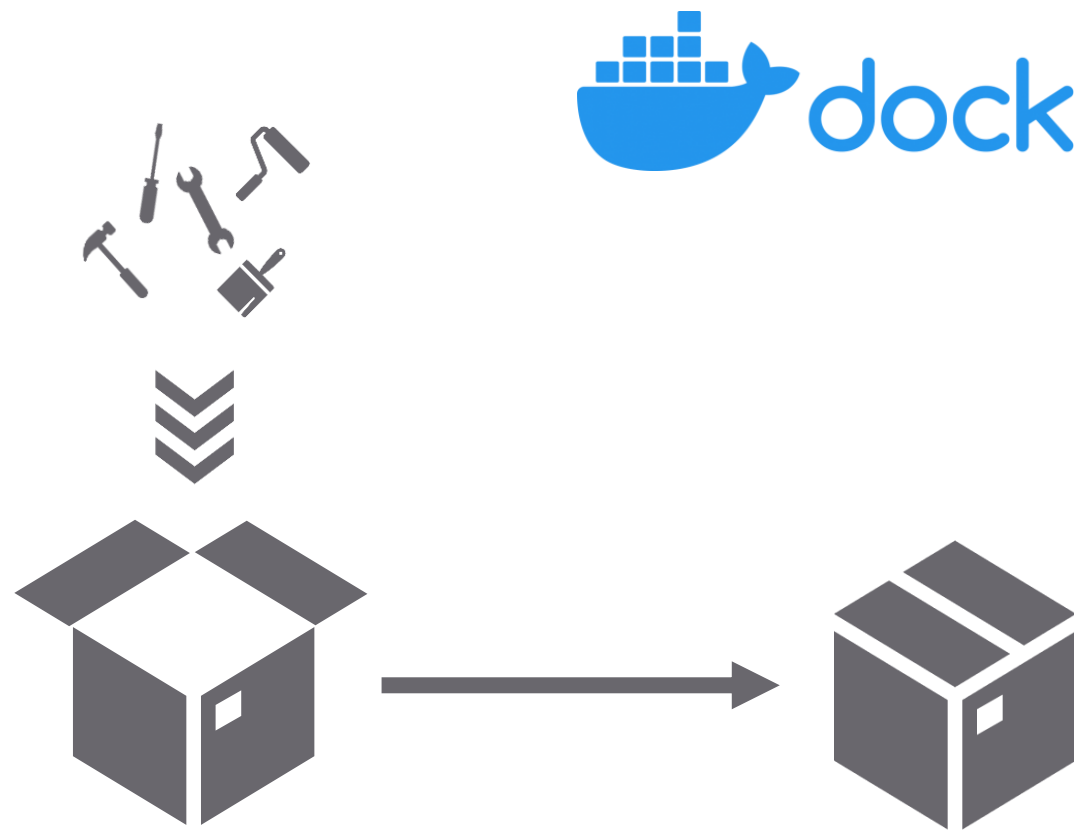
REPRODUCIBILITY

- CONTAINERS

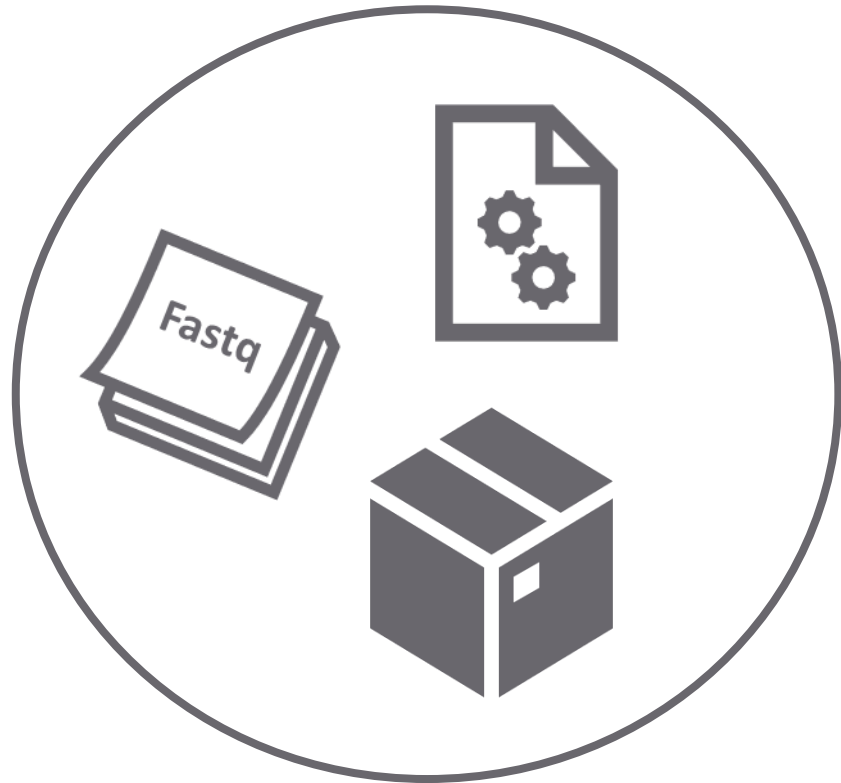
- Docker
- Singularity

- Softwares

- Versions



REPRODUCIBILITY



DATA DEPOSITORY

- Gene Expression Omnibus (NCBI)
- ArrayExpress (EMBL-EBI)



THANK YOU FOR YOUR ATTENTION