LIÈGE université
GIGA
Bioinformatics

GIGA Doctoral School for Health Sciences
Computational Biosciences Days

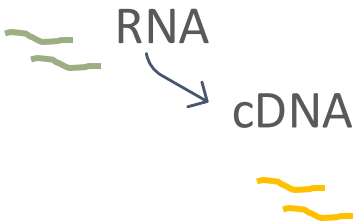
INTRODUCTION TO NEXT-GENERATION SEQUENCING DATA PROCESSING

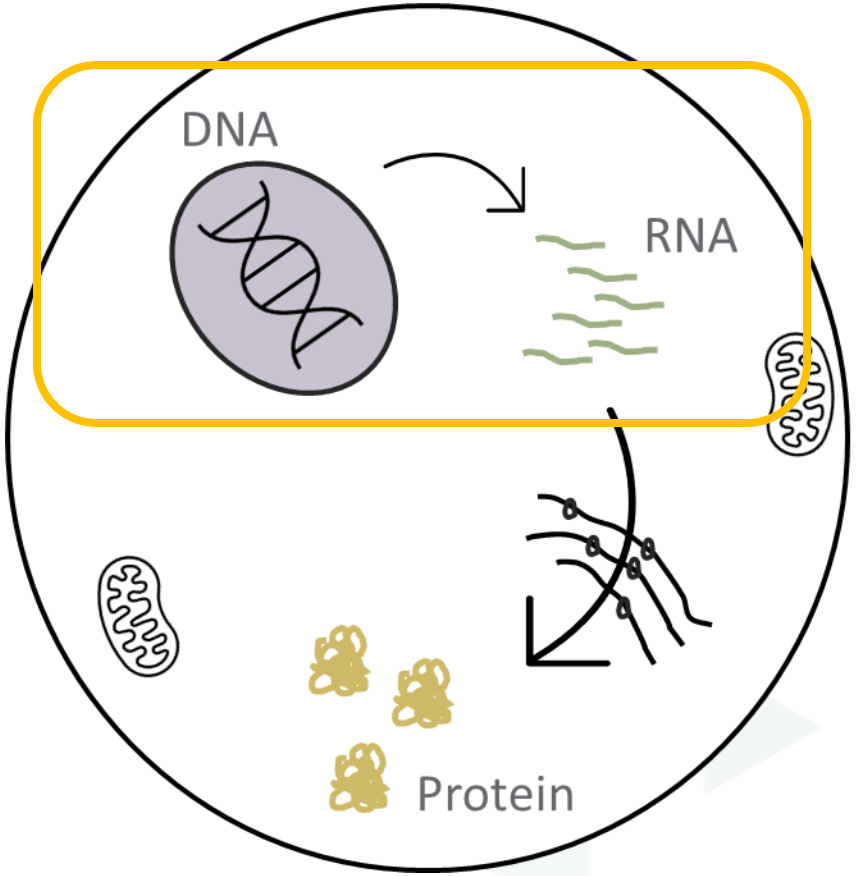
ARNAUD LAVERGNE, PhD

GIGA-Bioinformatics

CLASSIFICATION

« Transcriptomics »

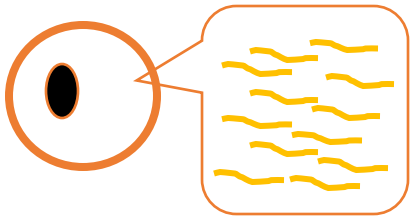




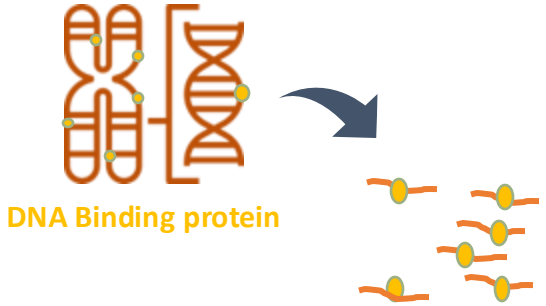
« Genomics »

SAMPLING

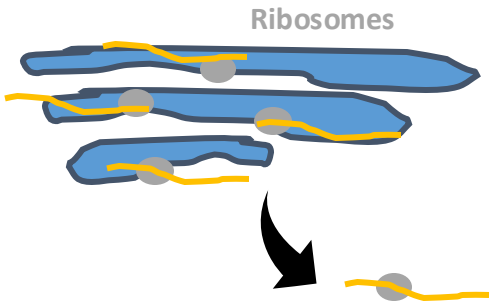
RNA-Seq



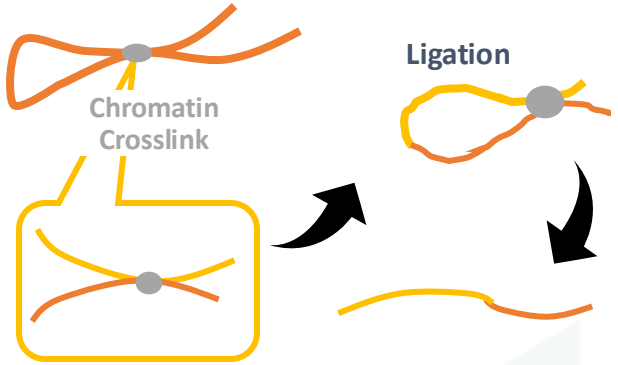
ChIP-Seq



ATAC-Seq



Ribo-Seq



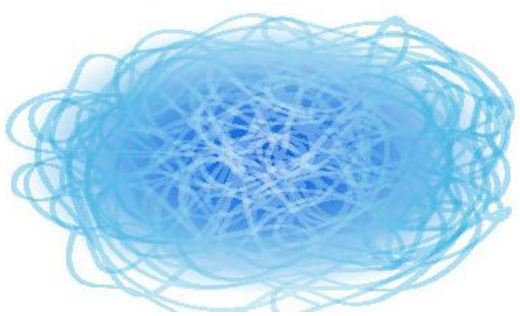
Hi-C

CLIP-Seq

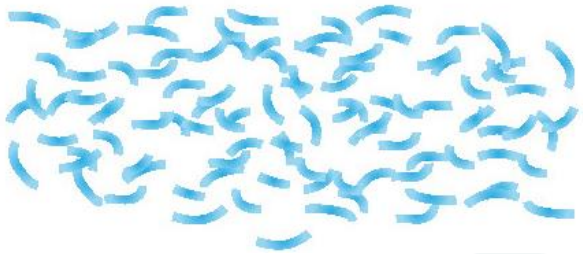
NEXT-GENERATION SEQUENCING

= High-Throughput Sequencing (HTS)

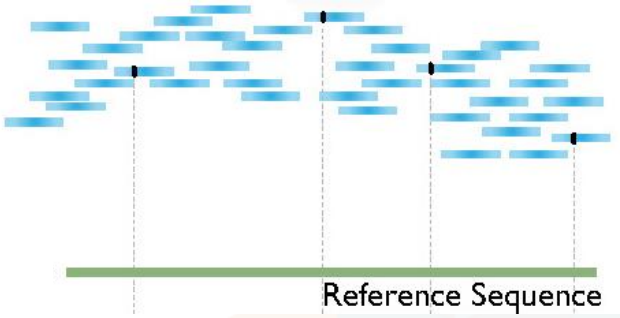
« Genome-wide »



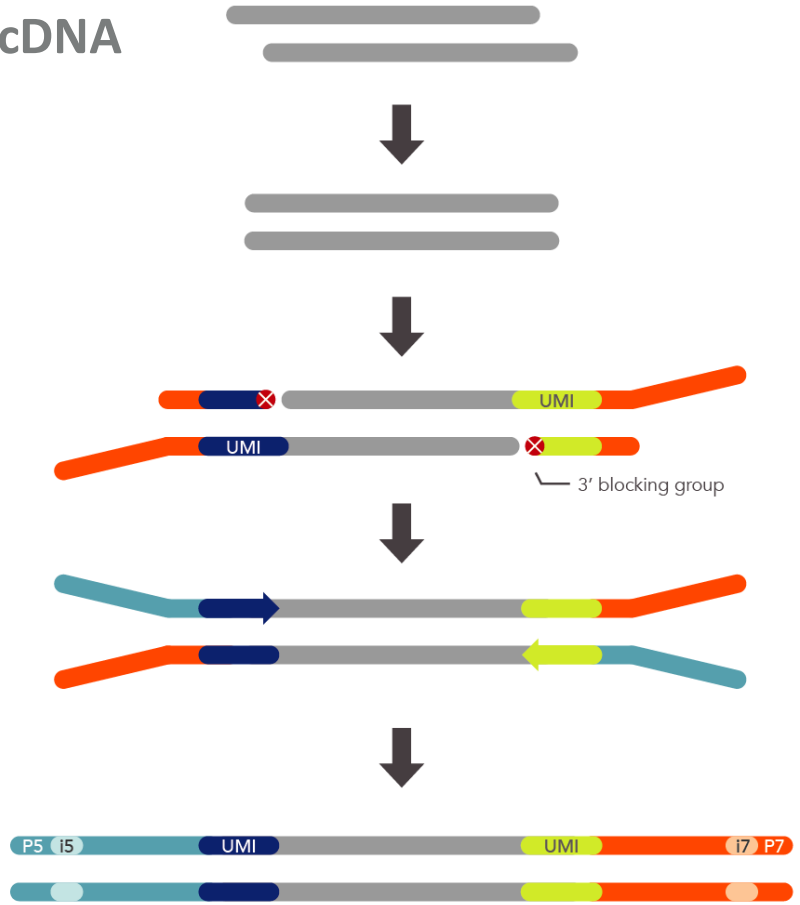
Massive parallel sequencing



References



LIBRARY PREPARATION



- Fragmented input
- End-repair
- Adding adaptors
  - Priming sequences
  - Indexes
  - UMI
  - FlowCell complementary sequences
- Amplification

