DATA PROCESSING IN GENOMICS: FROM THE SEQUENCERS TO THE RESULTS

GIGA Doctoral School for Health Sciences
Introduction to scientific computing

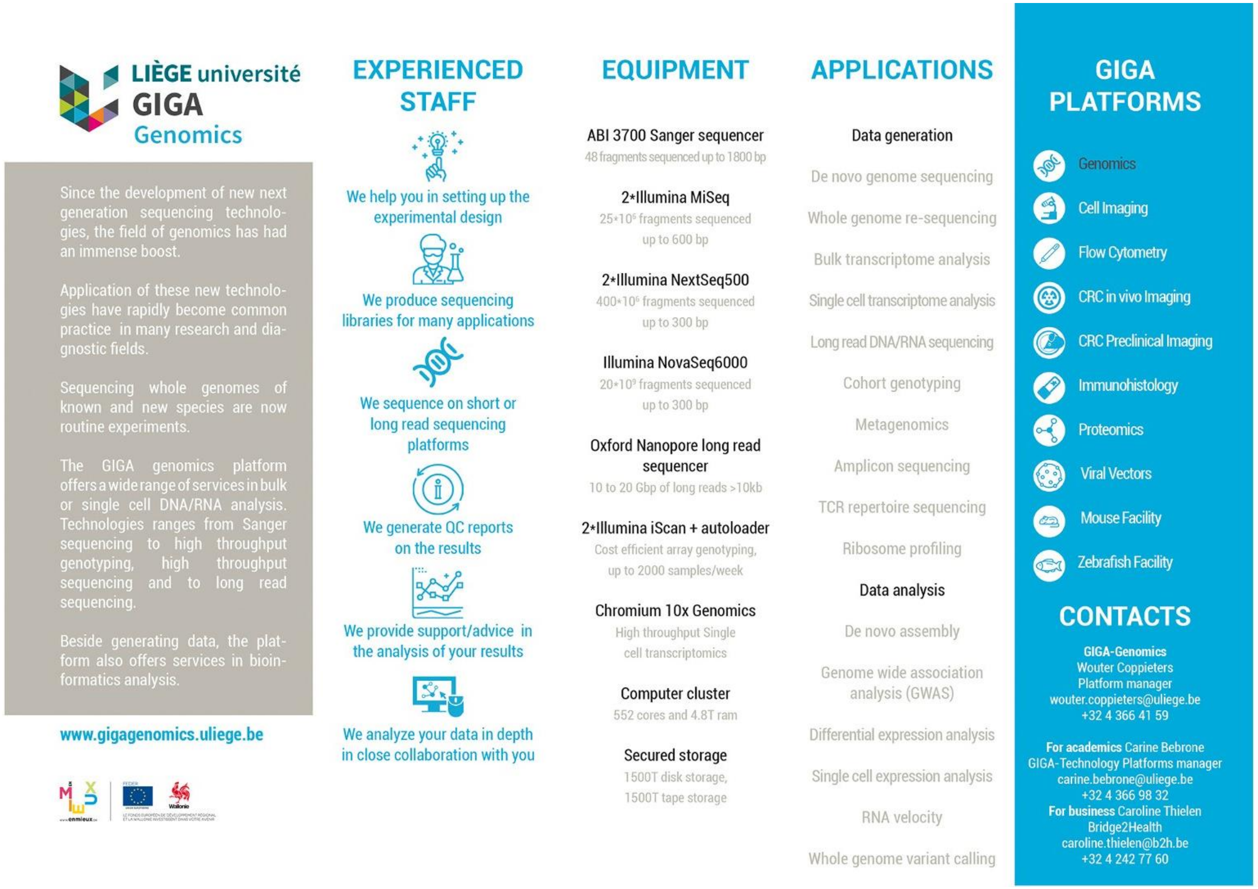
ARNAUD LAVERGNE
GIGA-Genomics
Bioinformatic team

LIÈGE université
GIGA institute

LIÈGE université
GIGA
Genomics

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 ...



LIÈGE université
GIGA
Genomics

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
$25*10^6$ fragments sequenced up to 600 bp

2*Illumina NextSeq500
$400*10^6$ fragments sequenced up to 300 bp

Illumina NovaSeq6000
$20*10^9$ 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

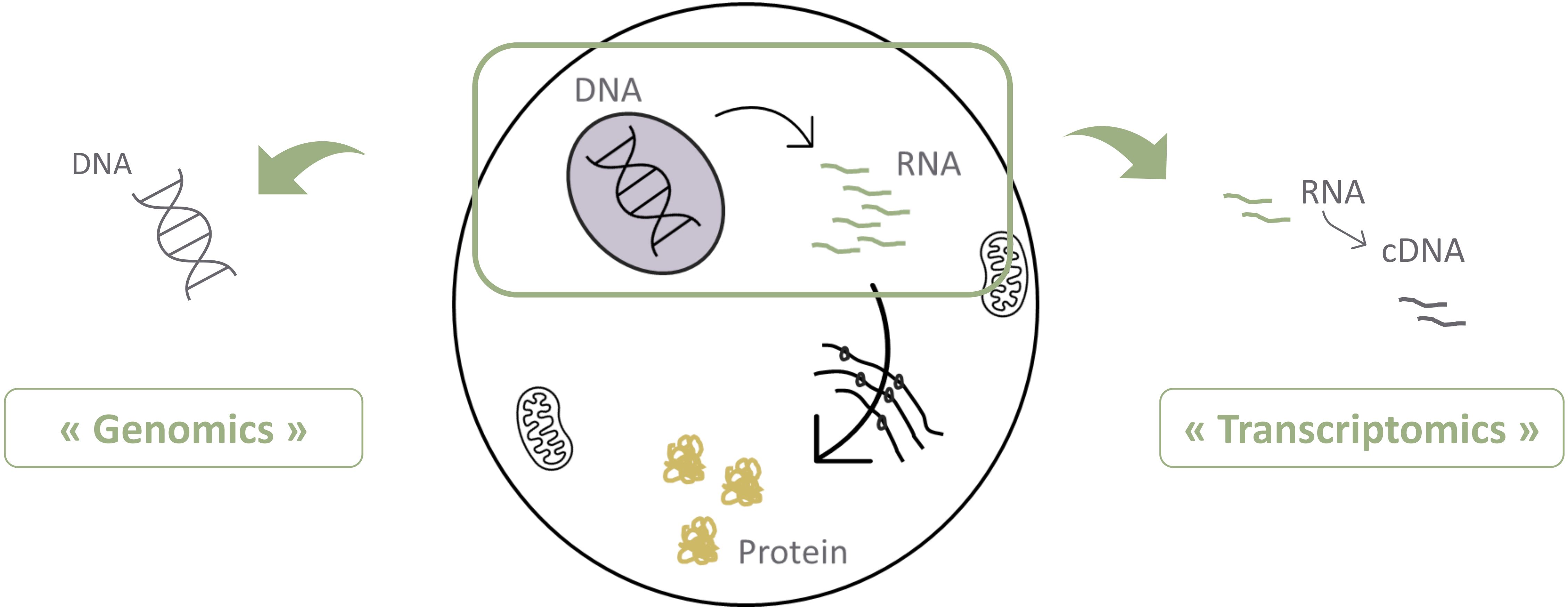
- Genomics
- Cell Imaging
- Flow Cytometry
- CRC in vivo Imaging
- CRC Preclinical Imaging
- Immunohistology
- Proteomics
- Viral Vectors
- Mouse Facility
- Zebrafish Facility

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

SAMPLING

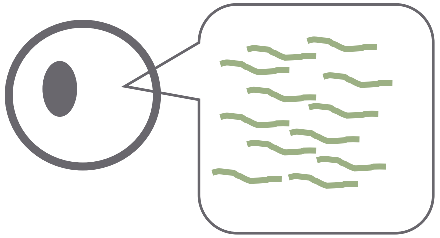


« Genomics »

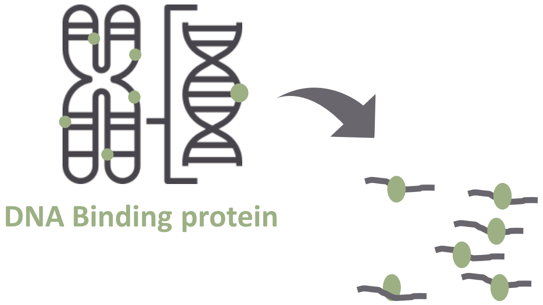
« Transcriptomics »

