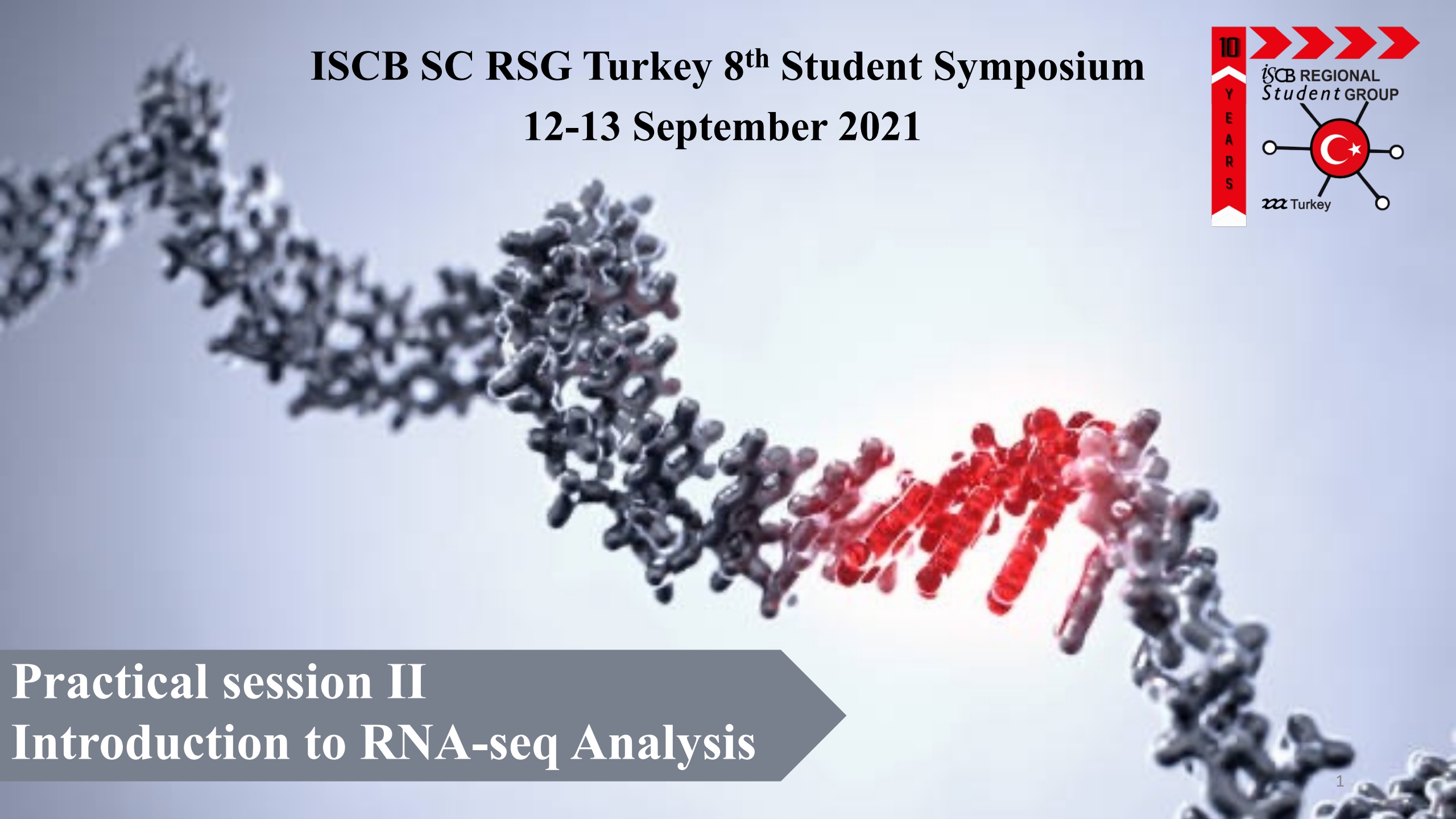
ISCB SC RSG Turkey 8th Student Symposium

12-13 September 2021



Practical session II
Introduction to RNA-seq Analysis

E. Ravza Gur
DPhil student

Center for Computational Biology
MRC Molecular Haematology Unit
MRC Weatherall Institute of Molecular Medicine
Radcliffe Department of Medicine
University of Oxford