DATA ACQUISITION

SEQUENCING

SEQUENCERS

illumina®

Oxford NANOPORE Technologies

PacBio

Element Biosciences

SEQUENCERS

Adapted from Illumina

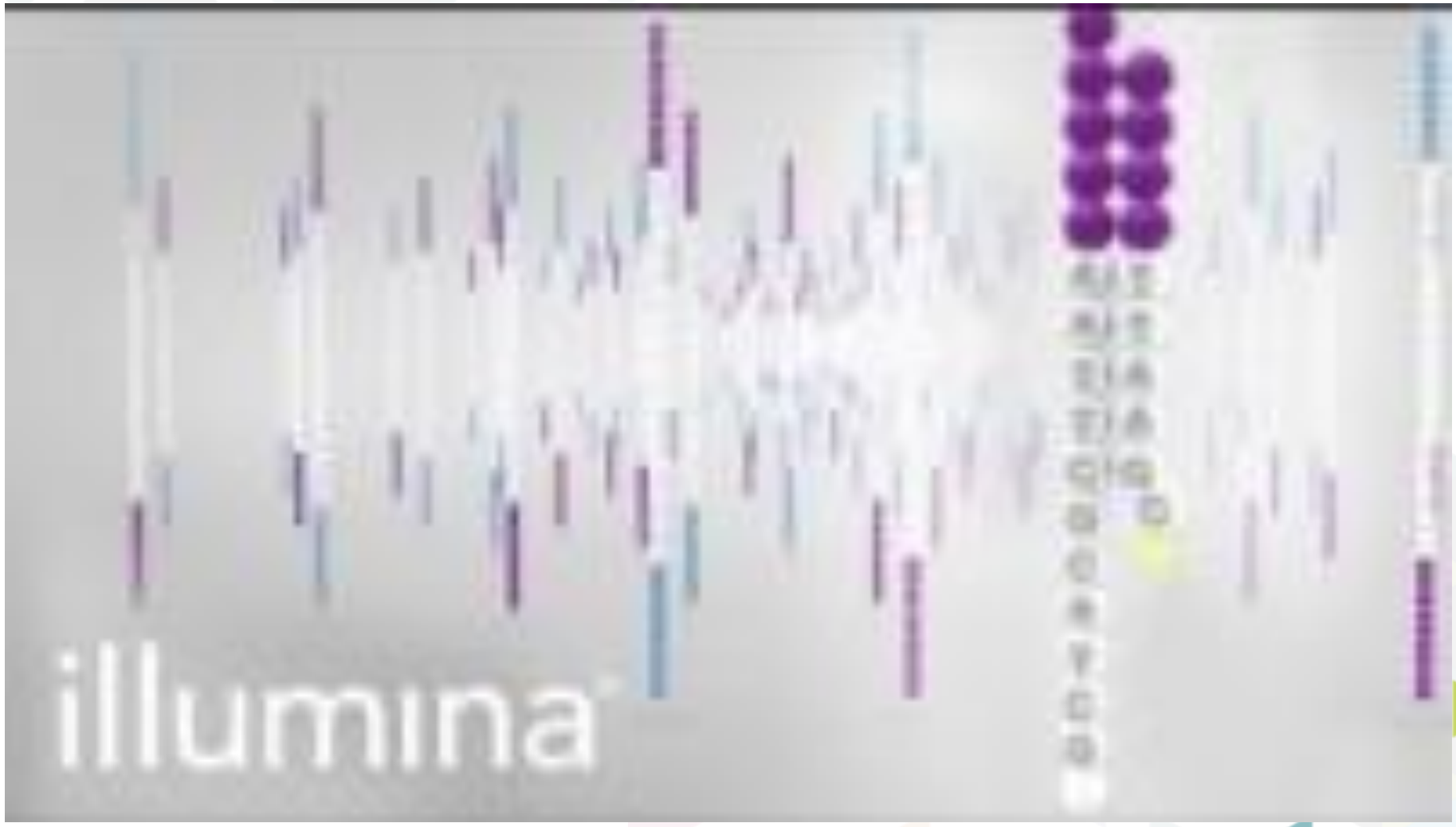
MiSeq
540 Mb -15 Gb
4 – 56 hours

HiSeq
105 Gb - 1,5 Tb
1 – 3,5 days

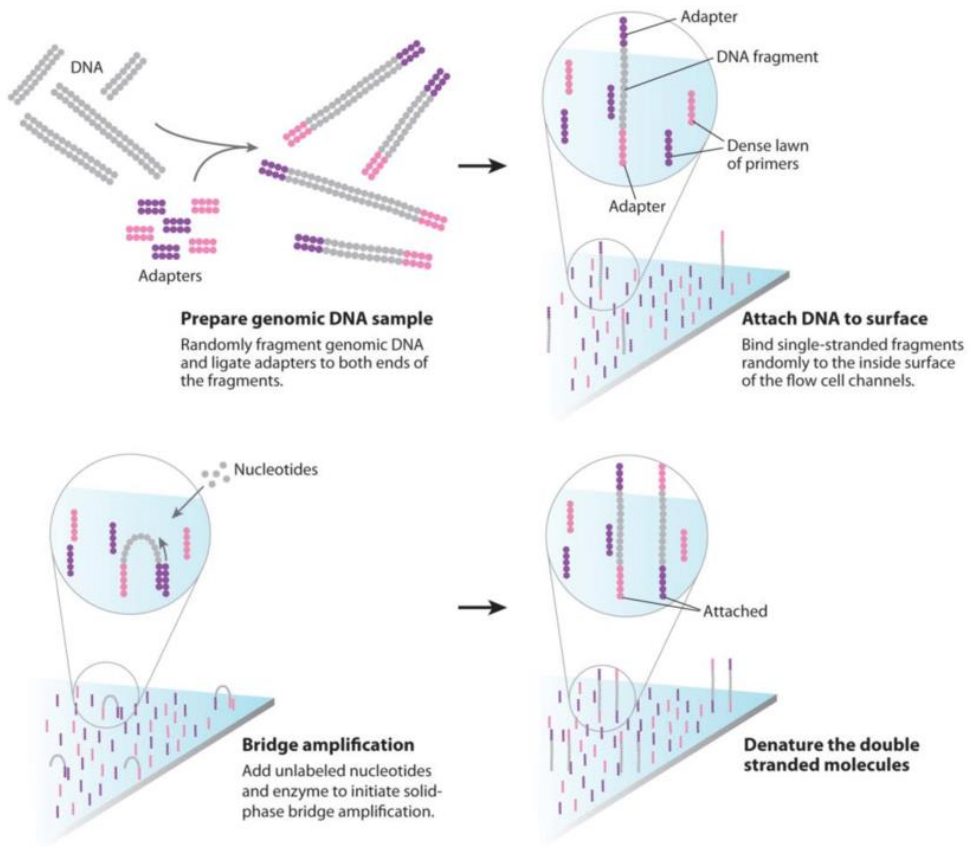
NextSeq
16,25 Gb - 120 Gb
11 – 29 hours

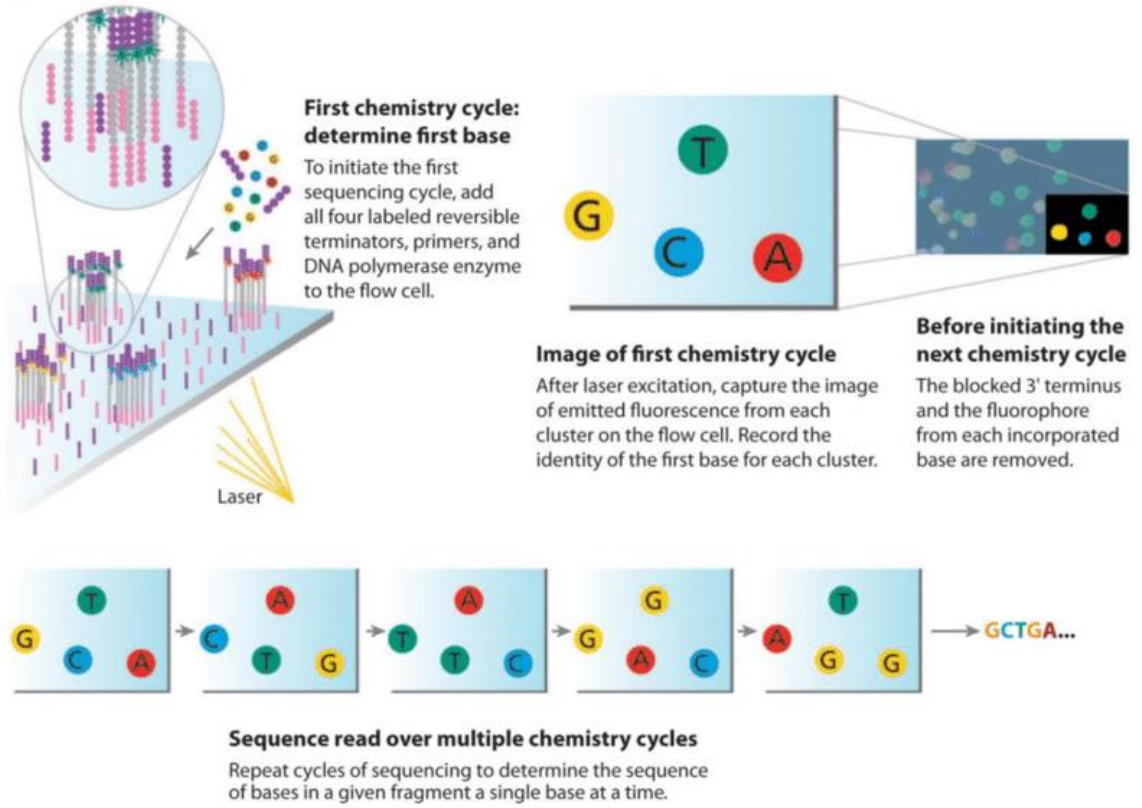
NovaSeq
65 Gb – 3 Tb
13 – 44 hours

ILLUMINA SEQUENCING BY SYNTHESIS



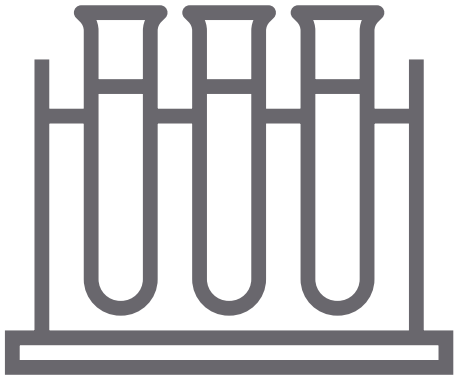
ILLUMINA SEQUENCING BY SYNTHESIS





Adapted from Illumina

SEQUENCING



SEQUENCING

SEQUENCING

DIGITALIZATION OF THE INFORMATION

SEQUENCING

DATA ANALYSIS – RAW DATA

DATA FORMAT & QC SEQUENCING

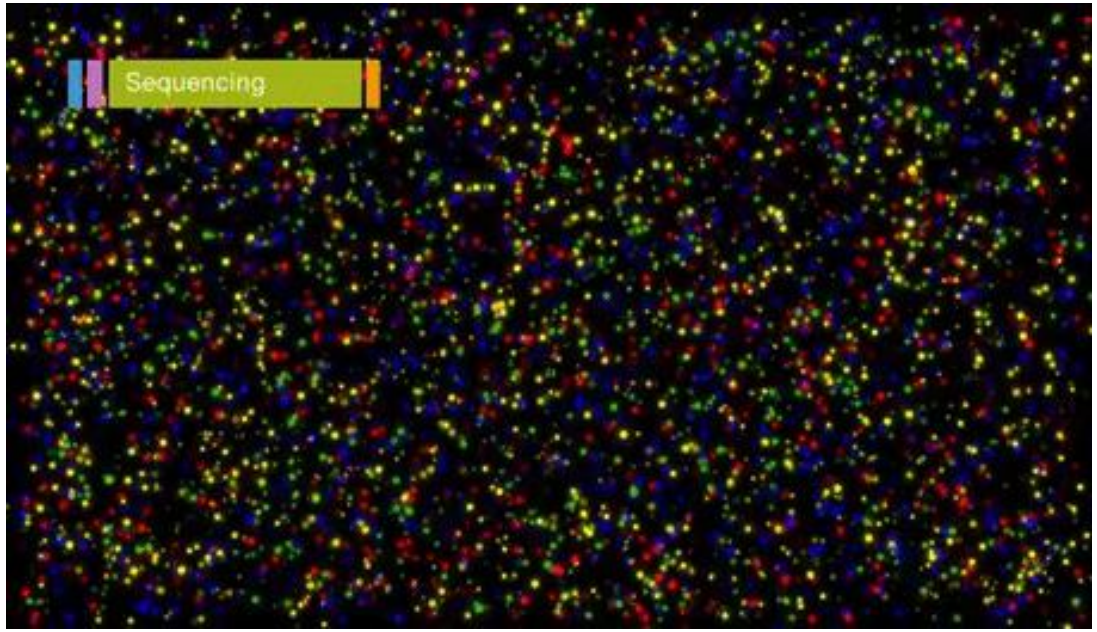
SEQUENCERS

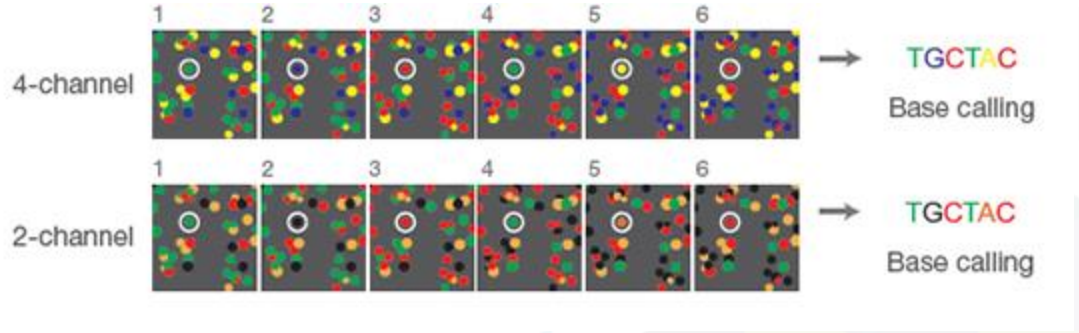
- 1.2 Tb of data
- 12.8 Billions of reads
- Hundreds of samples
  - Multiple experiments
  - Unique combinations of indexes

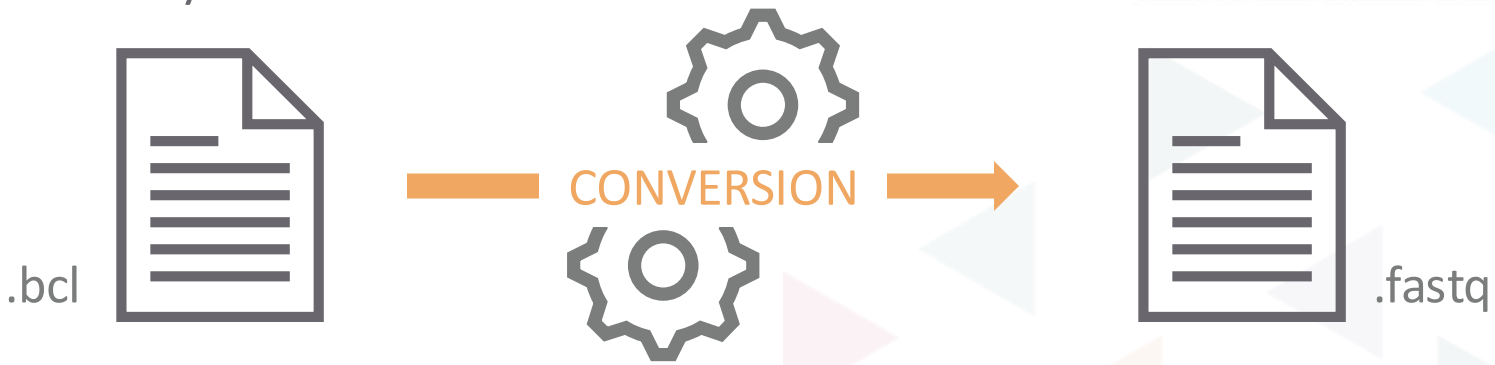
- « Run »
- No storage on device

DATA FORMAT

- Bcl files
  - Each position
  - Each cycle
    - Base calls
    - Base call quality scores
  - Raw data files
    - Binary format







DATA FORMAT

- DEMULTIPLEXING
  - Reads → Samples
    - Indexes
  - Fastq format
    - Identifier
    - Sequence
    - Separator
    - Quality score


```
@A00801:49:H2VGKDSXY:2:1101:1624:1016 1:N:0:NGCTTAAG+TCGTGACC
CTTCTGGAGAGGAGTTCCTCTGATATGAATTAAGGTTTTCCCTTCTGTCATGACCAGAAGAAGGTTTTATCTGTGCCACACTACTTTTCATTTTCTGTTGCCAGTTGGTCCAATAAATCAAA
GATGNTTCAAACTGGTCCAAATAACAAGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFF#FFFFFFFFFFFFFFFFFFFFFFF
FF
@A00801:49:H2VGKDSXY:2:1101:2022:1016 1:N:0:NGCTTAAG+TCGTGACC
TGACAAAAAGATACCTCATTTATGGGGAAATTGAGGAAGATACATATACAAGCACCCCAACCCATATTTAACATATTTGGCAATAACTCCCTTCCCATTCTTCCCCCTCCAATTTTCAAATA
GTAGNTTTTTAAAAAATTAAAGACATGTC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFF,FFFFFF:FFFFFFF,#FFFFFFF,FFFFFFFF:F:FF,FF
@A00801:49:H2VGKDSXY:2:1101:4137:1016 1:N:0:NGCTTAAG+TCGTGACC
TTTTTTTGCCTTTTCAAGTGTTTATTTTTATACATTTTTTTGTATTAAAAAGAAAAGCATAATTACCACAAATTACAAAGGACTAAAGCAGGACTAGAATAATGAATGAATCACTTCAGCCT
GGAANGCAGATACTCTCAATAATATTAAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFF:FFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFF:FFF,FFFFFFFFFFF:FFF#FFFFF:FFFFFFFFFFFFFFFFF
```