APPLICATIONS

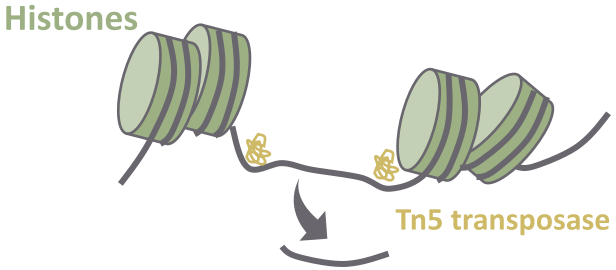
RNA-Seq

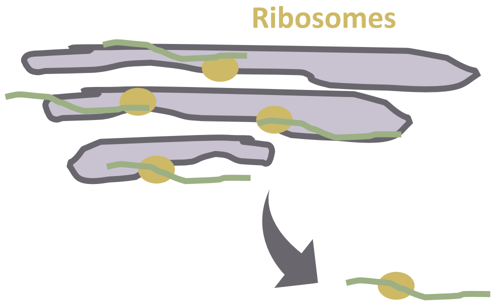


ChIP-Seq

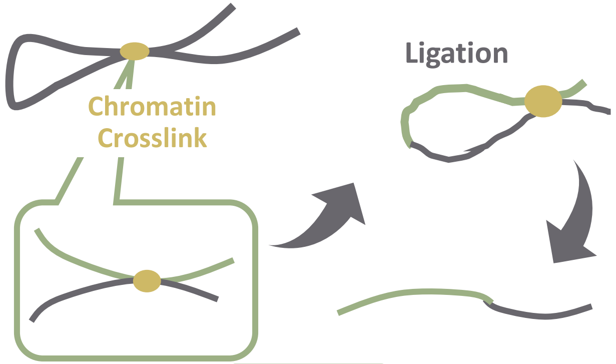


ATAC-Seq

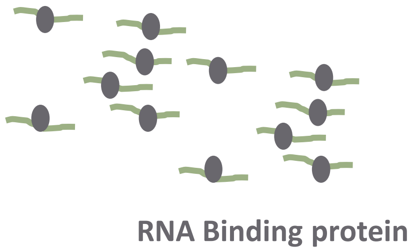




Ribo-Seq



Hi-C



CLIP-Seq

SEQUENCERS

MiSeq
540 Mb -15 Gb
4 – 56 hours

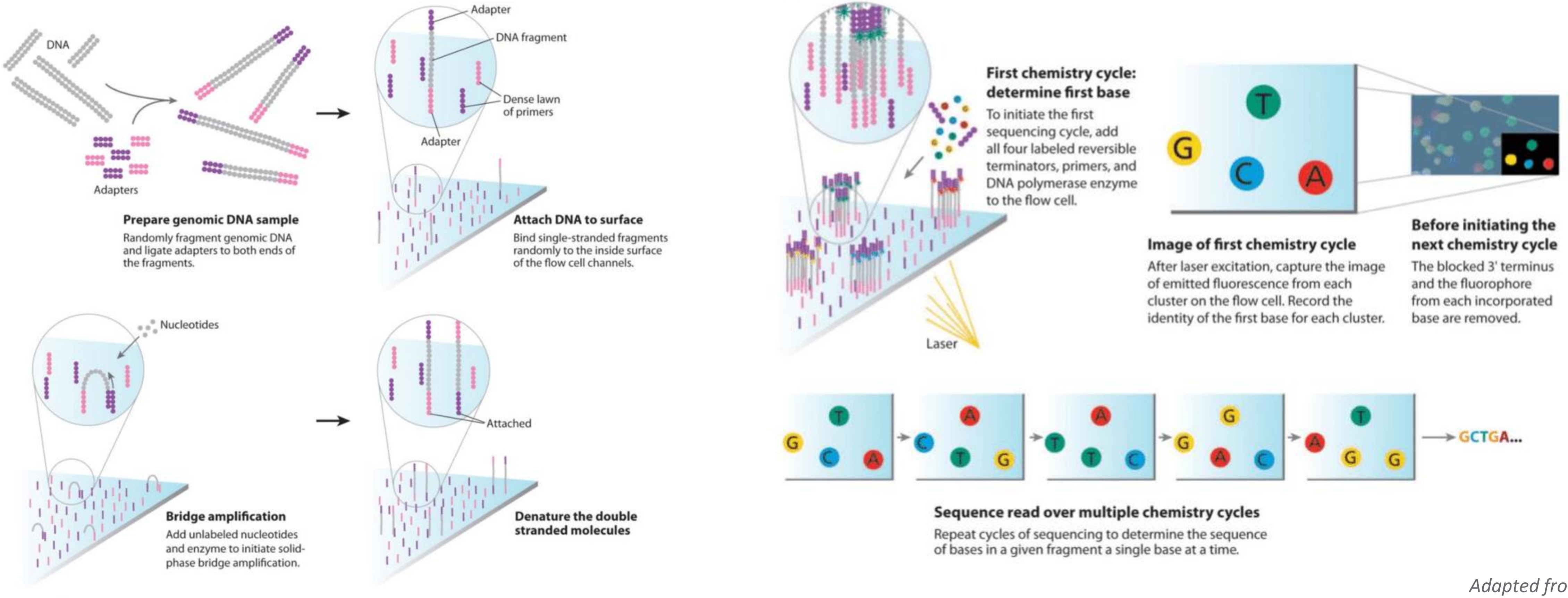
HiSeq
105 Gb - 1,5 Tb
1 – 3,5 days

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

NovaSeq
65 Gb – 3 Tb
13 – 44 hours

Adapted from Illumina

SEQUENCING (HTS/NGS)



Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Denature the double stranded molecules

First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

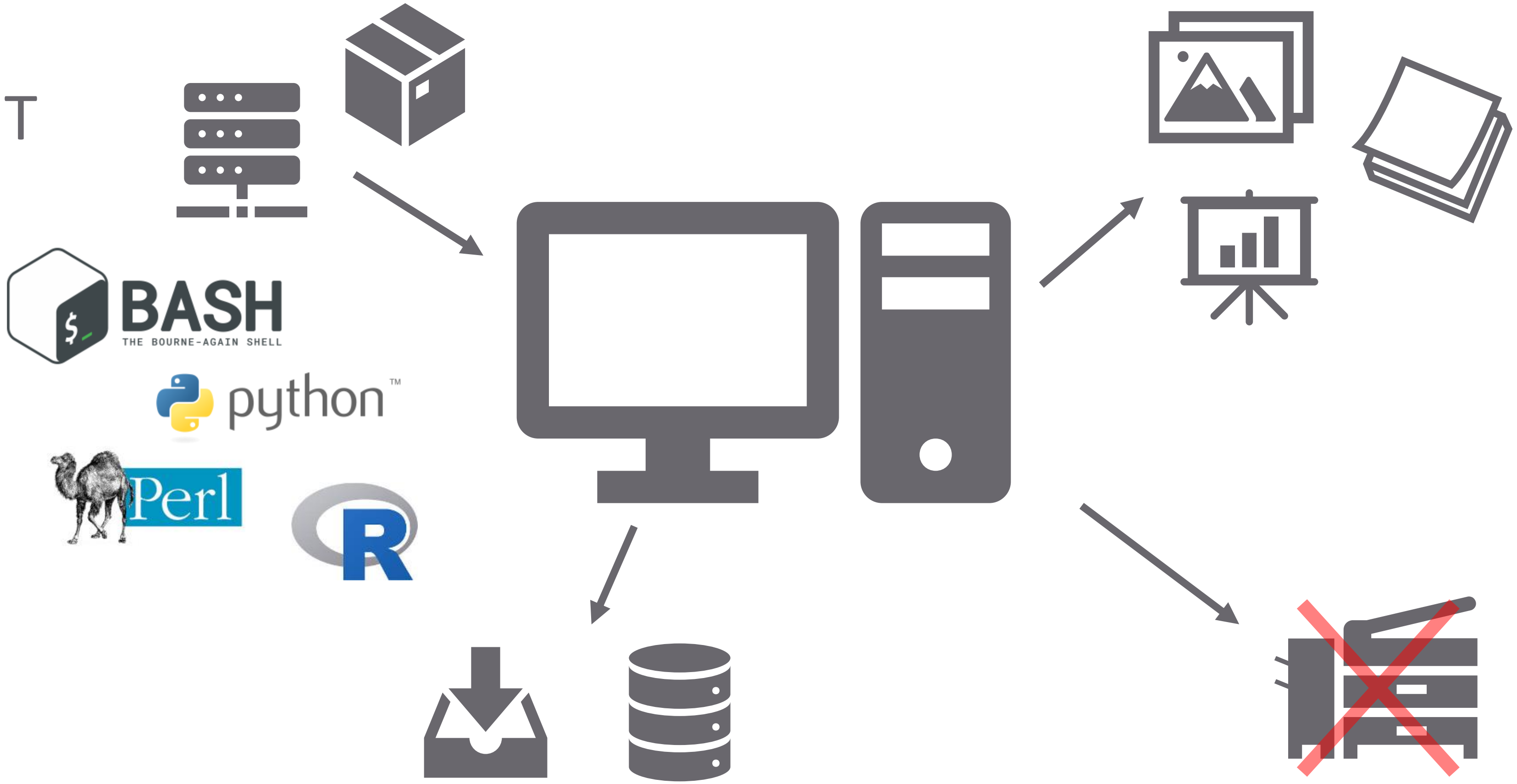
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.

Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

Adapted from Illumina

COMPUTING

- DATA MANAGEMENT
  - DATA PROCESSING
  - STORAGE
  - RESULTS
- REPRODUCTIBILITY
  - PIPELINES
  - CONTAINERS



DATA MANAGEMENT

PROCESSING & RESULTS

DATA MANAGEMENT

LAPTOP	LAB COMPUTER	CLUSTER/STORAGE
250Gb – 1Tb	1 - 5 Tb	1.5 Pb
4-8 Go RAM	32 - 128 Go RAM	256 Go RAM
4-8 CPUs	8-16 CPUs	32 CPUs

DATA MANAGEMENT

Example:

- 1.2 Tb of data
- 12.8 Billions of reads
- 612 samples
  - Multiple experiments
  - Unique combinations of indexes

- « Run »
- No storage on device

DATA FORMAT

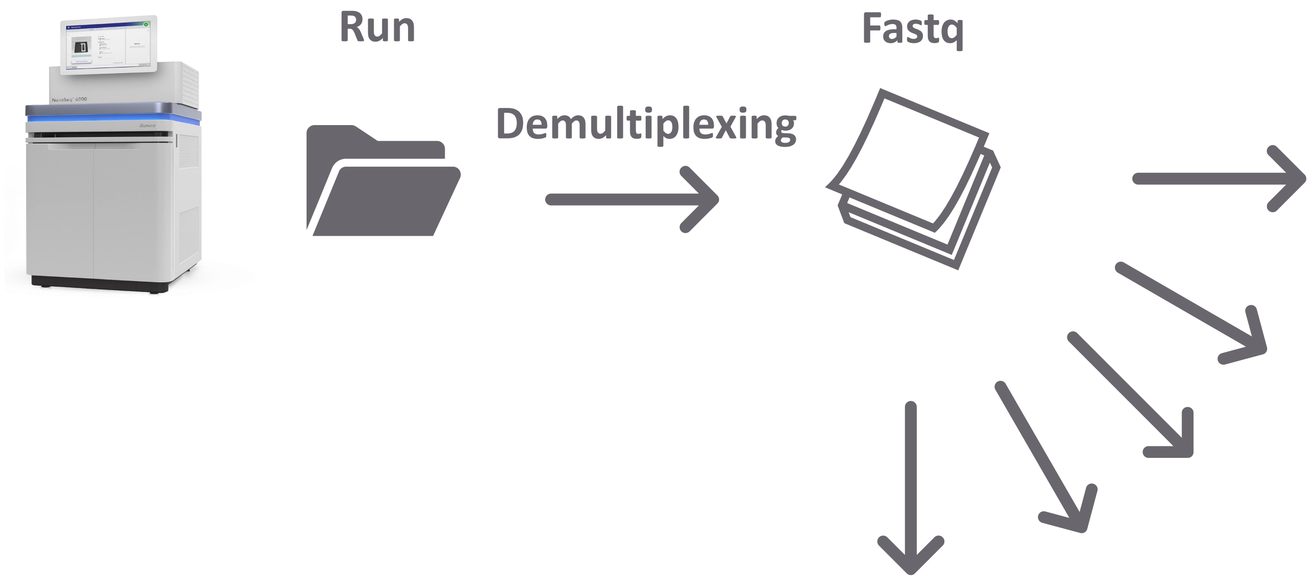
- DEMULTIPLEXING
  - Reads → Samples
  - ~ 20M / sample
  - « Fastq »
    - Identifier
    - Sequence
    - Separator
    - Quality score
      - Ascii +33
  - ~ 1.2 Tb x2 => 2.4 Tb

```
@A00801:49:H2VGKDSXY:2:1101:1624:1016 1:N:0:NGCTTAAG+TCGTGACC
CTTCTGGAGAGGAGTTCCTCTGATATGAATTAAGGTTTTCCCTTCTGTCATGACCAGAAGAAGGTTTTATCTGTGCCACACTACTTTTCATTTTCTGTTGCCAGTTGGTCCAATA
AATCAAAGATGNTTCAAACTGGTCCAAATAACAAGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFF#FFFFFFFFFFFFFFFF
FFFFFFFF
@A00801:49:H2VGKDSXY:2:1101:2022:1016 1:N:0:NGCTTAAG+TCGTGACC
TGACAAAAAGATACCTCATTTATGGGGAAATTGAGGAAGATACATATACAAGCACCCCAACCCATATTTAACATATTTGGCAATAACTCCCTTCCCATTCTTCCCCCTCCAATTT
TCAAATAGTAGNTTTTTAAAAAATTAAAGACATGTC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFF,FFFFFF:FFFFFFF,#FFFFFFF,FFFFFFFF:F
:FF,FF
@A00801:49:H2VGKDSXY:2:1101:4137:1016 1:N:0:NGCTTAAG+TCGTGACC
TTTTTTTGCCTTTTCAAGTGTTTATTTTTATACATTTTTTTGTATTAAAAAGAAAAGCATAATTACCACAAATTACAAAGGACTAAAGCAGGACTAGAATAATGAATGAATCAC
TTCAGCCTGGAANGCAGATACTCTCAATAATATTAAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFF:FFFFF:FFFFFFFFFFFFFFFFFFFFFFF:FFF,FFFFFFFFFFF:FFF#FFFFF:FFFFFFFFFFF
FFFFFF
```