Instructor at ERES Biotechnology

Twitter: ozturkravzae
LinkedIn: ravzagur
Instagram: biocomputationalist

- Genome Biology Group
- Functional Genomics & Machine Learning Group

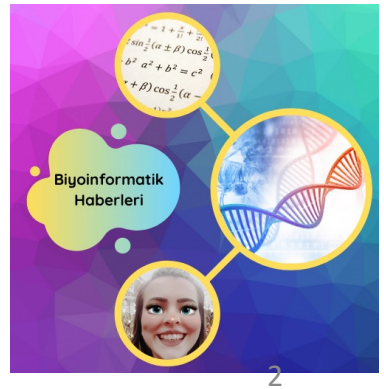
DeepC: predicting 3D genome folding using megabase-scale transfer learning

LANCEOTRON
deep learning-based peak caller

Octopus
variant caller

NG Capture-C **MCC**
Capture-C

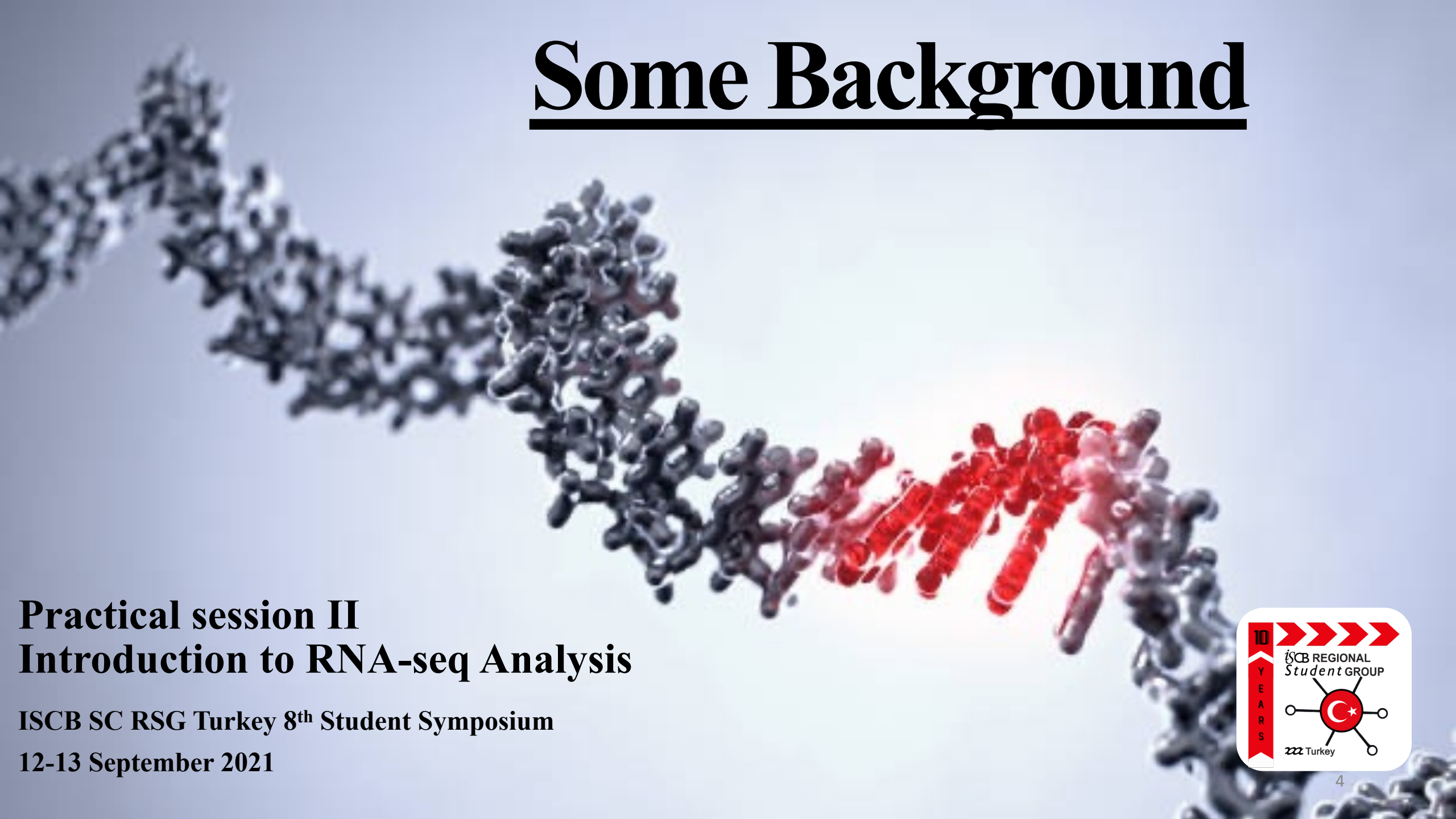
biocomputationalist



Content

- What are transcriptomics and RNA-seq?
- Getting familiar with technical concepts and data file formats
- Mapping sequencing reads to a reference genome using an aligner (STAR)
- Understanding the alignment file as an output of the mapping step
- Manipulating biological files via samtools
- Counting number of reads that mapped to genes
- Visualizing bam file via Integrative Genomics Viewer (IGV)