DATA FORMAT

- DEMULTIPLEXING
  - Reads → Samples
    - Indexes
  - Fastq format
    - Identifier
    - Sequence
    - Separator
    - Quality score
      - Phred +33 encoded using ASCII

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%

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			

OVERVIEW

Run

Demultiplexing

Fastq

DATA TYPE

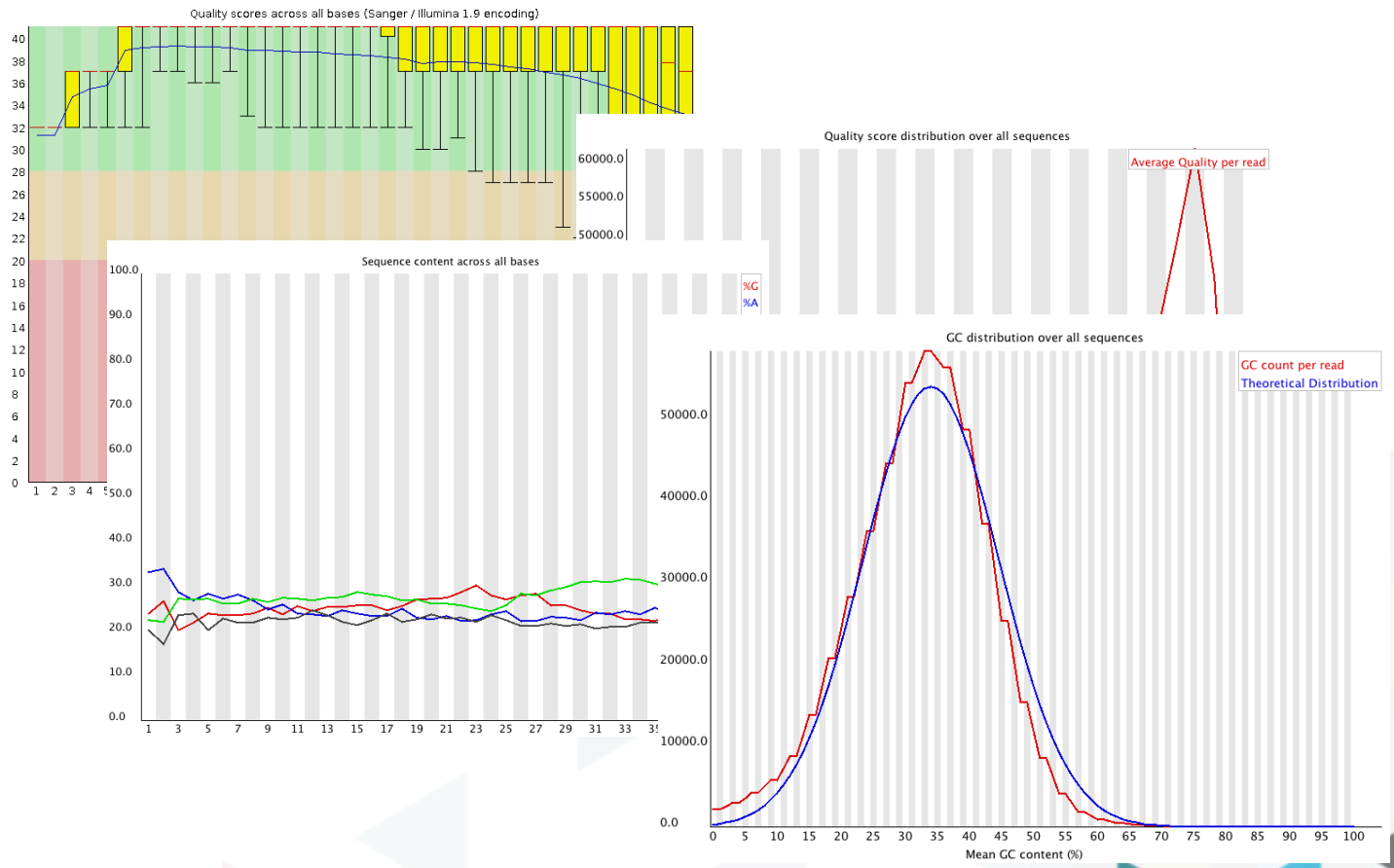
EXPERIMENTAL DESIGN

Fastq.gz

QC SEQUENCING

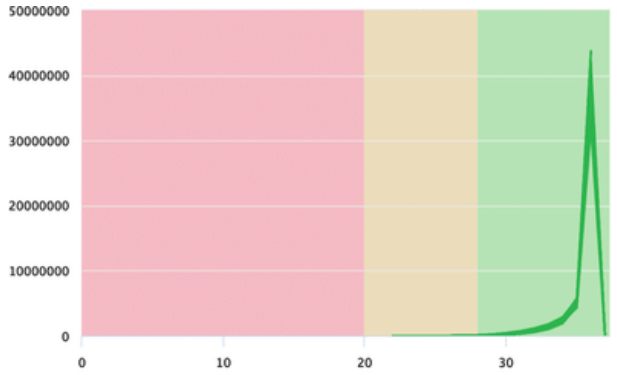
FastQC

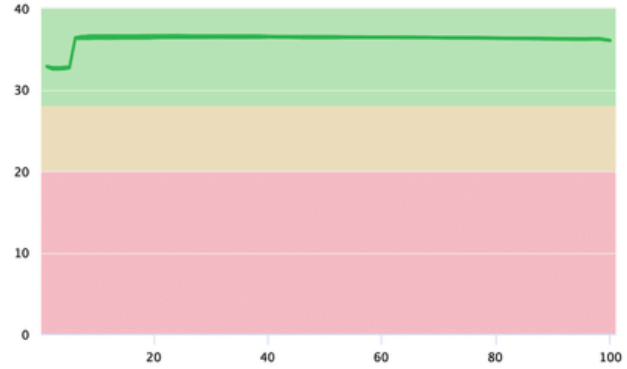
- Number of reads
- Base calling quality
- Sequence quality
- GC content
- Sequence length
- Duplication levels
- Adapter content
- Overrepresented sequences
- …

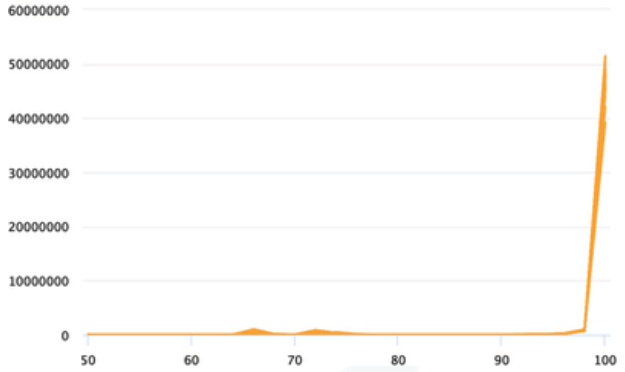


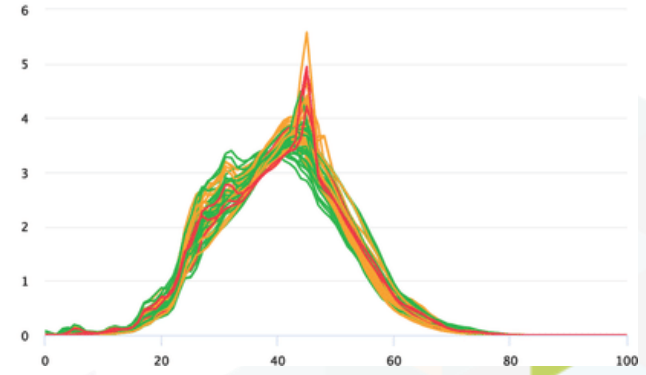
QC SEQUENCING

MultiQC





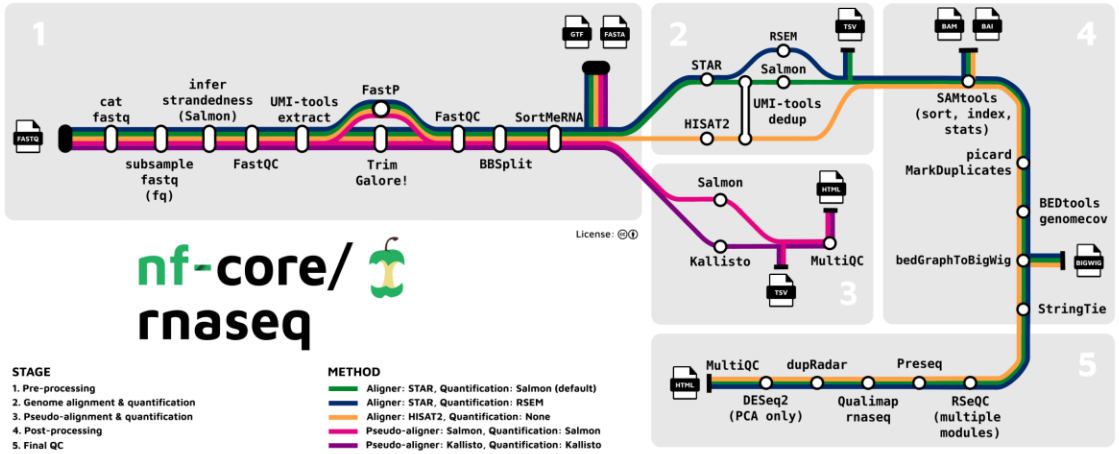


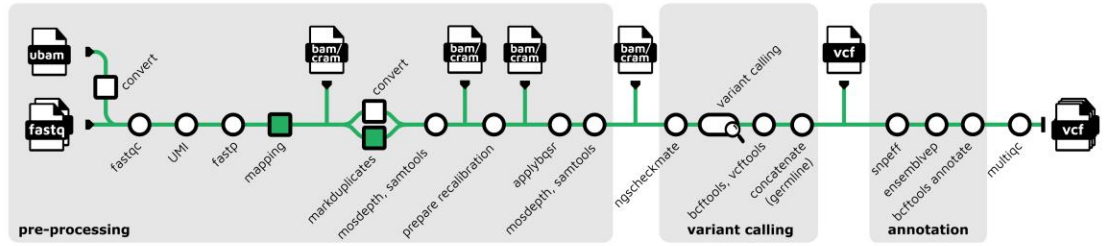


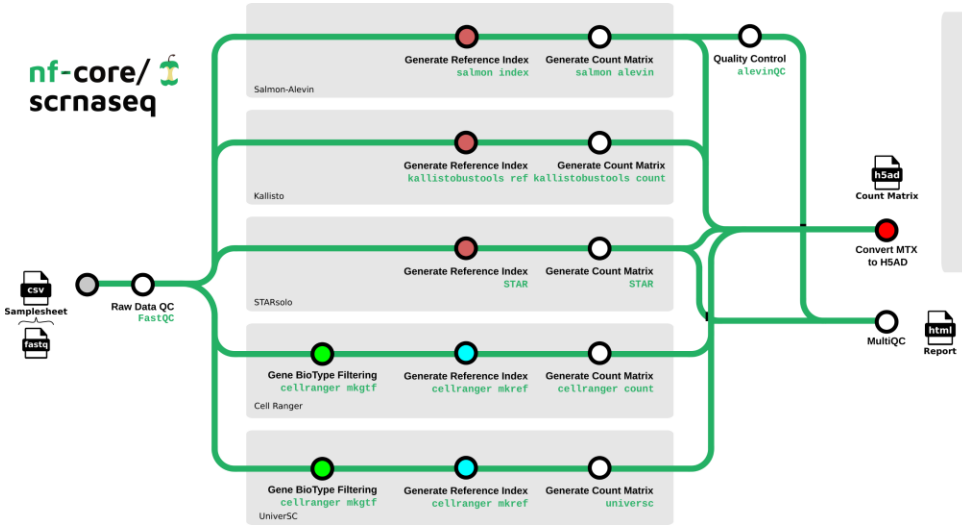
DATA ANALYSIS - PROCESSING

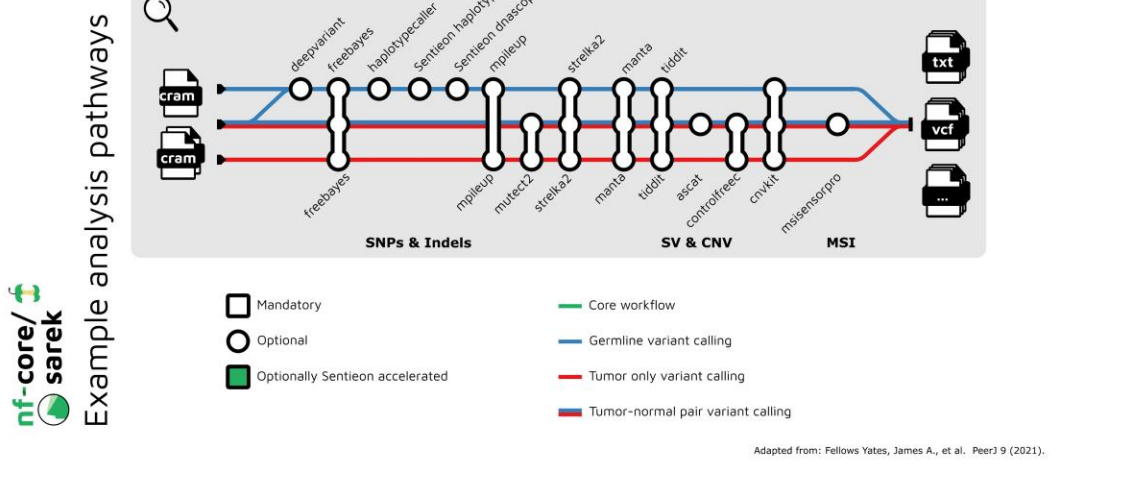
MAPPING, BAM FILES & QC MAPPING

PIPELINES


nf-core/
rnaseq
STAGE
1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC
METHOD
Aligner: STAR, Quantification: Salmon (default)
Aligner: STAR, Quantification: RSEM
Aligner: HISAT2, Quantification: None
Pseudo-aligner: Salmon, Quantification: Salmon
Pseudo-aligner: Kallisto, Quantification: Kallisto
cat fastq
infer strandedness (Salmon)
UMI-tools extract
FastP
FastQC
SortMeRNA
subsample fastq (fq)
FastQC
Trim Galore!
BBSplit
STAR
RSEM
Salmon
HISAT2
UMI-tools dedup
Salmon
Kallisto
MultiQC
SAMtools (sort, index, stats)
picard MarkDuplicates
BEDtools genomecov
bedGraphToBigWig
StringTie
MultiQC
dupRadar
Preseq
DESeq2 (PCA only)
Qualimap rnaseq
RSeQC (multiple modules)


ubam
fastq
convert
fastqc
UMI
fastp
mapping
markduplicates
convert
mosdepth, samtools
prepare recalibration
applybqsr
mosdepth, samtools
ngscheckmate
variant calling
bcftools, vcftools
concatenate (germline)
snpeff
ensemblvep
bcftools annotate
multiqc
vcf
pre-processing
variant calling
annotation