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%

OVERVIEW



Run

Demultiplexing

Fastq

⚠ Fastq.gz ⚠

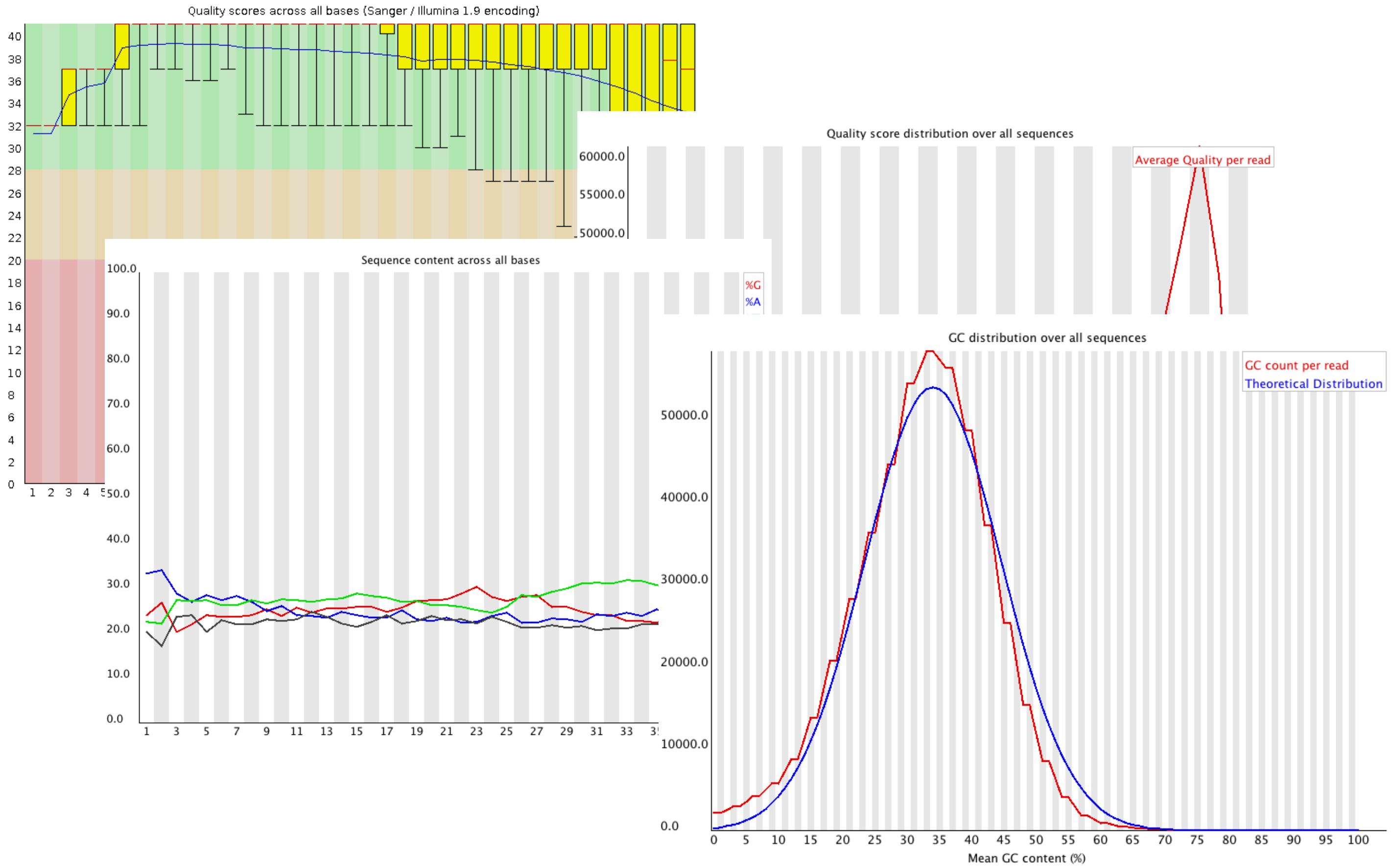
DATA TYPE

EXPERIMENTAL DESIGN

QC SEQUENCING

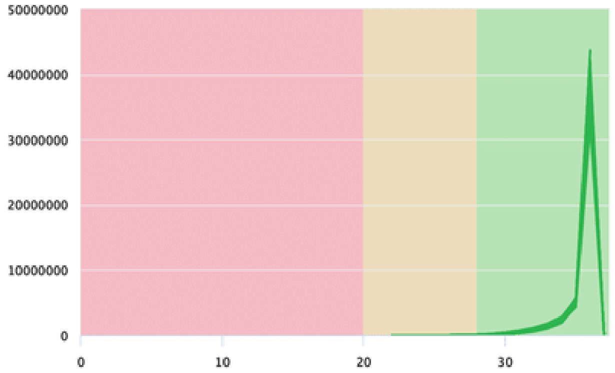
FastQC

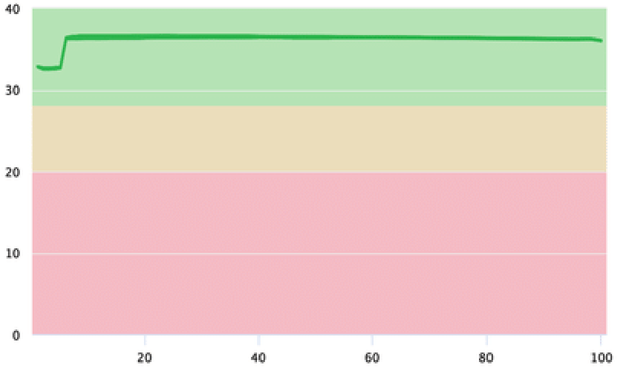
- Number of reads
- Base calling quality
- Sequece quality
- GC content
- Sequence length
- Duplication levels
- Adapter content
- Overrepresented sequences
- …

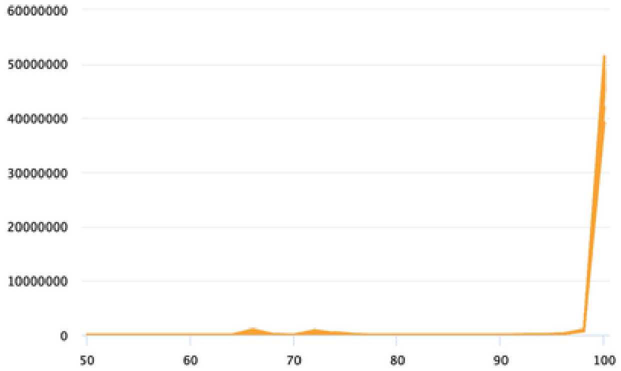


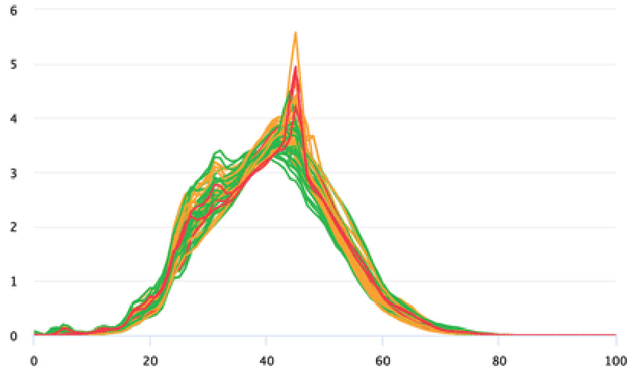
QC SEQUENCING

MultiQC





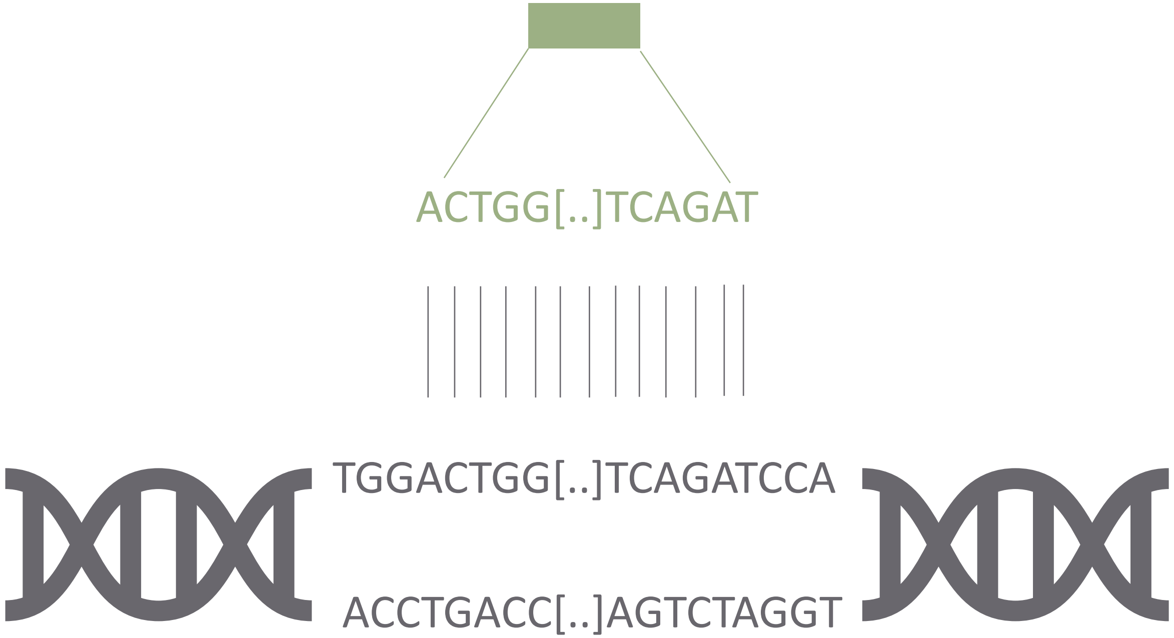




READS

Reference

- Genome sequence
- Gene set



ACTGG[..]TCAGAT

TGGACTGG[..]TCAGATCCA

ACCTGACC[..]AGTCTAGGT

Alignment / mapping

e!Ensembl

UCSC

REFERENCE

Genome (FASTA)

```
>1 dna:chromosome chromosome:GRCh38:1:1:248956422:1 REF
CCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCCAACCCCAACCCCAACCCCAACCCCAACCCCA
ACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCT
AACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGGTCT
GACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACA
ACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGC
CGTTGCAAAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGC
AGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGG
CGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGC
CGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCA
GAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGT
GGAGGCGTGGCGCAGGCGCAGAGAGGCGCACCGCGCCGGCGCAGGCGCAGAGACA
CATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAGACGCAAGCCTACG
GGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGGGCTG
GGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGG
CGCAGAGACGGGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTT
GCAGCCGGGCACTACAGGACCCGCTTGCTCACGGTGCTGTGC
```

Gene Set (GTF)

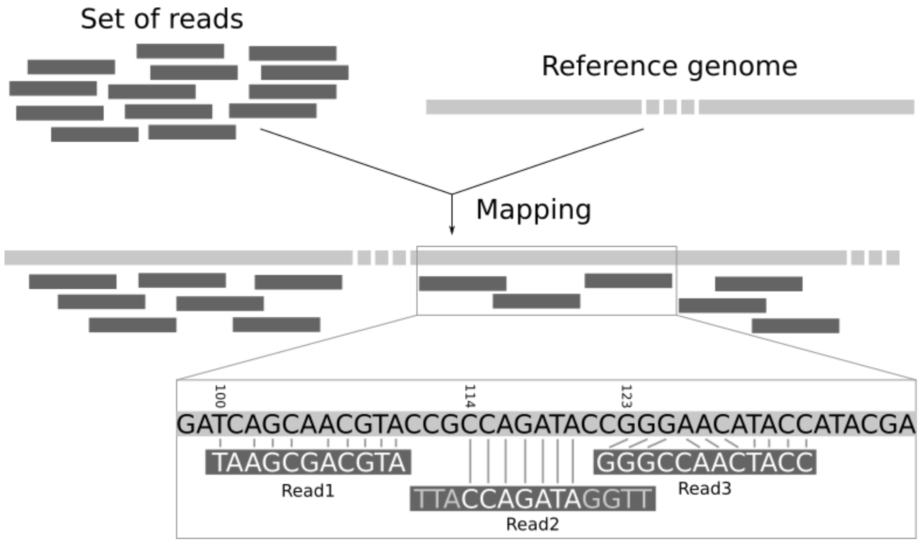
```
#!genome-build GRCh38.p12
#!genome-version GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.27
#!genebuild-last-updated 2019-03
1	havana	gene	11869	14409	.	+	.
gene_id "ENSG00000223972"; gene_version "5"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
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";
```

INDEXING

- Genome Indexing
  - Quick queries
  - « 20M reads »
- High RAM/CPUs
- H.Sapiens ~30 Gb