Some Background

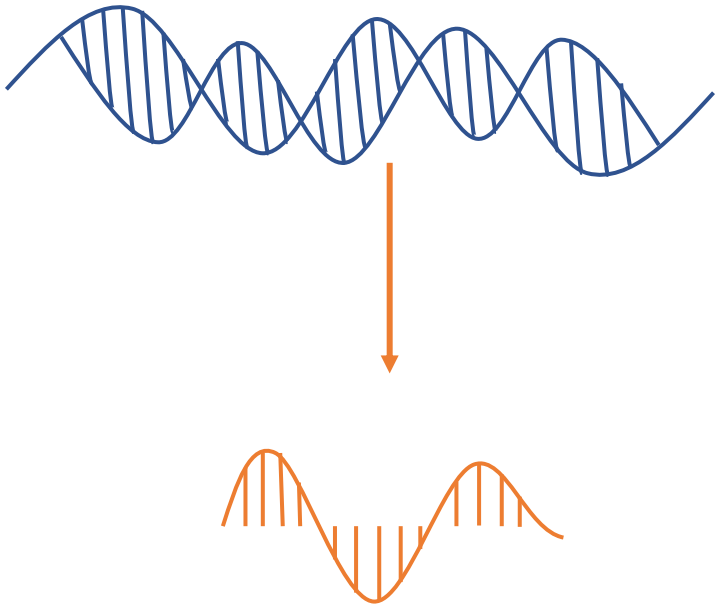


Practical session II
Introduction to RNA-seq Analysis

ISCB SC RSG Turkey 8th Student Symposium

12-13 September 2021

Transcriptomics and RNA-Seq



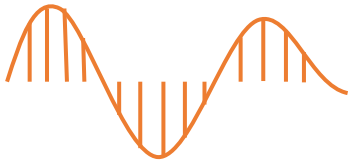
DNA ——→ All DNA

Genome

RNA ——→ All RNA

Transcriptome

Transcriptomics and RNA-Seq



RNA ——→ All RNA

Transcriptome