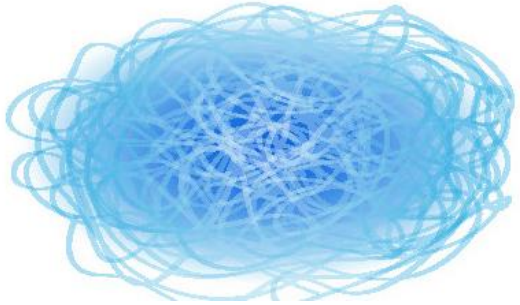

nf-core/
scrnaseq
Generate Reference Index
salmon index
Generate Count Matrix
salmon alevin
Salmon-Alevin
Quality Control
alevinQC
Generate Reference Index
kallistobustools ref
Generate Count Matrix
kallistobustools count
Kallisto
Generate Reference Index
STAR
Generate Count Matrix
STAR
STARsolo
Gene BioType Filtering
cellranger mkgtf
Generate Reference Index
cellranger mkref
Generate Count Matrix
cellranger count
Cell Ranger
Gene BioType Filtering
cellranger mkgtf
Generate Reference Index
cellranger mkref
Generate Count Matrix
universc
UniverSC
Samplesheet
Raw Data QC
FastQC
Count Matrix
Convert MTX to H5AD
MultiQC
Report
Legend
Start
End
Input:
Reference Annotation
Genome
Reference Annotation
Genome


nf-core/
sarek
Example analysis pathways
deepvariant
freebayes
haplotypecaller
Sentieon haplotyper
Sentieon dnascope
mpileup
strelka2
manta
tiddit
cram
cram
freebayes
mpileup
mutect2
strelka2
manta
tiddit
ascat
controlfreec
cnvkit
msisensorpro
txt
vcf
SNPs & Indels
SV & CNV
MSI
Mandatory
Optional
Optionally Sentieon accelerated
Core workflow
Germline variant calling
Tumor only variant calling
Tumor-normal pair variant calling
Adapted from: Fellows Yates, James A., et al. PeerJ 9 (2021).

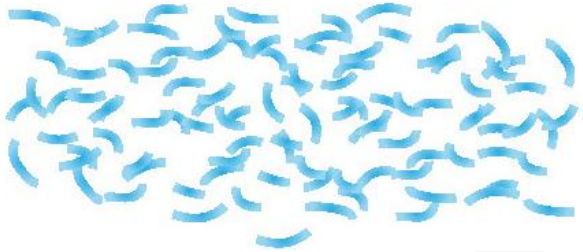
NEXT-GENERATION SEQUENCING

= High-Throughput Sequencing (HTS)

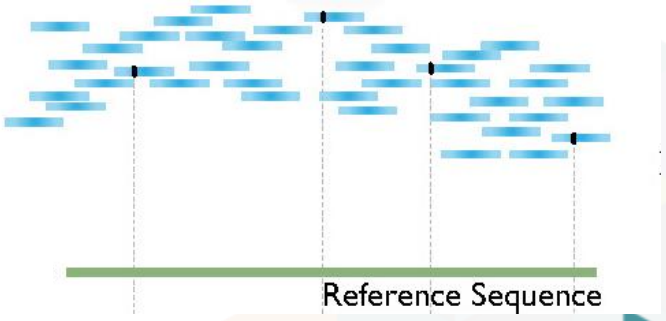
« Genome-wide »



Massive parallel sequencing



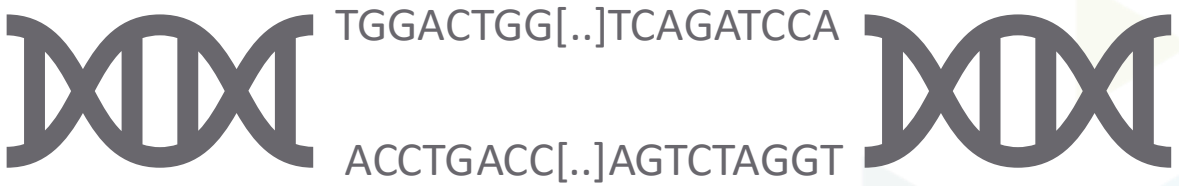
References



MAPPING

Reference

- Genome sequence
- Gene set



Alignment / mapping

REFERENCE

Genome (FASTA)

>1 dna:chromosome chromosome:GRCh38:1:1:248956422:1 REF

CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAAC
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";

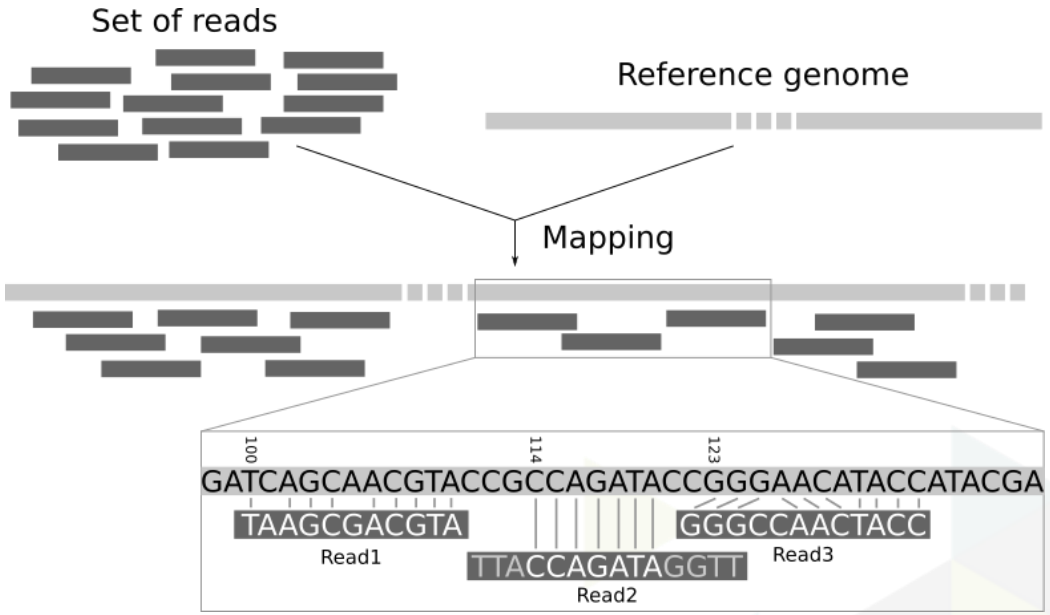
INDEXING

- Genome Indexing
  - Quick queries
  - « 20M reads »
- High RAM/CPUs

- H.Sapiens ~30 Gb

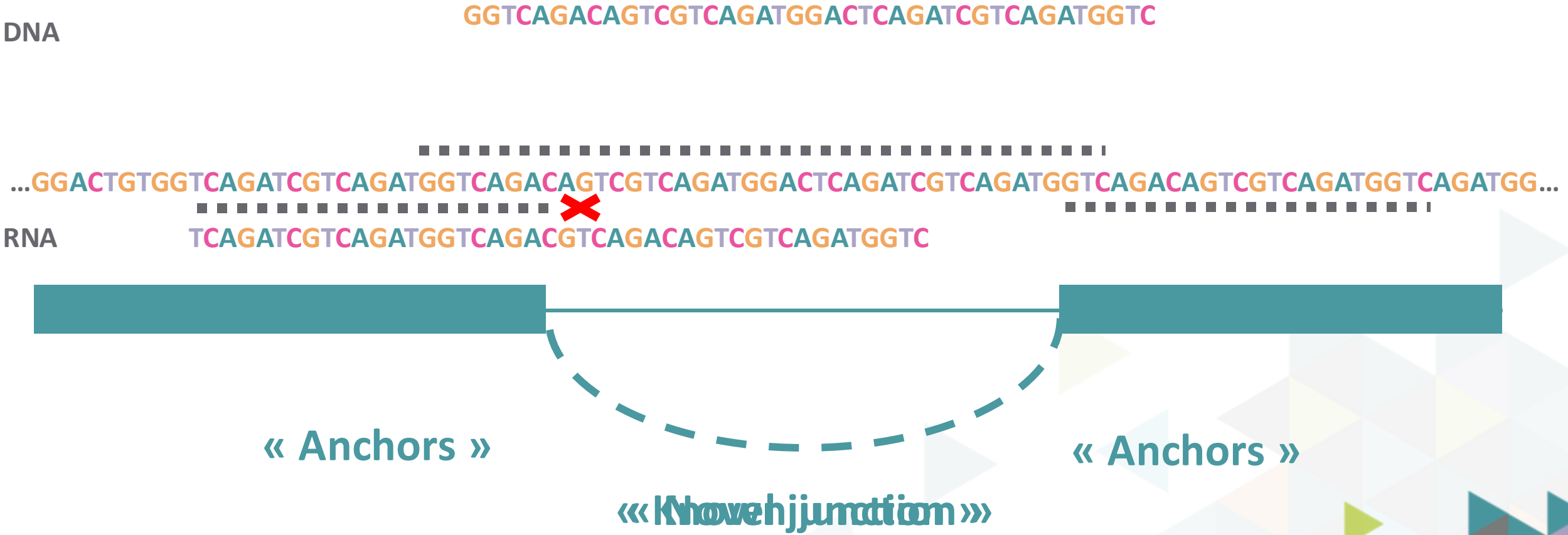
MAPPING

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



- High RAM/CPUs

MAPPING


DNA
GGTCAGACAGTCGTCAGATGGACTCAGATCGTCAGATGGTC
...GGACTGTGGTCAGATCGTCAGATGGTCAGACAGTCGTCAGATGGACTCAGATCGTCAGATGGTCAGACAGTCGTCAGATGGTCAGATGG...
RNA
TCAGATCGTCAGATGGTCAGACGTCAGACAGTCGTCAGATGGTC
« Anchors »
« Anchors »

BAM FILES

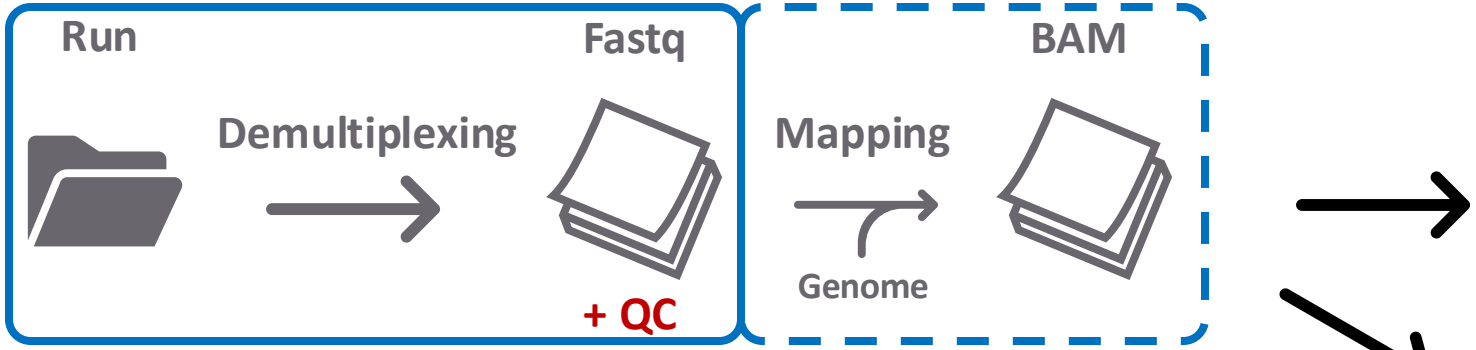
- SAM/BAM files
  - FLAG - Information
  - RNAME - Chromosome
  - POS – Location of 1st base
  - MAPQ – Quality score
  - CIGAR - Operations

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:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF: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