MAPPING

- Softwares
  - STAR
  - HISAT
  - Bowtie
  - BWA
  - …



- High RAM/CPUs

MAPPING

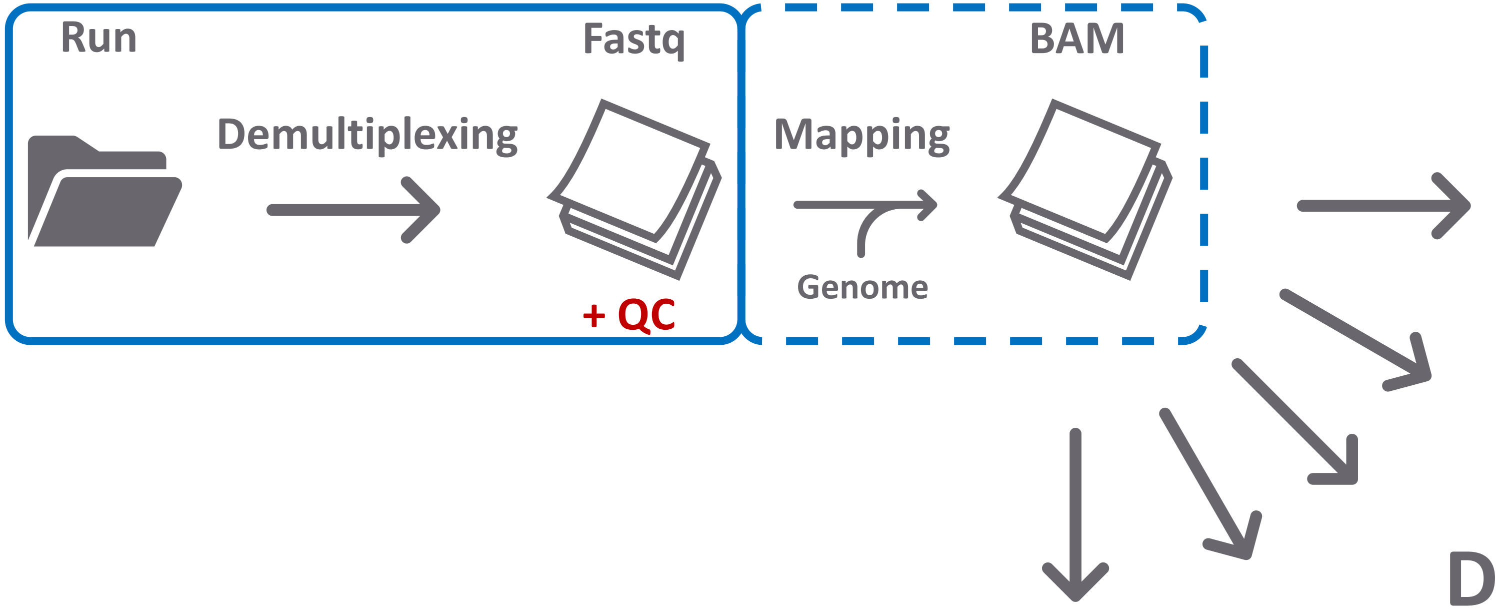
- SAM/BAM files
  - FLAG - Information
  - RNAME - Chromosome
  - POS – Location of 1st base
  - MAPQ – Quality score
  - CIGAR - Operations