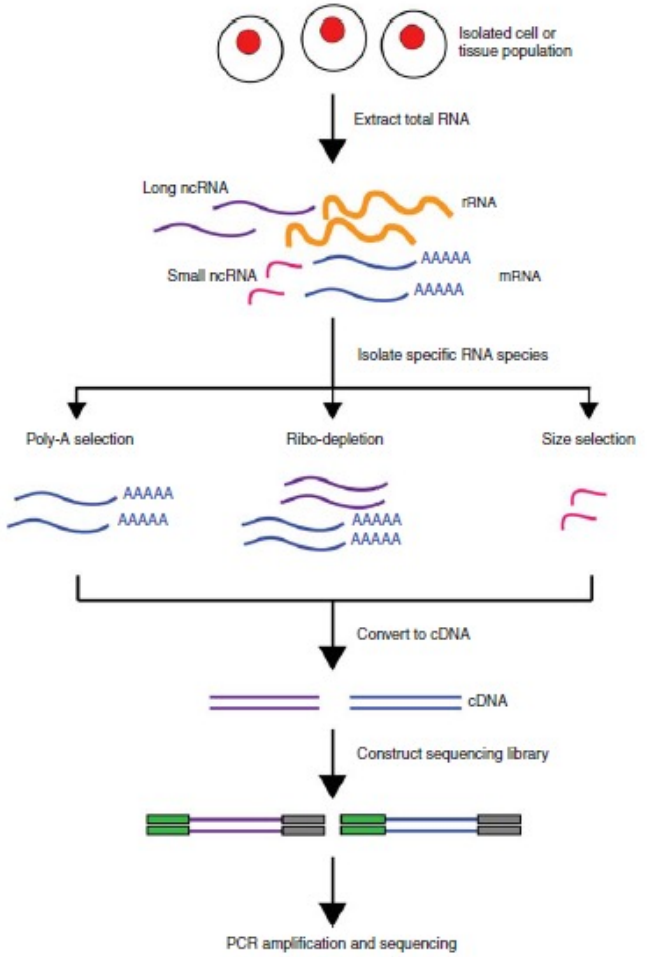
The aim of RNA-seq can be

- To quantify mRNA abundance,
- To determine start sites, 5' and 3' ends, alternate splicing
- To quantify the different expression levels of each transcript during development or in disease.

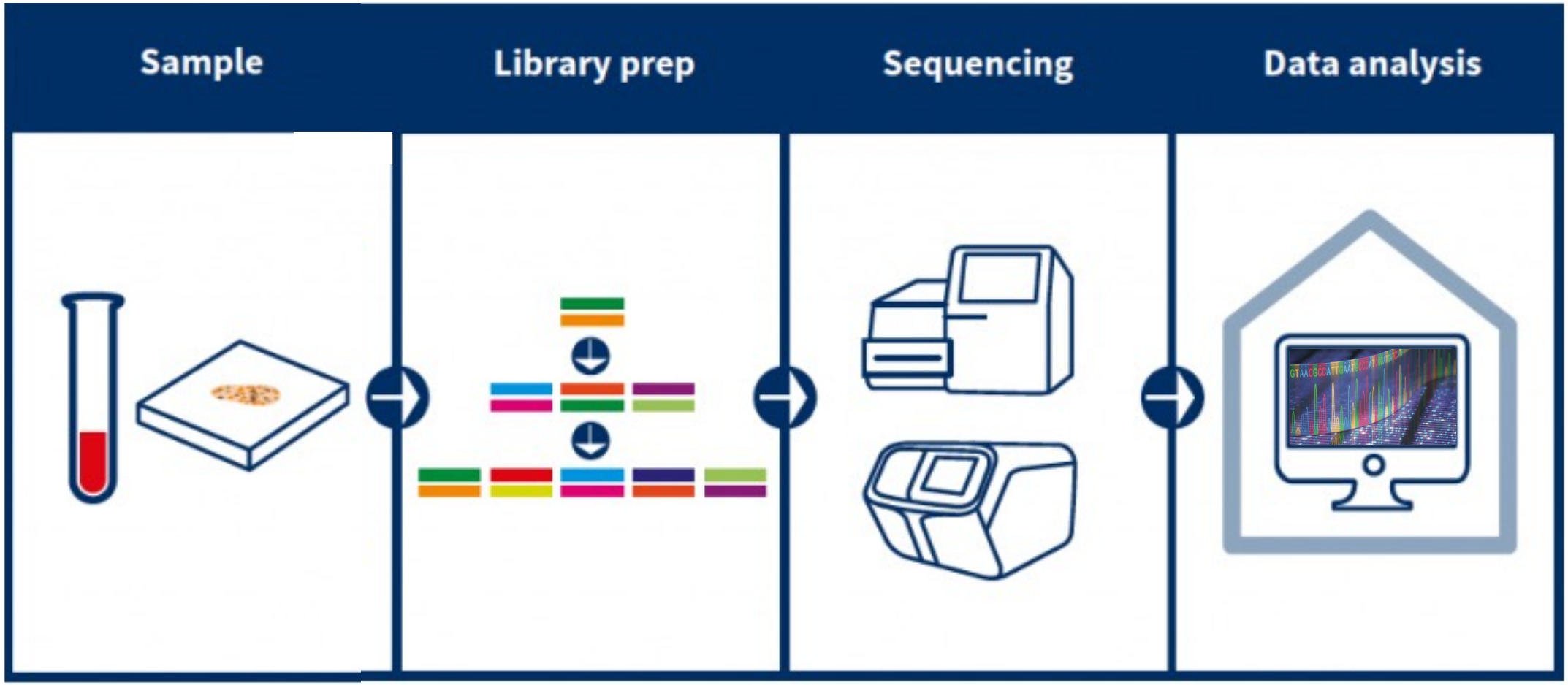
Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10,** 57–63 (2009). https://doi.org/10.1038/nrg2484