DOWNSTREAM ANALYSES

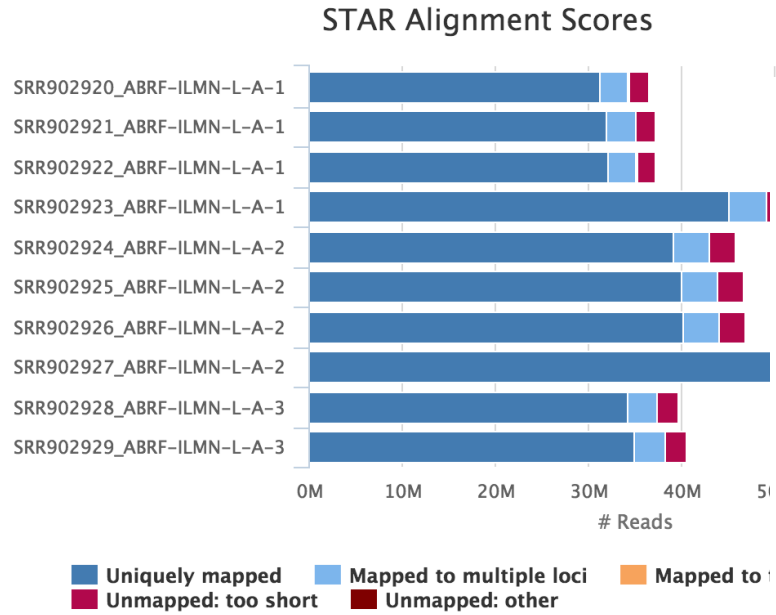
QC MAPPING

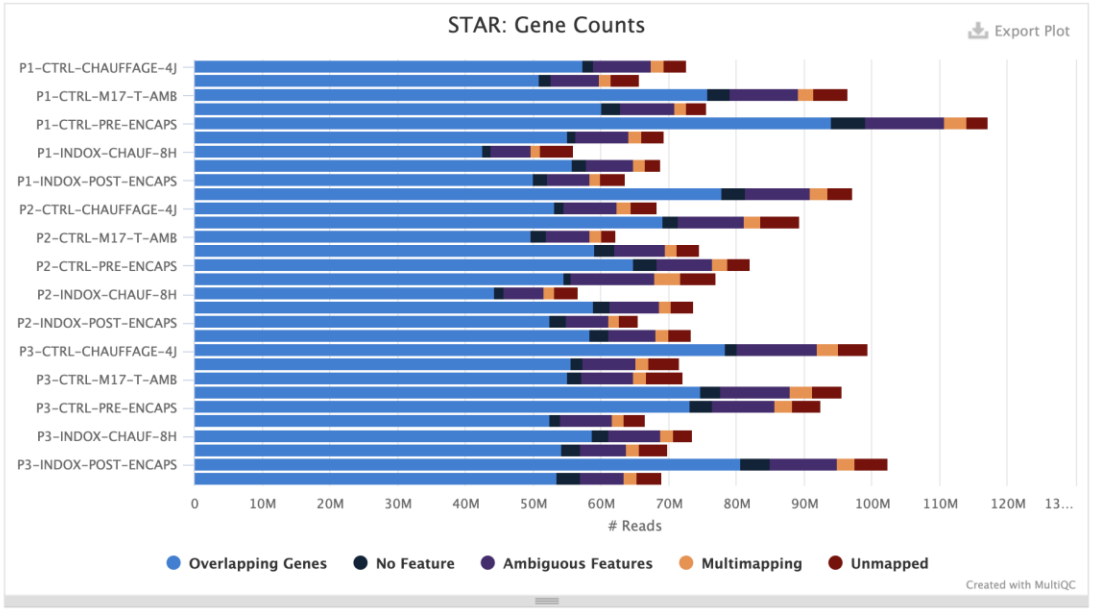
MultiQC

General Statistics

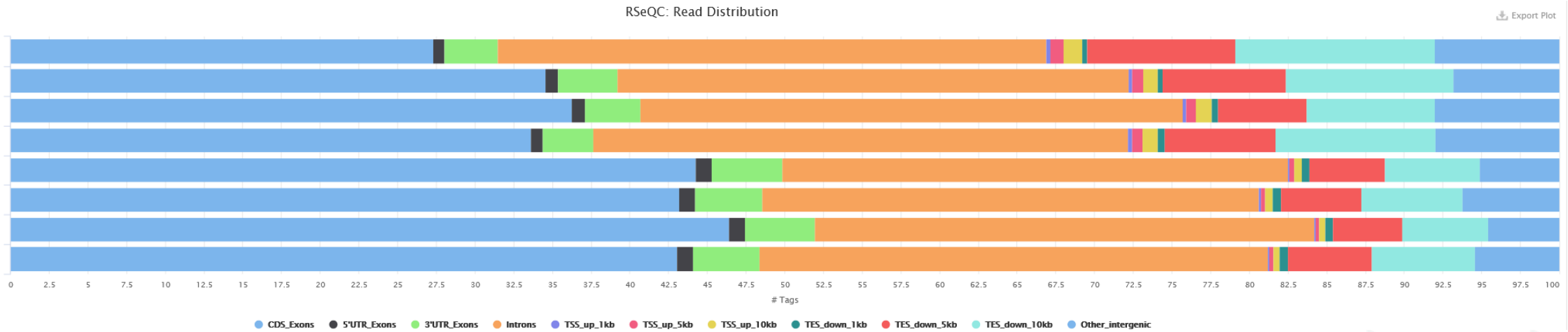
Copy table | Configure Columns | Plot | Showing 8/8 rows and 8/10 columns.

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





QC MAPPING



Mapping distribution

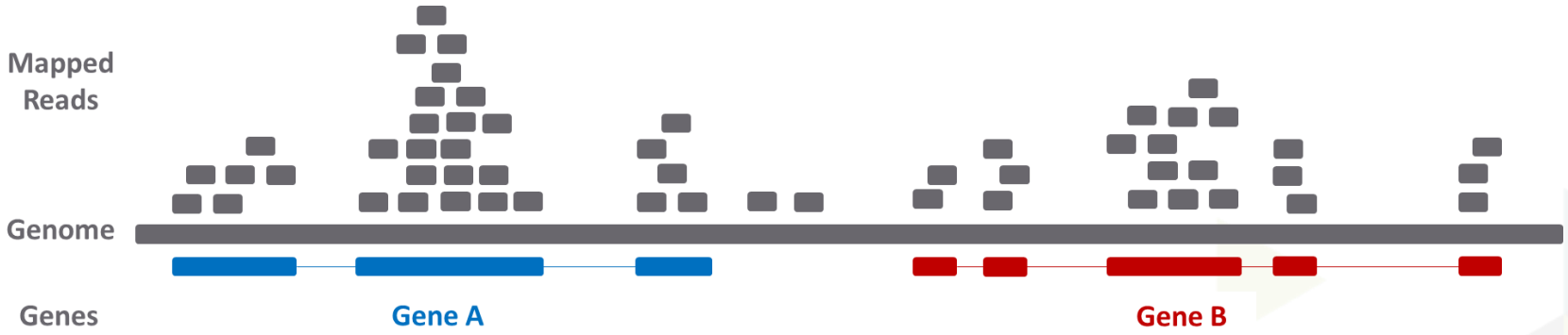
- Proportion of exonic, intronic and intergenic locations
- DNA ≠ total RNA ≠ mRNA
- Experiments