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
CTAGGAGATAGTAGGGATTGGGAAGCAACTACTGAAAGGTCTGTGTCTTCTTTGTGGATGATAAAATATTCTGGAATTATATTGTATGCTAGGCGCACAATCTTGTGACCATAGTACAGATATTCAACAGATAAATTTTGTGTGCTATGA
F:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFF:FFFFF
NH:i:1          HI:i:1          AS:i:299          nM:i:0          RG:Z:SV2-CTRL2_NGS20-O393_AHGJCYDSXY_S241_L004_R1_001
```

OVERVIEW



Run

Demultiplexing

Fastq

+ QC

Mapping

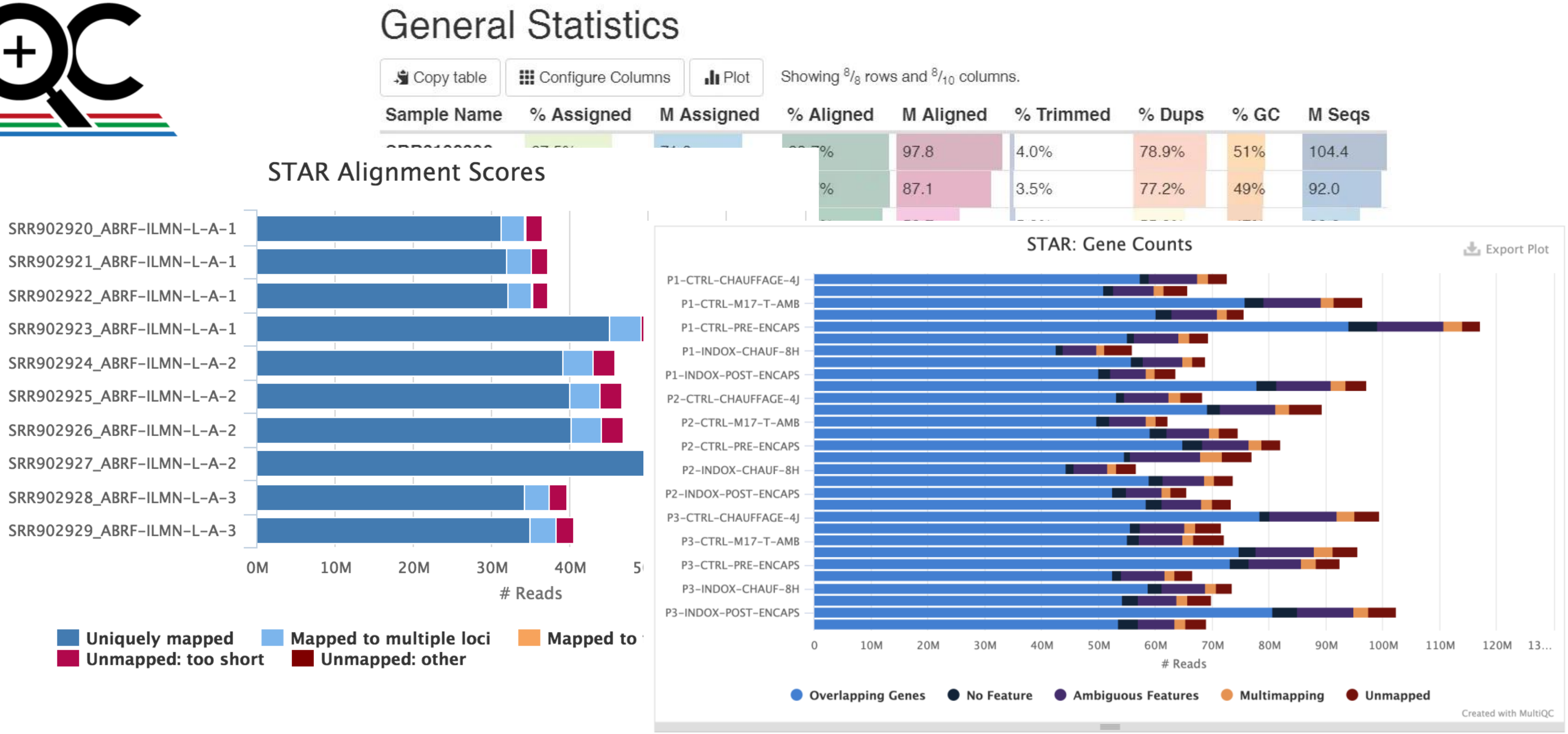
Genome

BAM

DOWNSTREAM ANALYSES

QC MAPPING

MultiQC



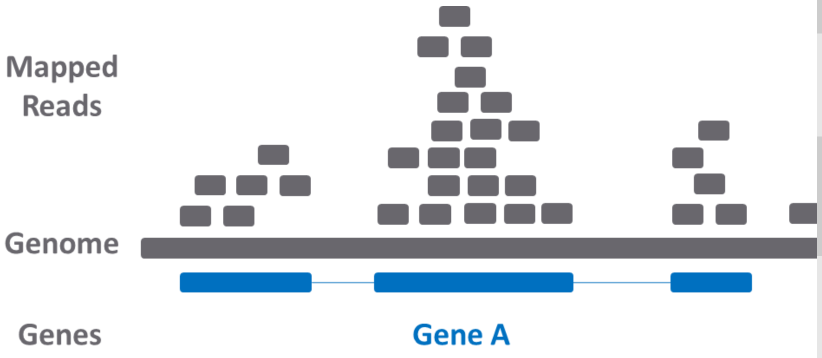
General Statistics

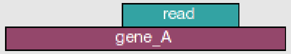
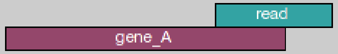
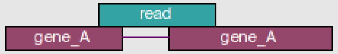
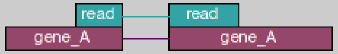
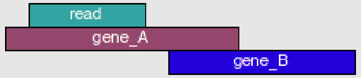
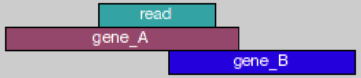
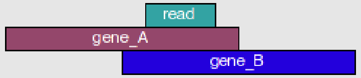
Showing $^8/_8$ rows and $^8/_{10}$ columns.

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