DATA ANALYSIS - RESEARCH

DOWNSTREAM ANALYSIS

RNA - QUANTIFICATION

- Gene Expression

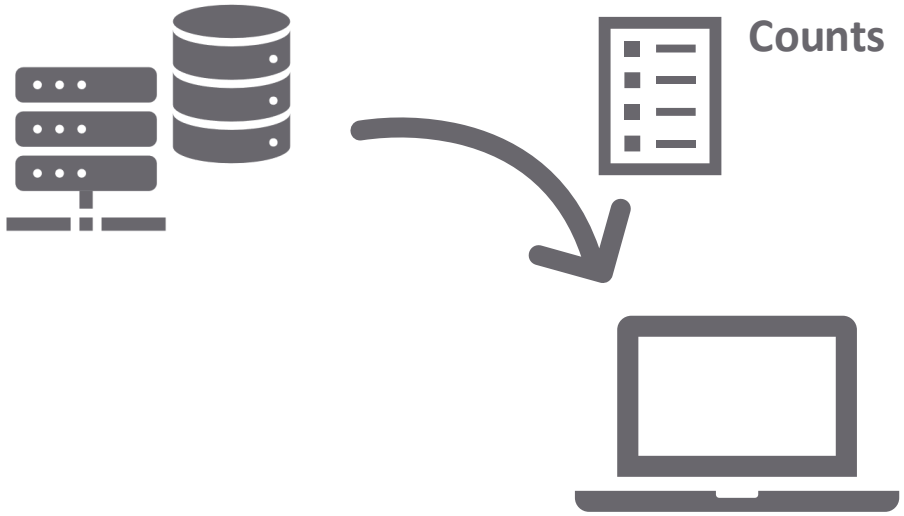


RNA - QUANTIFICATION

- Genes
  - Transcripts
    - Exons

RNA - QUANTIFICATION

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

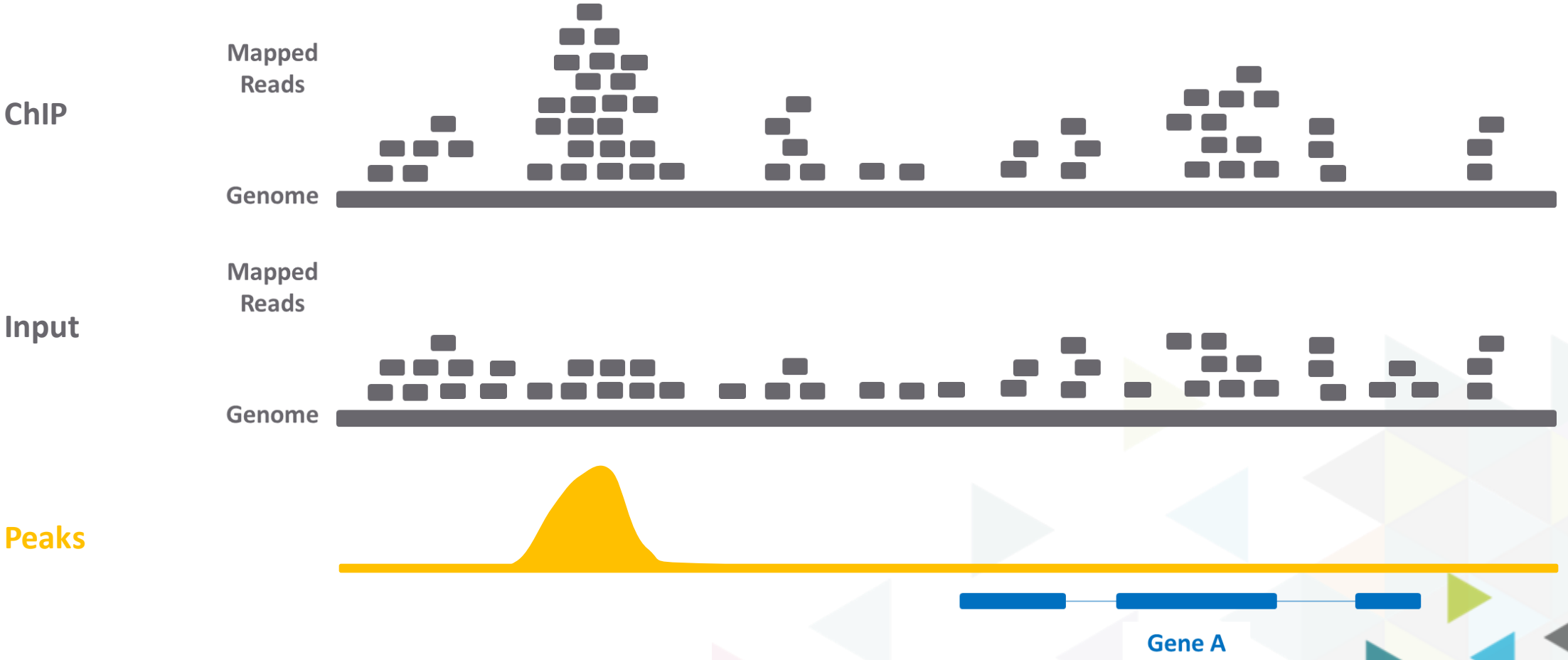


Each column is a sample

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



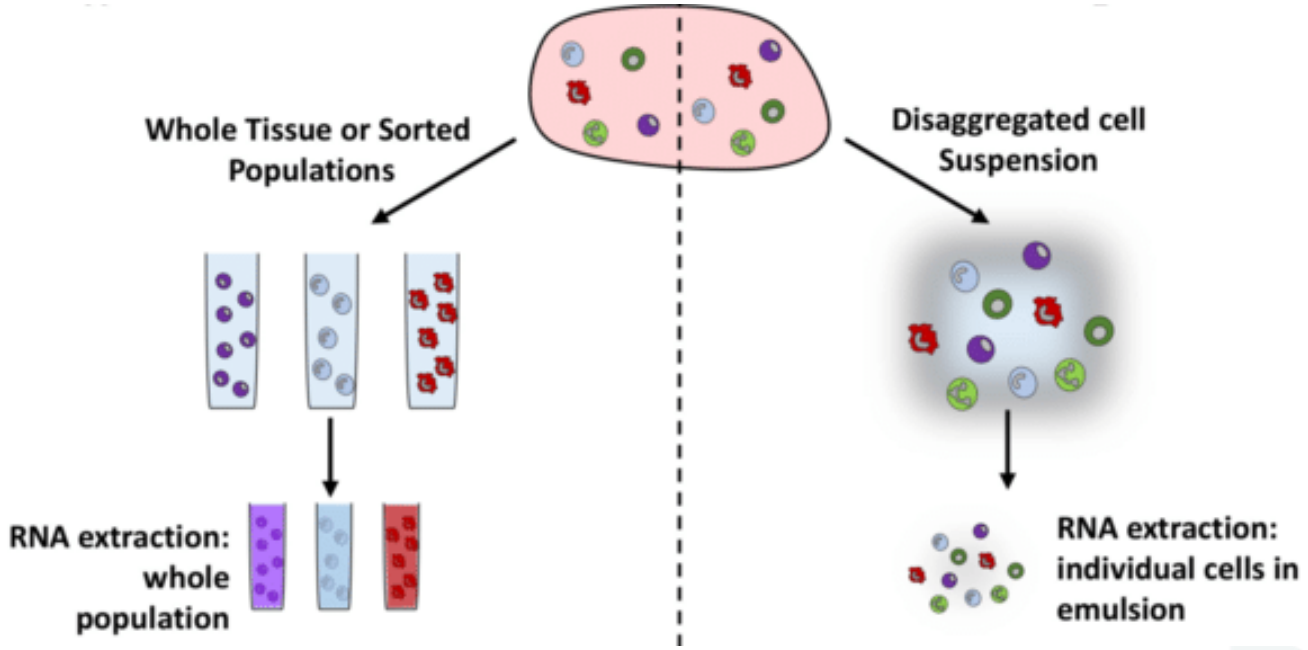
DNA - SNP

REFERENCE …TGGACTGGA..ATCGGCTCGAAGCTTGCATCA..GATCCA…

DATA …TGGAATGGA..ATCGGCTCTAAGCTTGCATCA..GATTCA…

→ **VARIANT / GENOTYPE / GWAS**

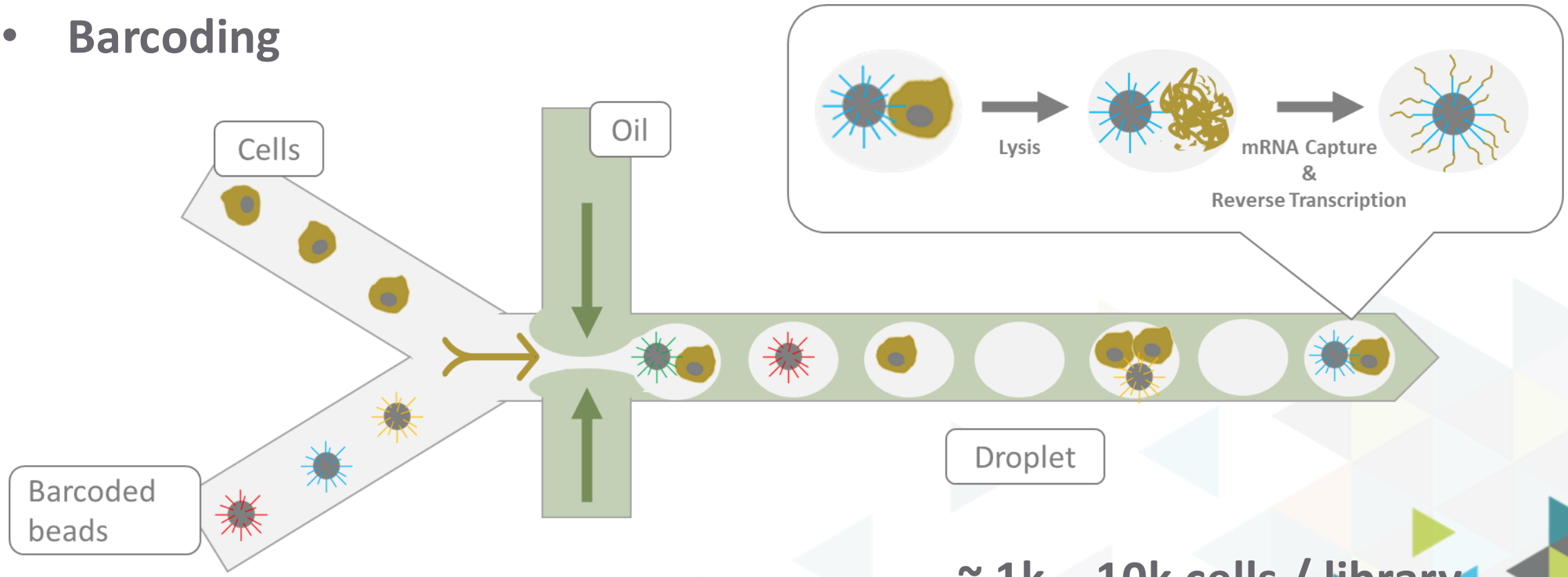
SINGLE-CELL RNASEQ



- Single cell
- No selection
- Heterogeneity
- « Dynamic »

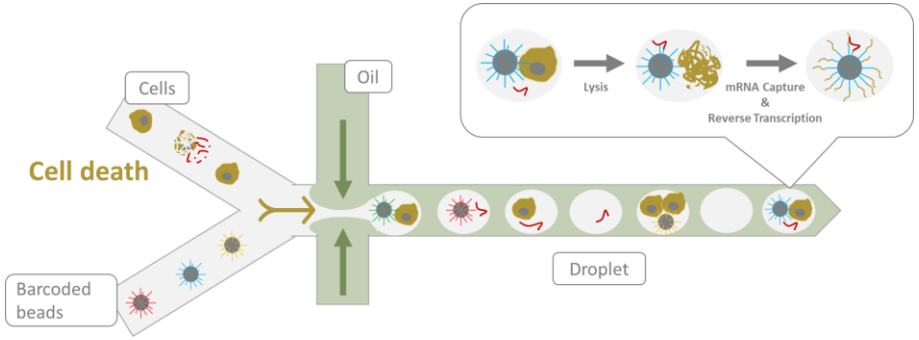
SINGLE-CELL – DROPLET-BASED

- **Encapsulation**
- **Barcoding**



~ 1k – 10k cells / library

CELL DEFINITION



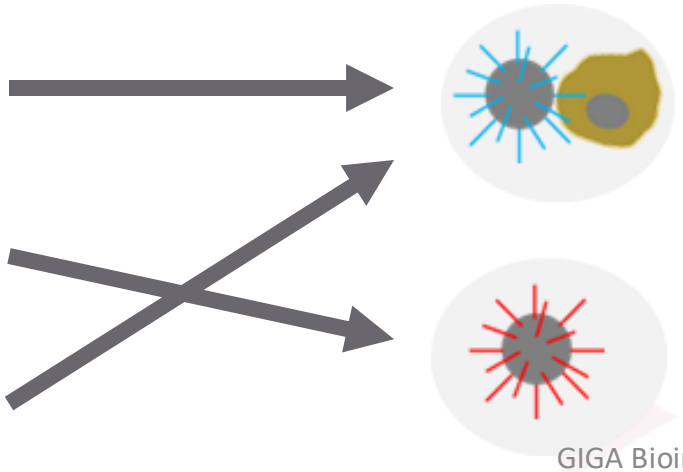
5,247
Estimated Number of Cells

28,918
Mean Reads per Cell

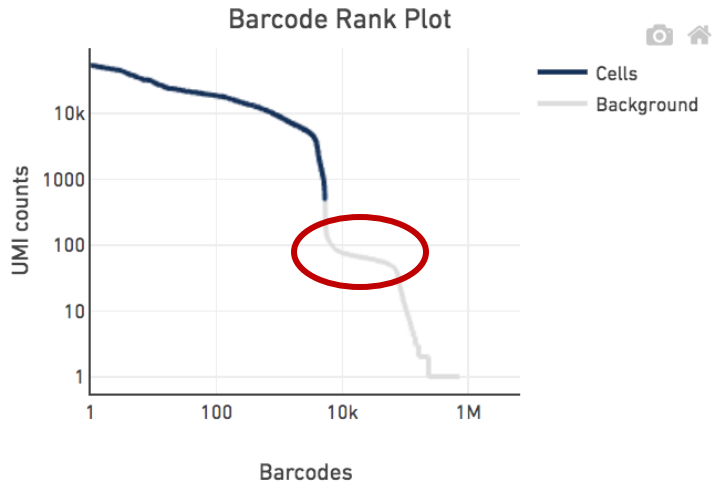
(AAACCCAAGGAGAGTA)
Barcode

(GGTCCCACTGAGAACT)
Barcode

(ATTCCGAAGGTCTGTA)
Barcode

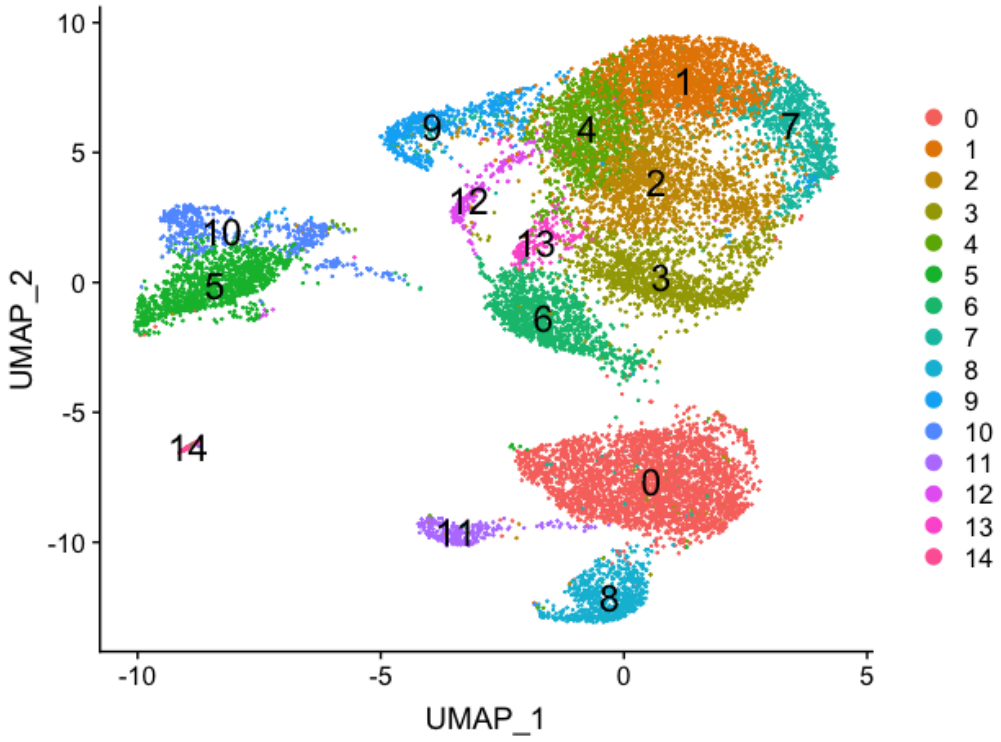


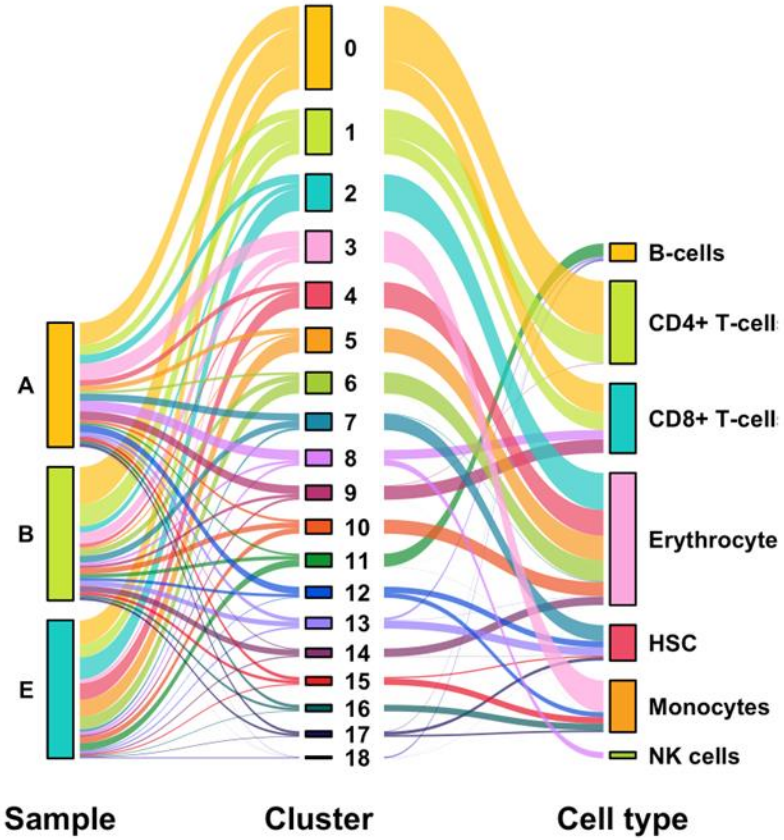
Cells



Estimated Number of Cells	5,247
Fraction Reads in Cells	87.7%
Mean Reads per Cell	28,918
Median Genes per Cell	1,644
Total Genes Detected	20,822
Median UMI Counts per Cell	5,496

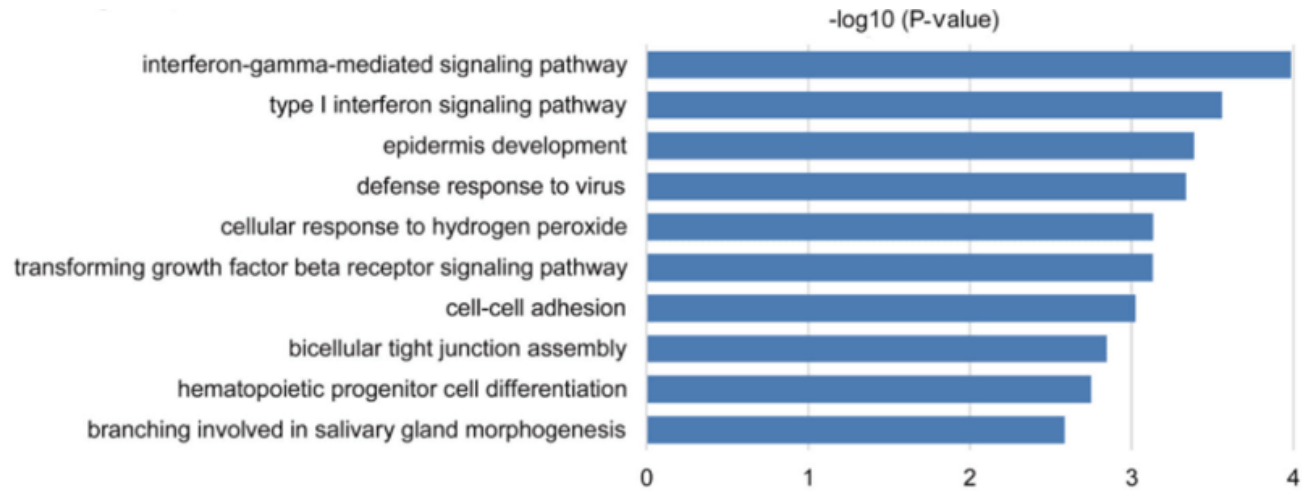
CLUSTERING & ANNOTATIONS





DOWNSTREAM ANALYSIS

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



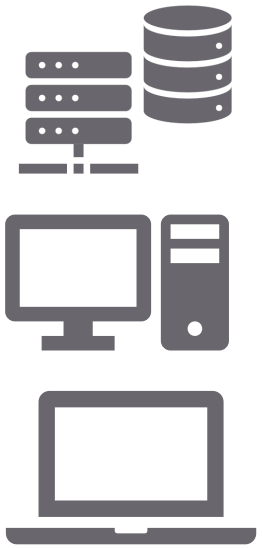
GSEA
Gene Set Enrichment Analysis

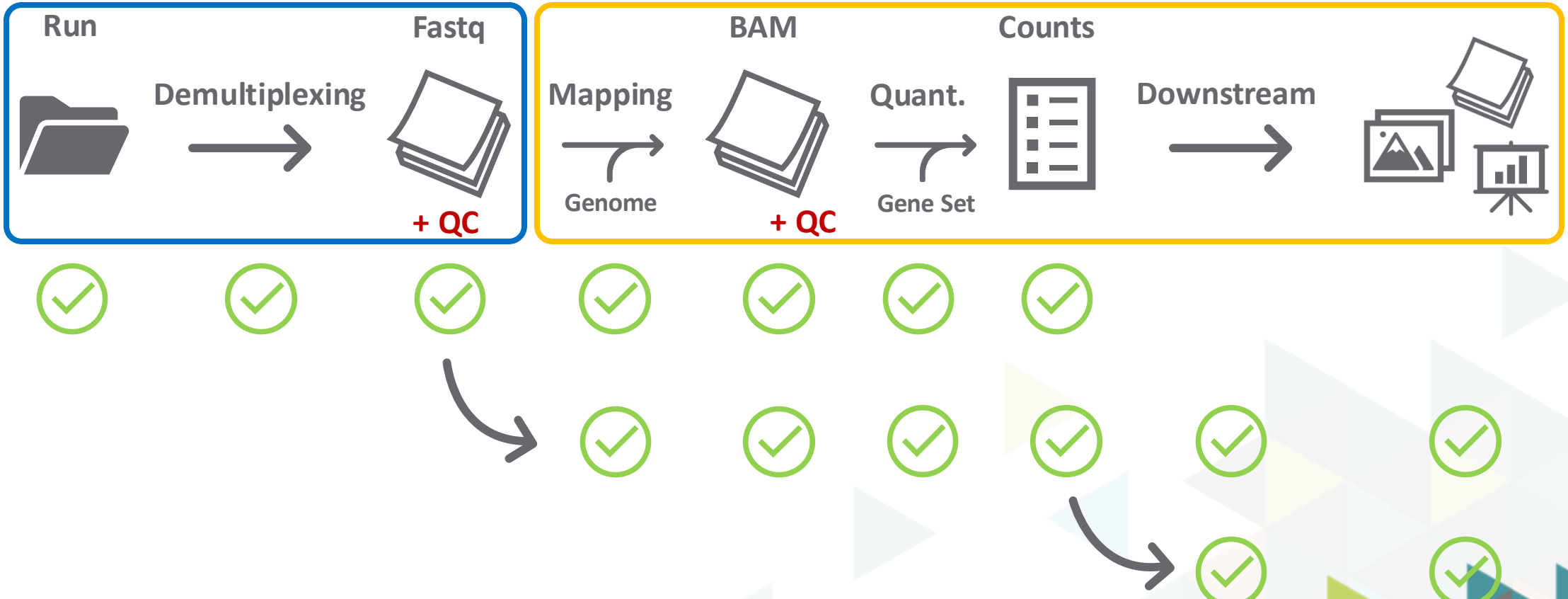
Enrichr

GORILLA
Gene Ontology enRIchment anaLysis and visuaLizAtion tool

PANTHER
Classification System

SUMMARY




Run
Fastq
Demultiplexing
+ QC
BAM
Counts
Mapping
Genome
+ QC
Quant.
Gene Set
Downstream

REPRODUCTIBILITY

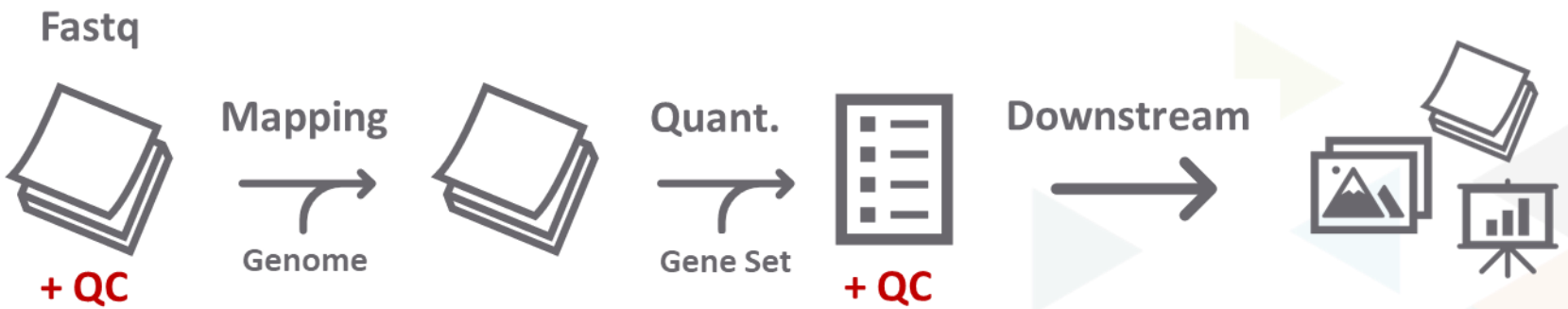
PIPELINES & CONTAINERS

REPRODUCTIBILITY

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

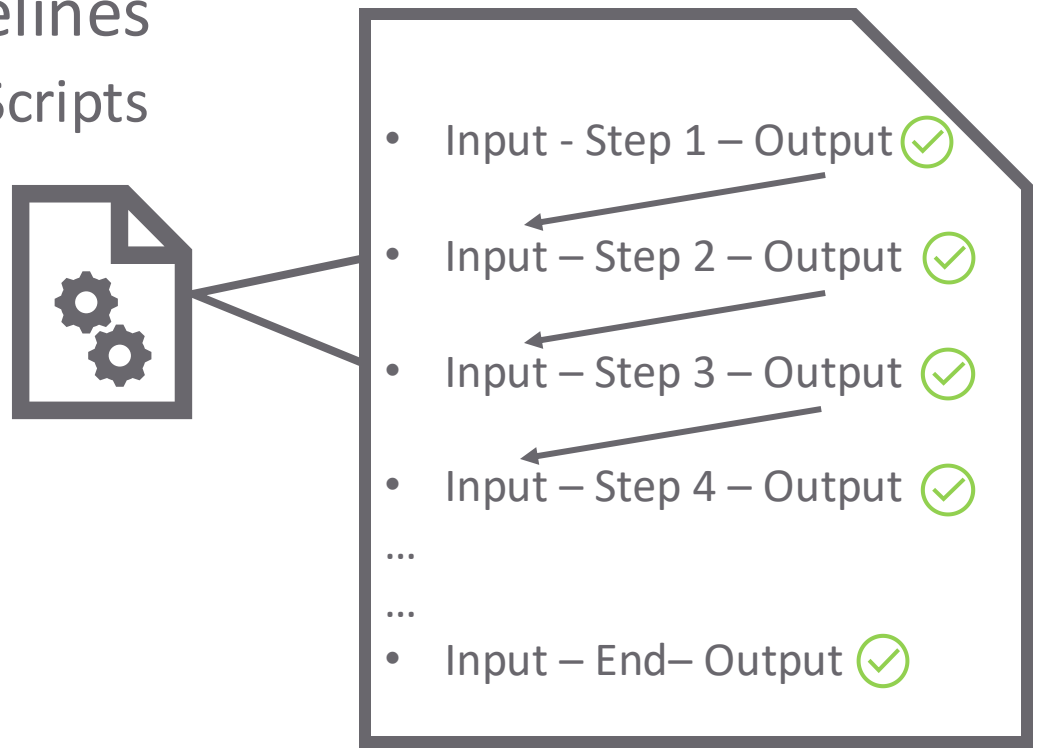
nextflow

Snakemake



REPRODUCTIBILITY

- Pipelines
  - Scripts



nextflow

Snakemake

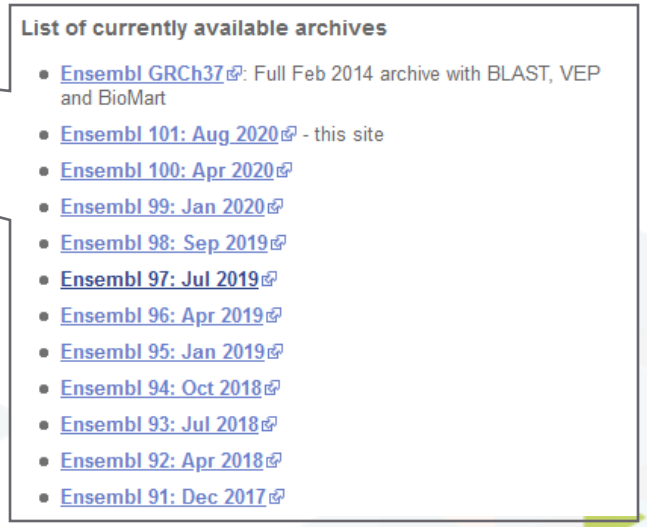
nf-core

REPRODUCTIBILITY

REPRODUCTIBILITY

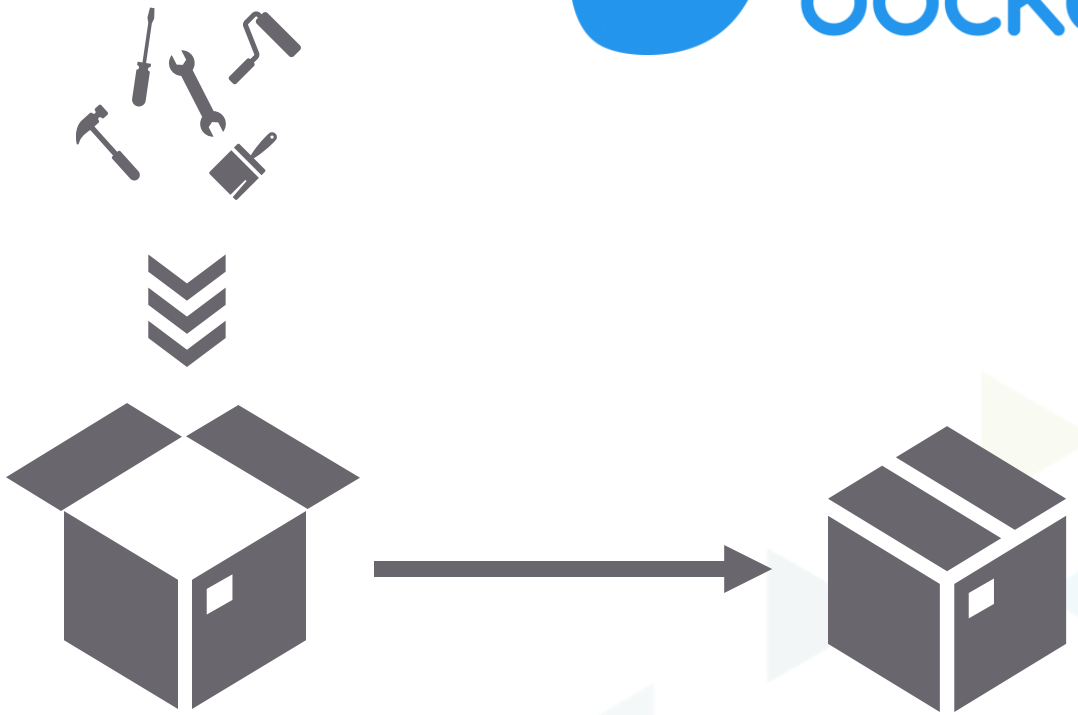
- Variability
  - Updates / Versioning
    - Softwares
    - References
  - Compatibility
  - Format
  - Knowledge