RNA - QUANTIFICATION

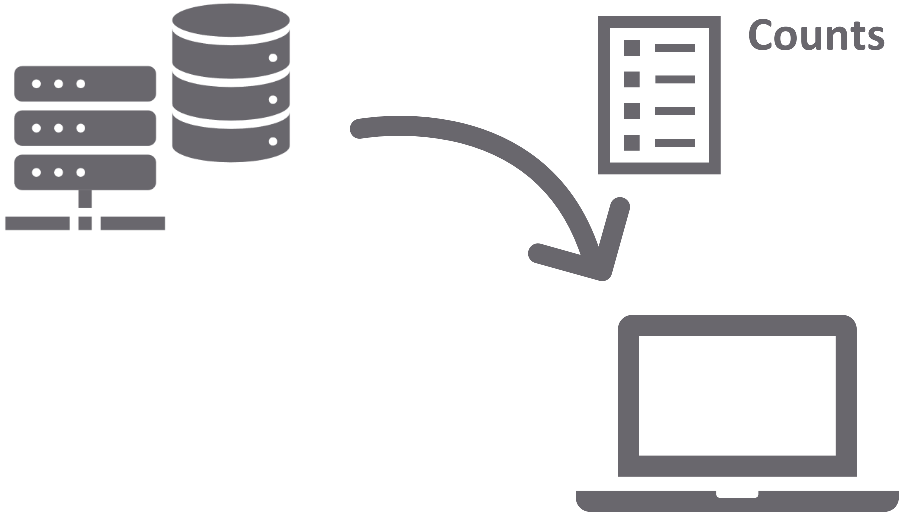
- Gene Expression



	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

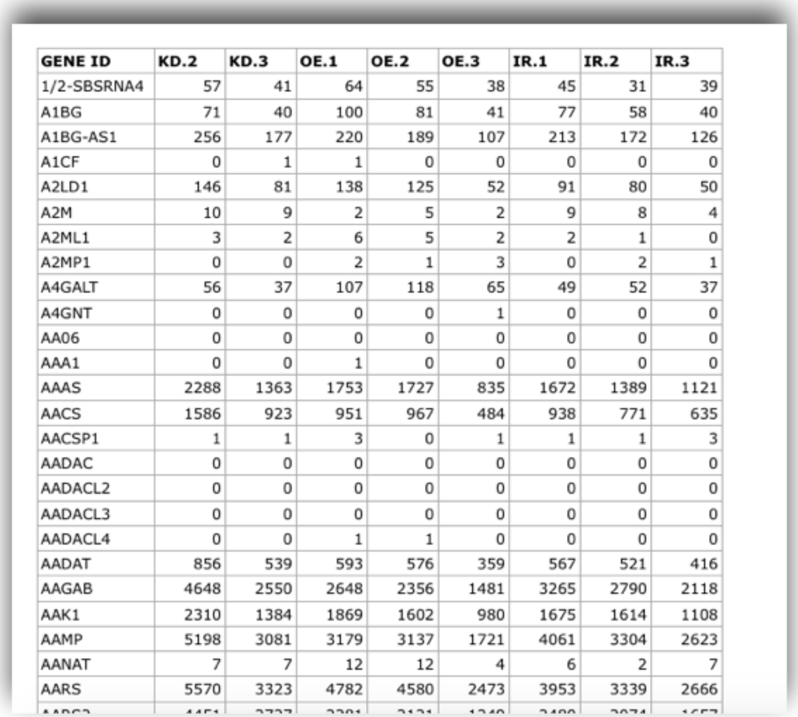
RNA - QUANTIFICATION

- Gene Expression
- « Count matrix » (3.2 Mb)
- <u>Major output</u>



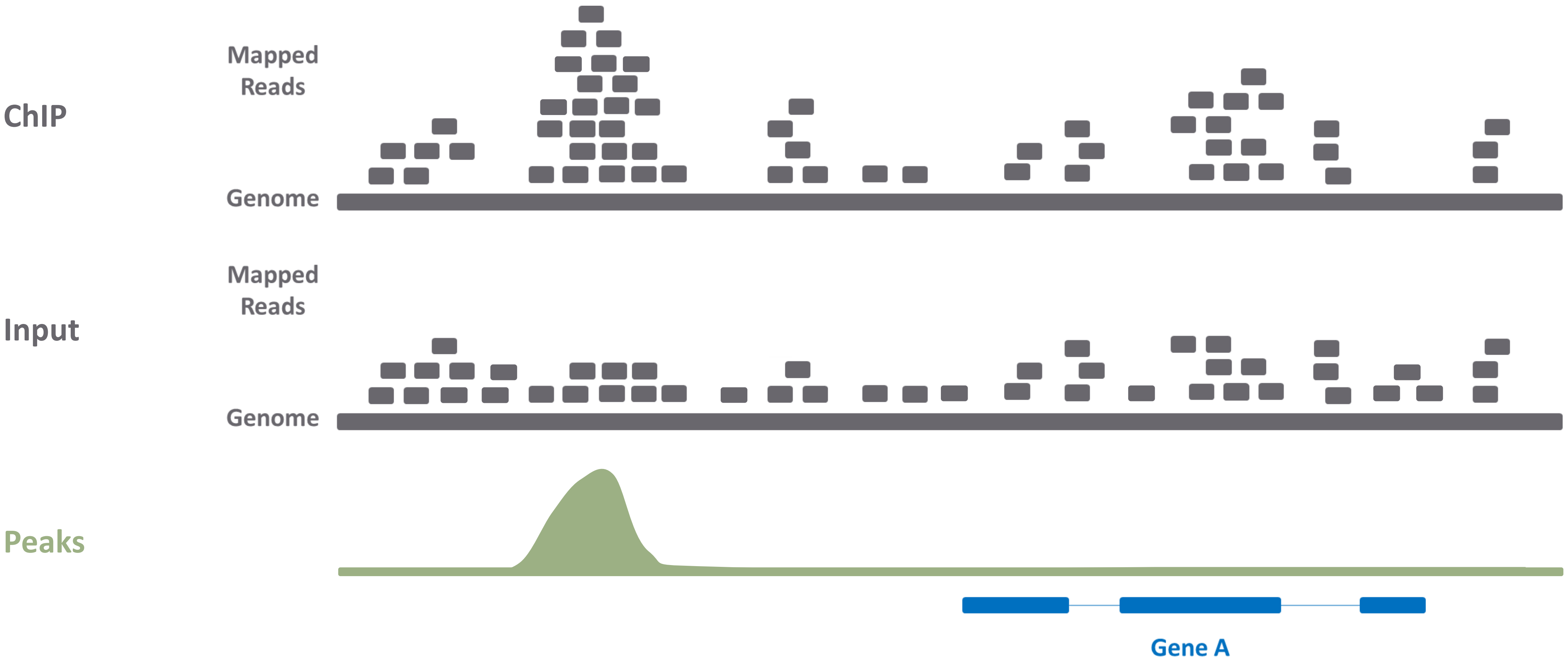
Each column is a sample

Each row is a gene



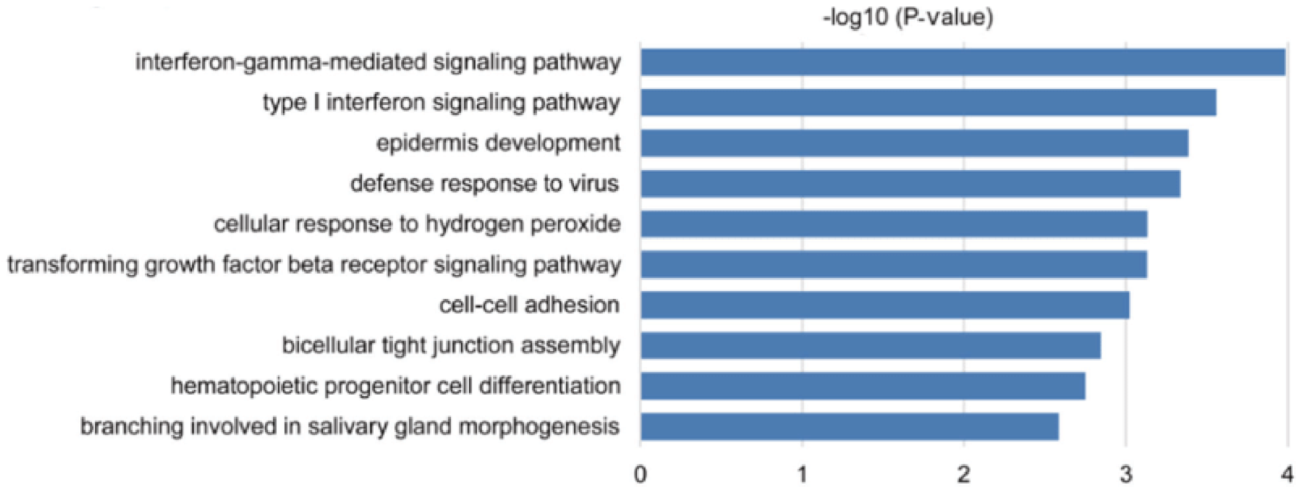
GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	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

DNA - PEAK CALLING



DOWNSTREAM ANALYSIS

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



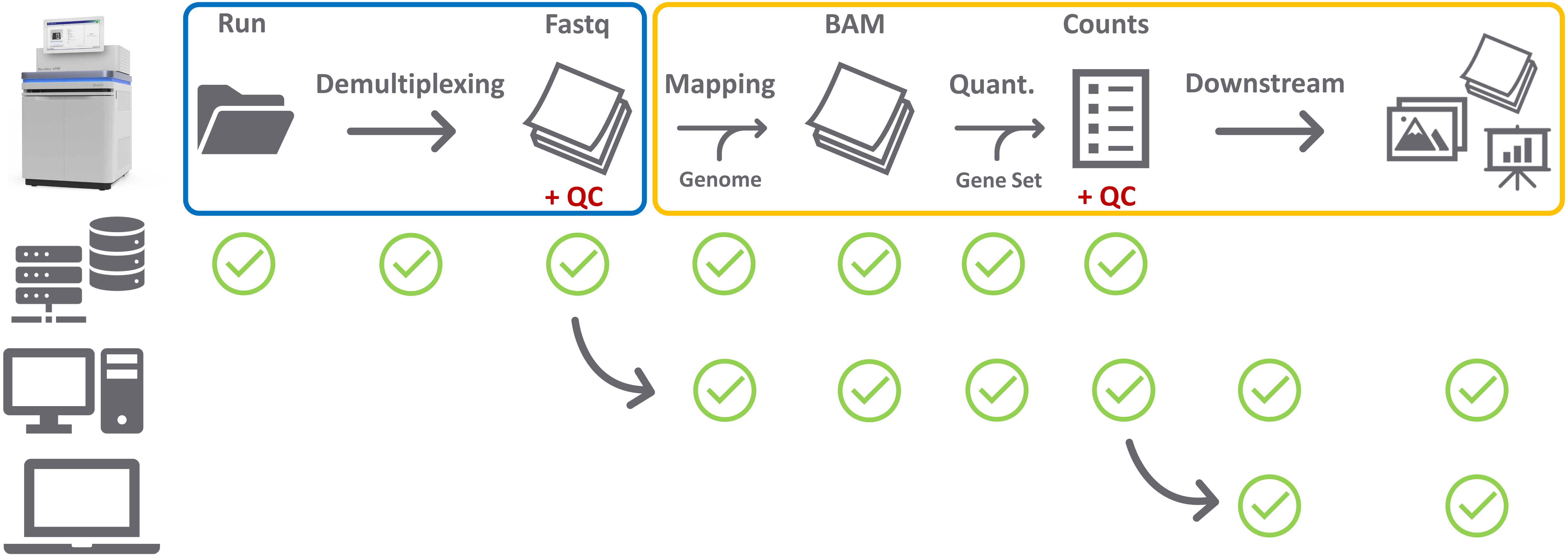
GSEA Gene Set Enrichment Analysis

Enrichr

GORILLA Gene Ontology enRIchment anaLysis and visuaLizAtion tool

PANTHER Classification System

SUMMARY



Run

Demultiplexing

Fastq

+ QC

BAM

Counts

Mapping

Genome

Quant.

Gene Set

+ QC

Downstream

REPRODUCTIBILITY

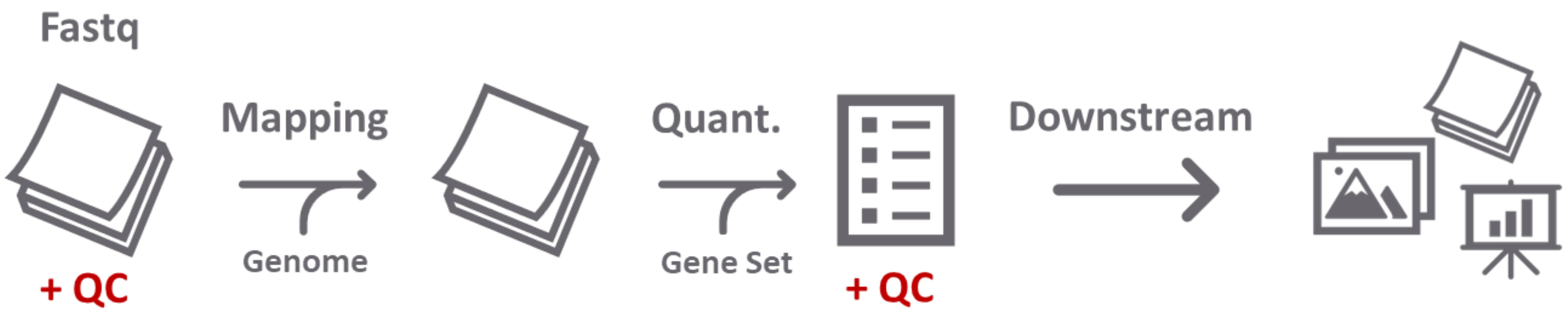
PIPELINES & CONTAINERS

REPRODUCTIBILITY