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

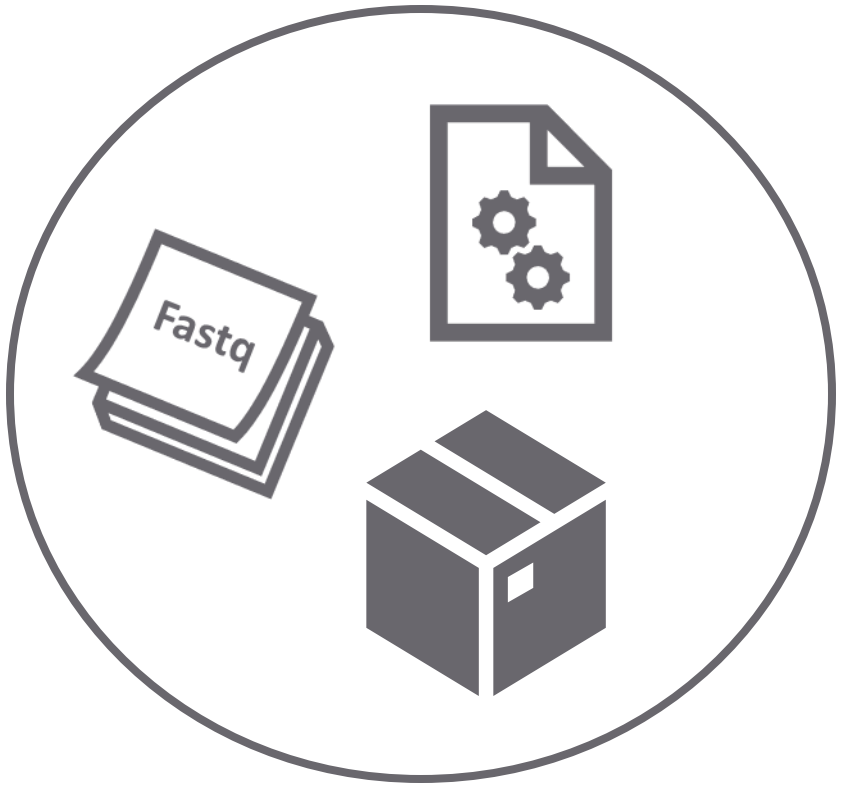


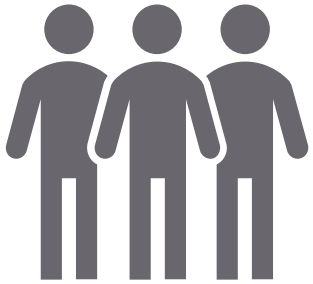
REPRODUCTIBILITY

- CONTAINERS
  - Docker
  - Singularity
- Softwares
  - Versions



REPRODUCTIBILITY





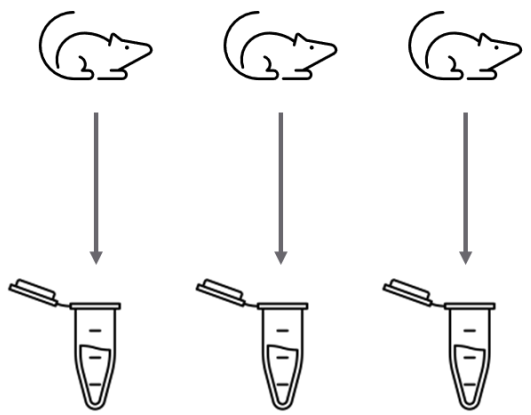
DATA DEPOSITORY

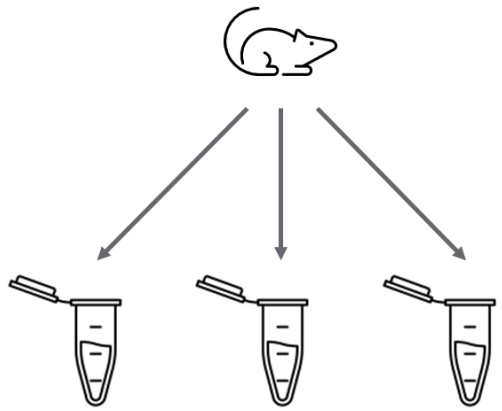
- Gene Expression Omnibus (NCBI)
- ArrayExpress (EMBL-EBI)

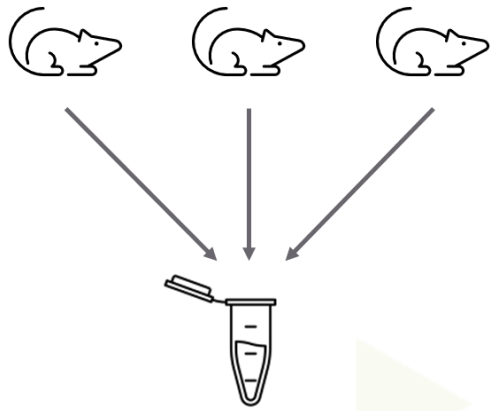
EXPERIMENTAL DESIGN

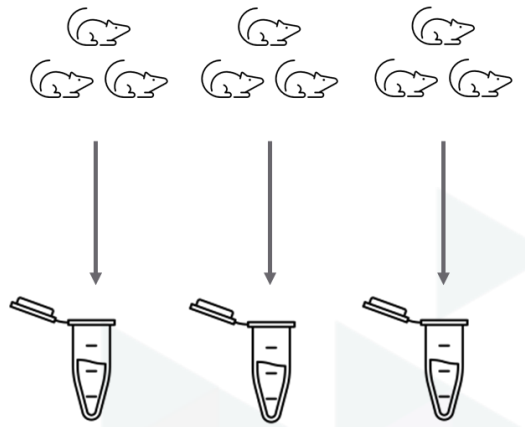
REPLICATES, POOLING & TRACKING

DESIGN

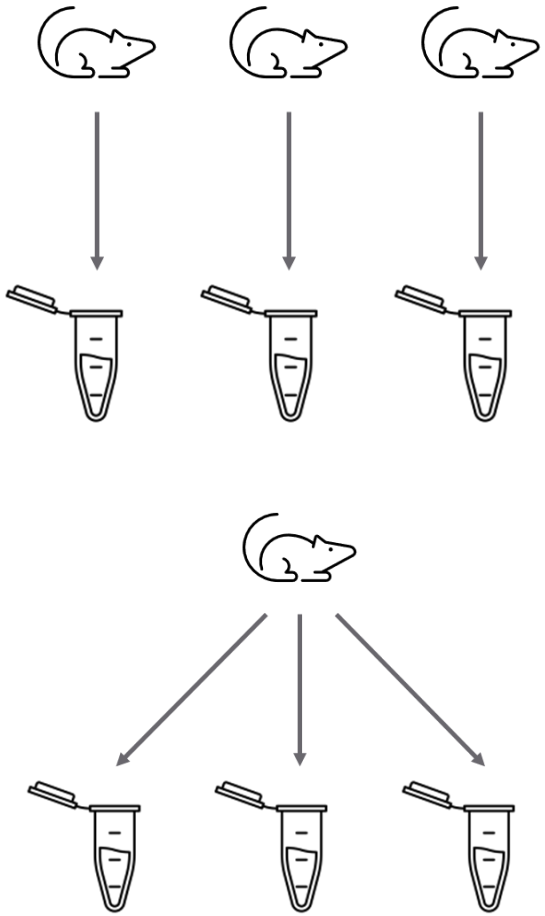


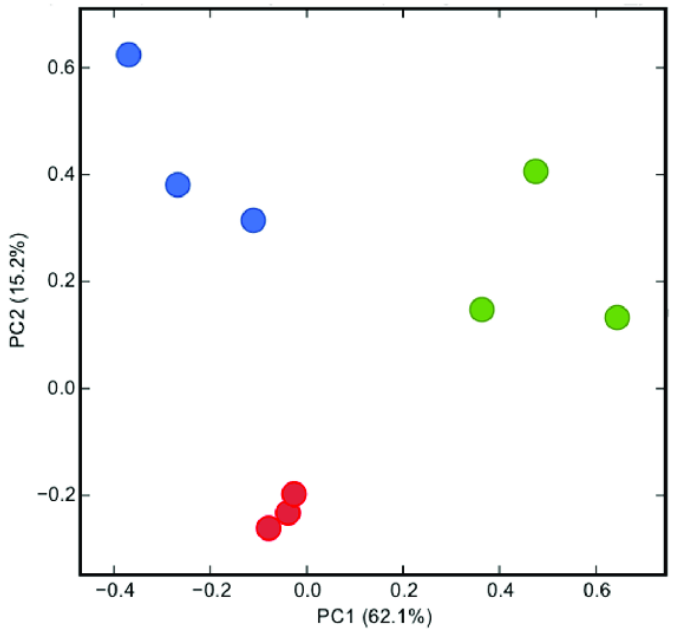






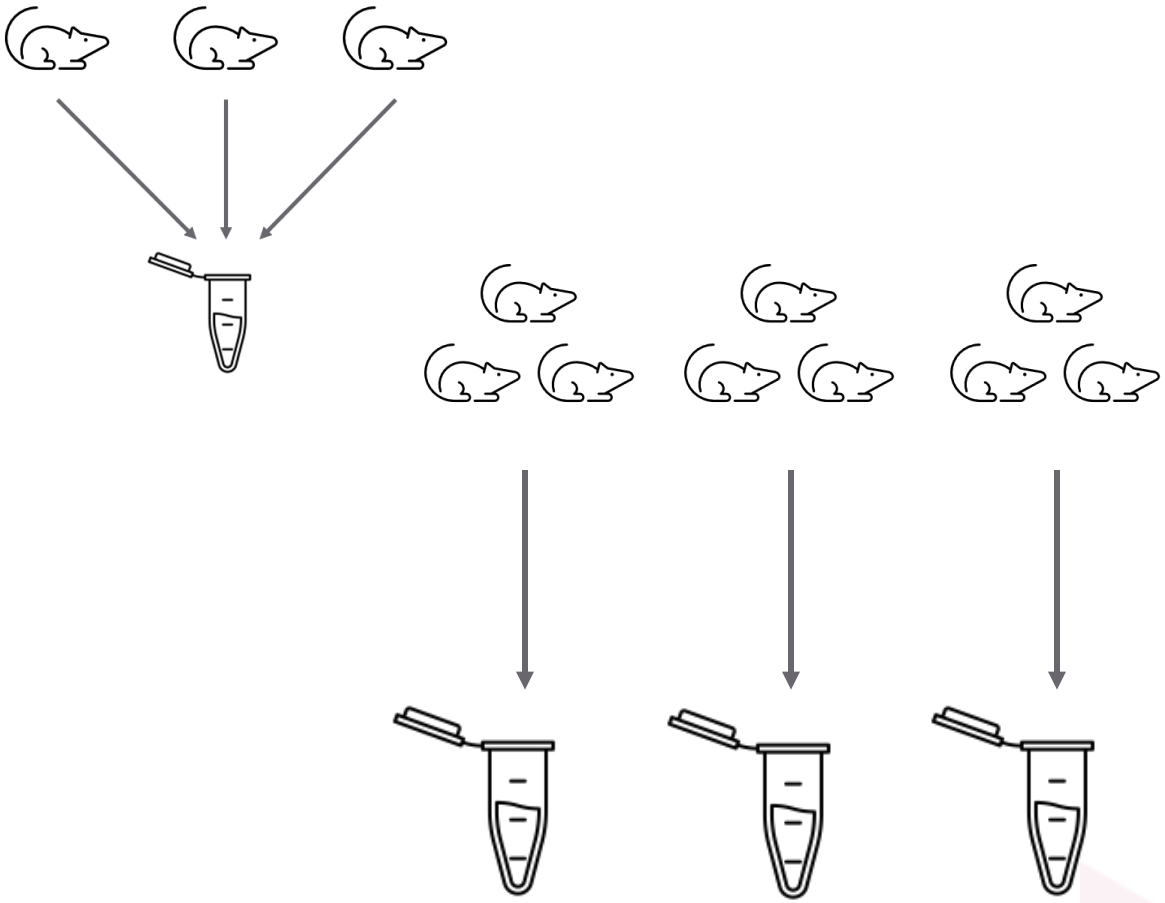
REPLICATES





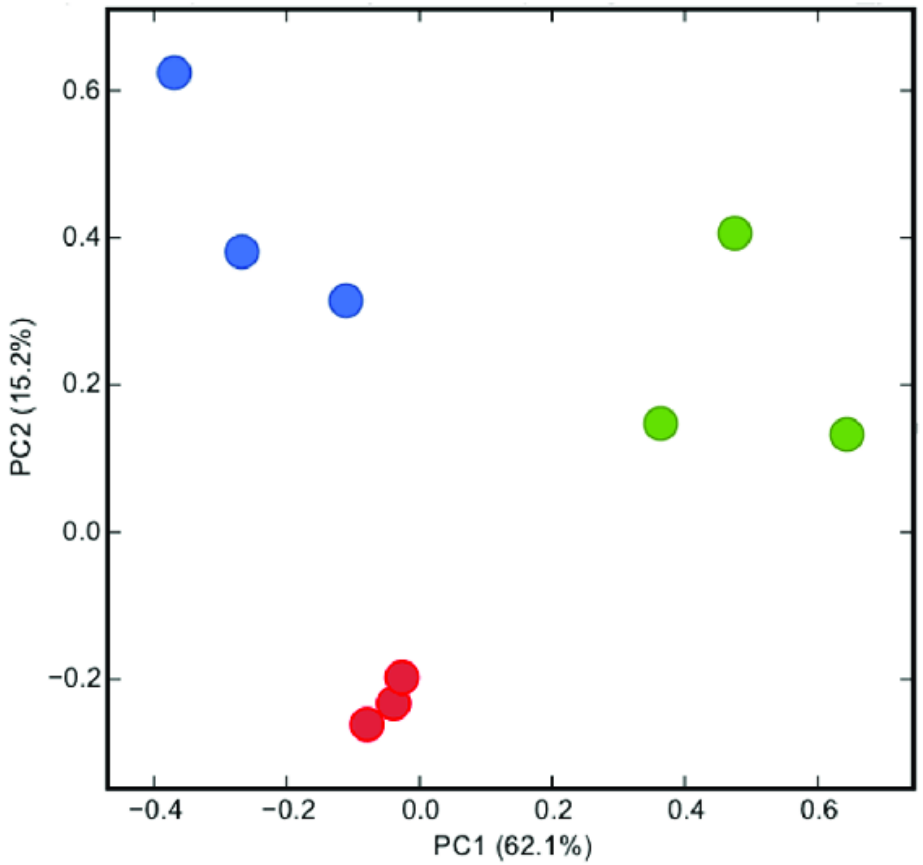
- 3 replicates / condition
  - Intra-variability < Inter-variability
  - Outliers
  - Different sources of variability
- The more, the better
  - Costs ++
- « Statistical test »
  - P-value
  - SE

POOLING



- Reduces sample variability
  - « More stability »
  - Reduces outlier impact
  - Costs --
- No individual informations
- 1 sample = no statistics

TRACKING



- « Sample history »
- Very sensitive

- Dissociate biological variability from experiment variability

- Age, Date, Time, Sex, Weigth, …
- Researcher, Kit, Instruments, …
- Observations

LIÈGE université
GIGA
Bioinformatics

THANK YOU FOR YOUR ATTENTION