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

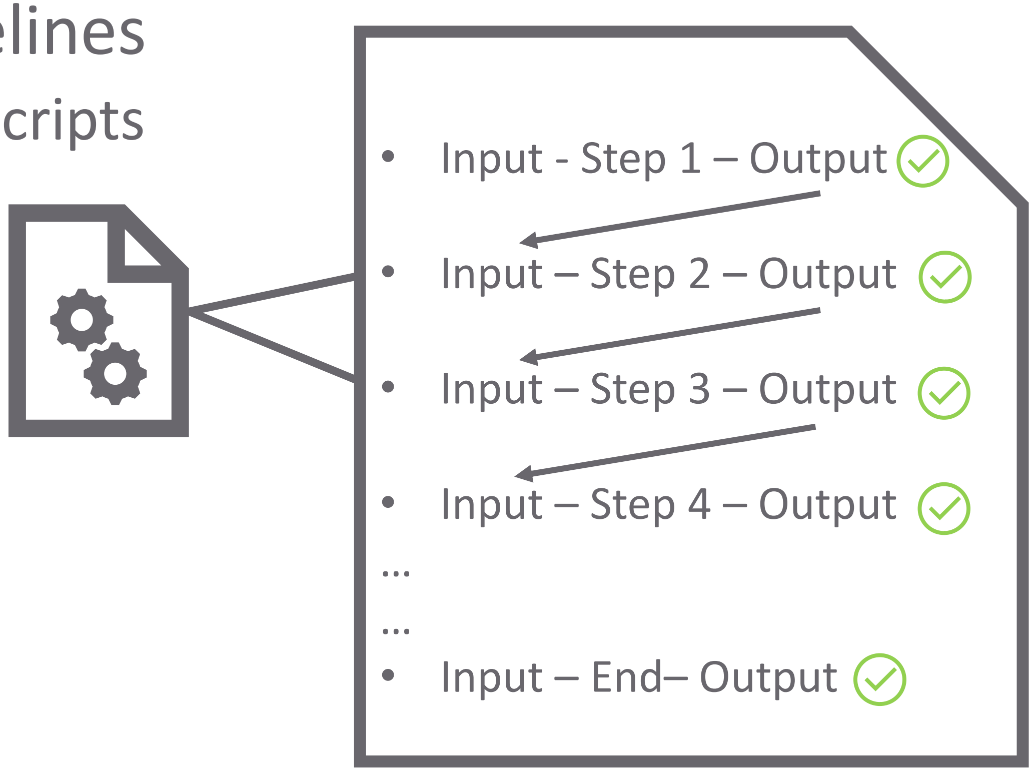
nextflow

Snakemake



REPRODUCTIBILITY

- Pipelines
  - Scripts



- Input - Step 1 – Output ✓
- Input – Step 2 – Output ✓
- Input – Step 3 – Output ✓
- Input – Step 4 – Output ✓

…

…

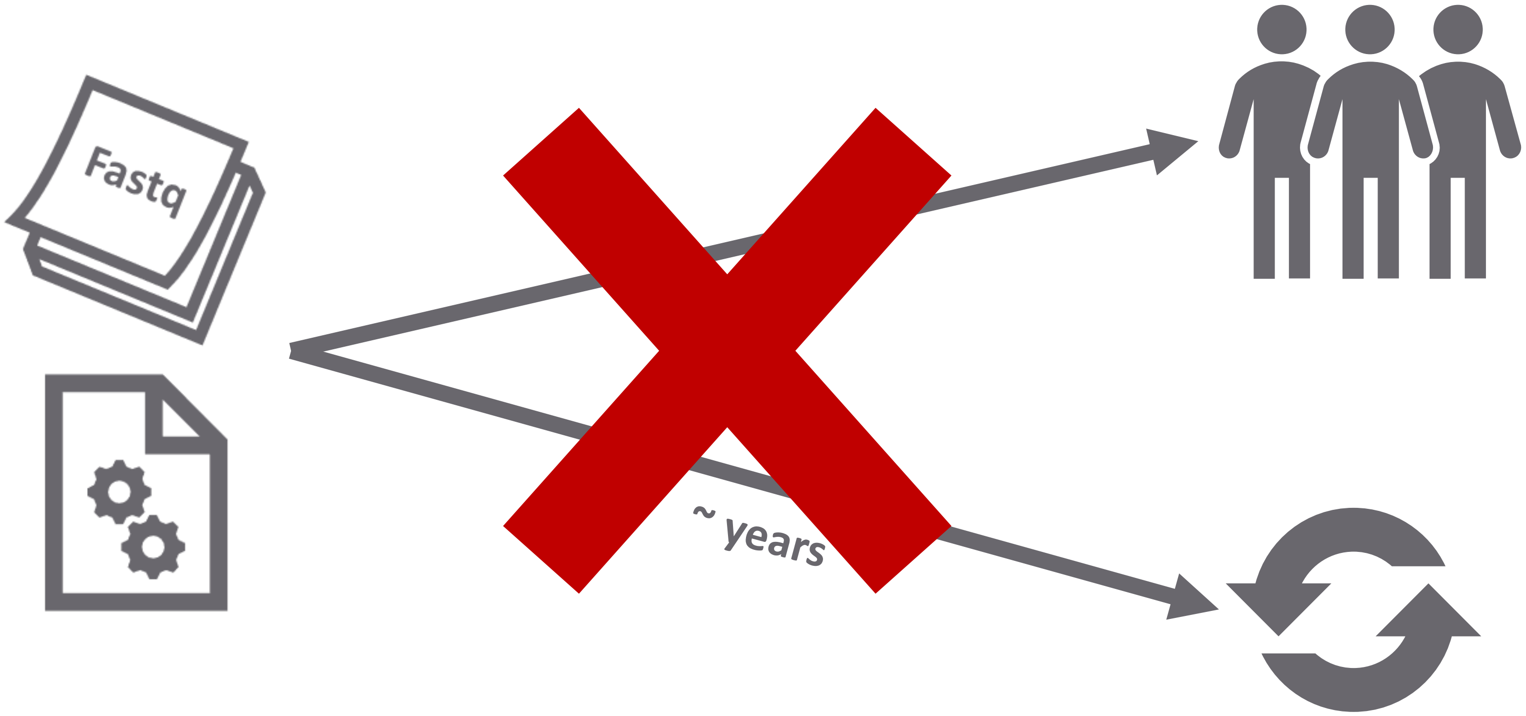
- Input – End– Output ✓

nextflow

Snakemake

nf-core

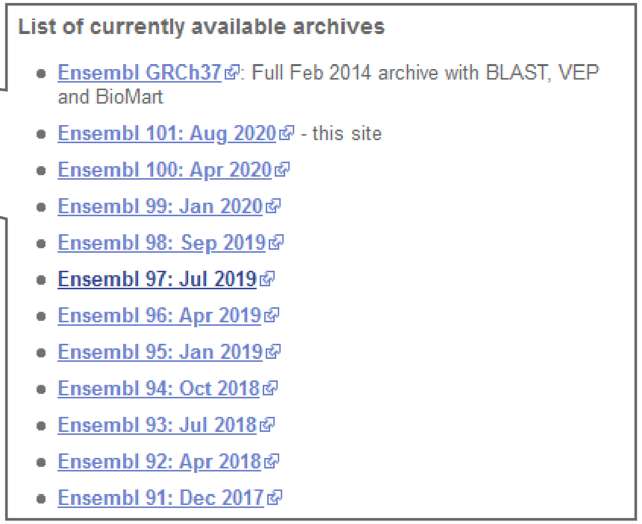
REPRODUCTIBILITY



REPRODUCTIBILITY

- Variability
  - Updates / Versioning
    - Softwares
    - References
  - Compatibility
  - Format
  - Knowledge

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



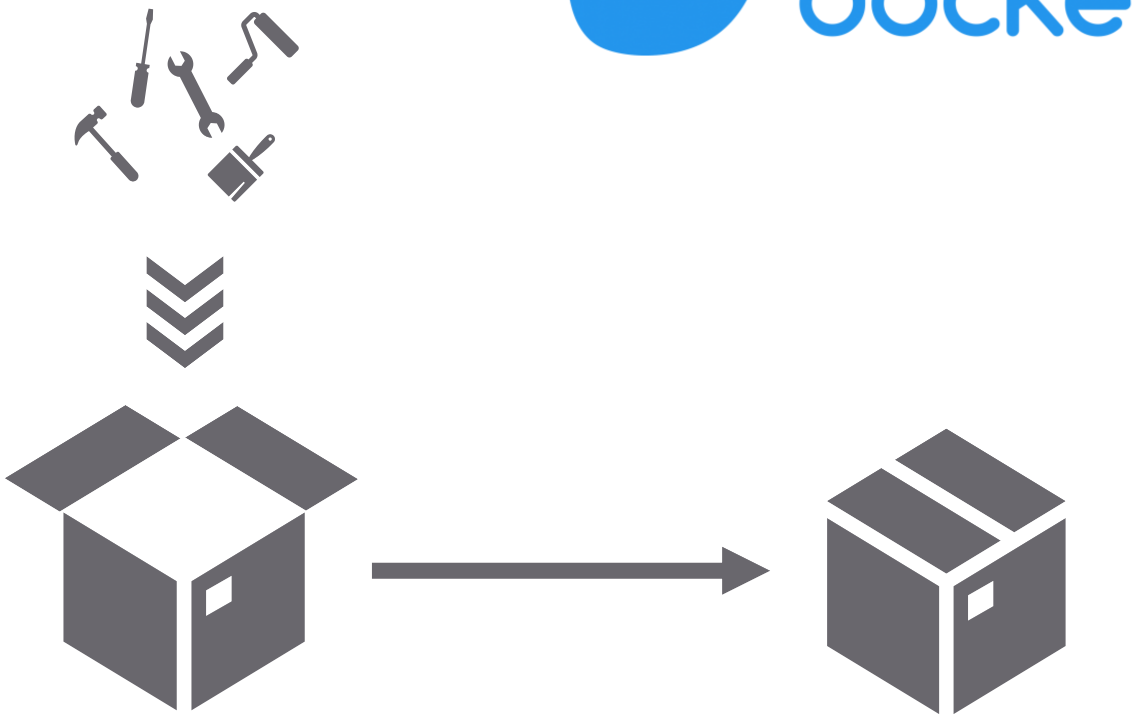
List of currently available archives

- 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

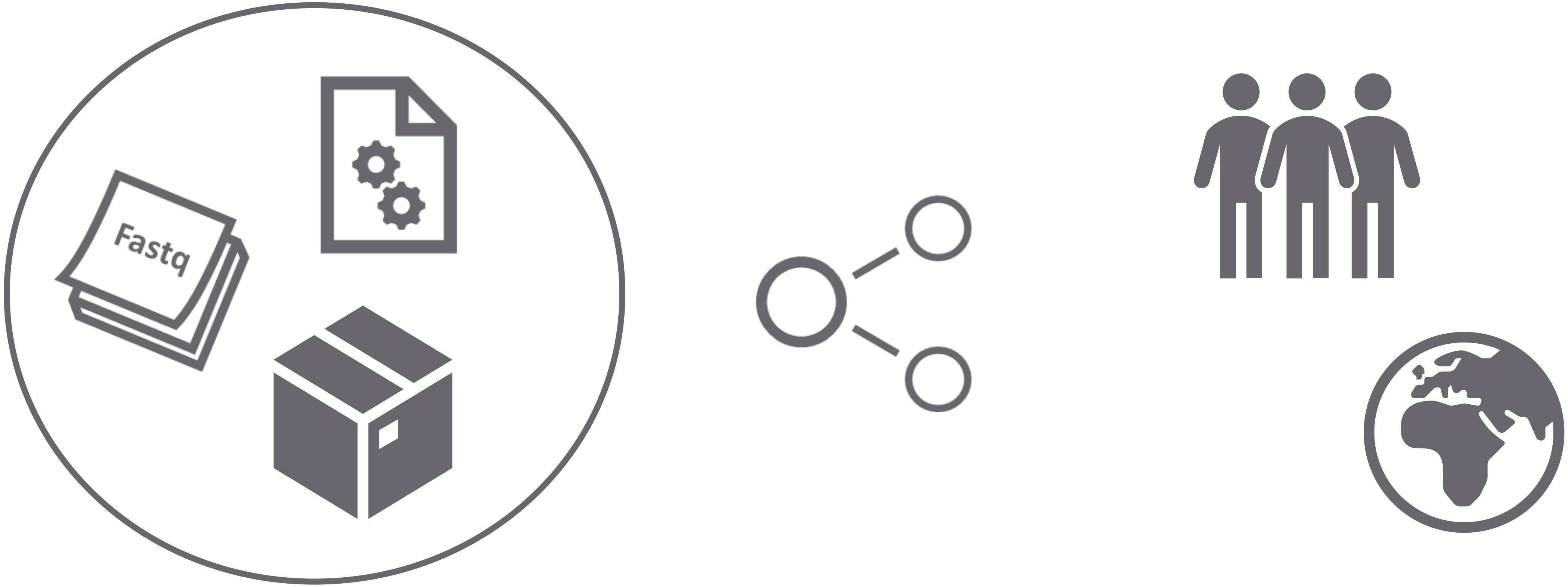
REPRODUCTIBILITY

- CONTAINERS
  - Docker
  - Singularity

- Softwares
  - Versions



REPRODUCTIBILITY



DATA DEPOSITORY

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

THANK YOU FOR YOUR ATTENTION