Types of RNA-Seq

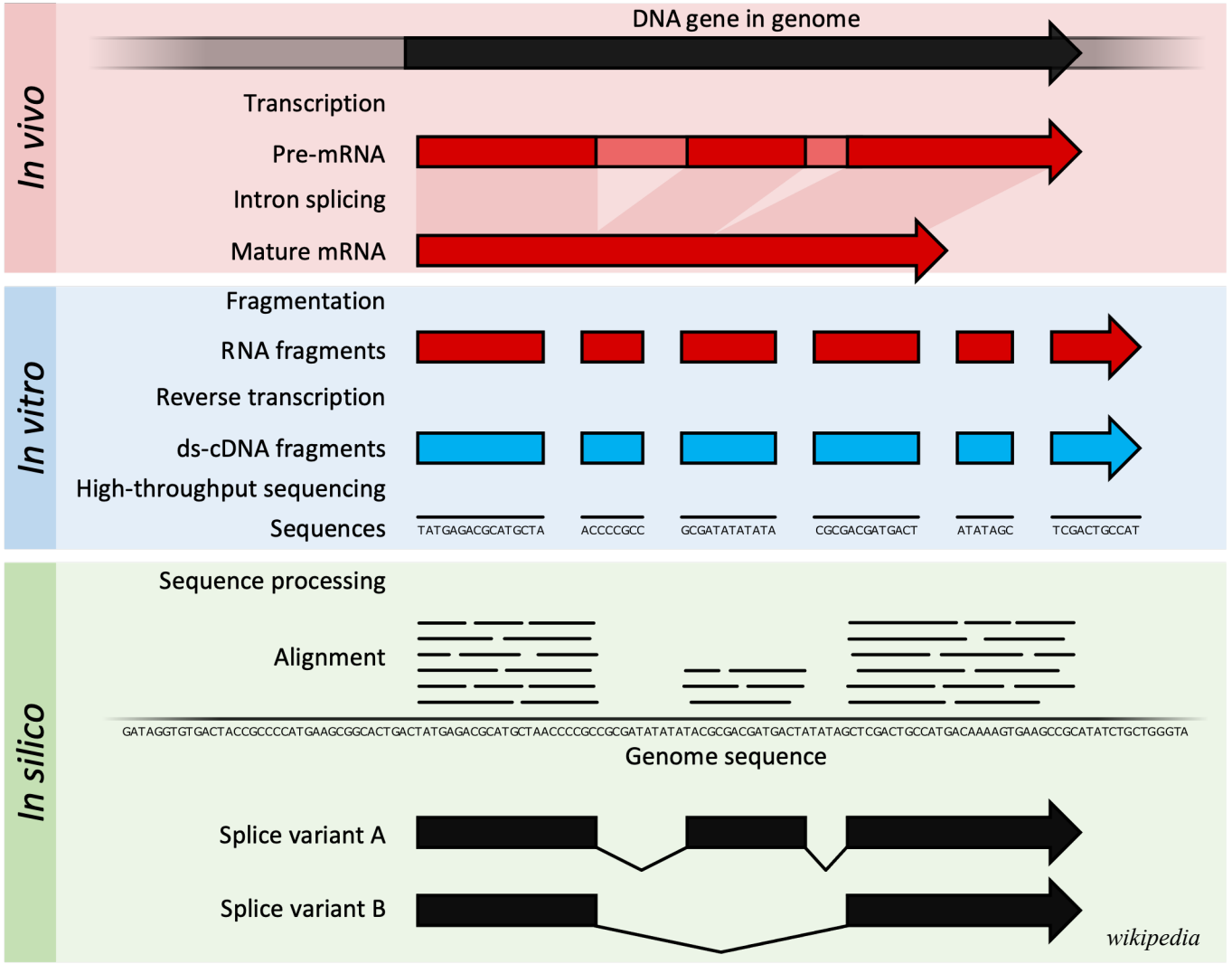
- Poly A +
- Poly A –
- Total RNA
- Nascent
- Nanocage/CAGE/CAP-Seq



RNA-Seq Workflow



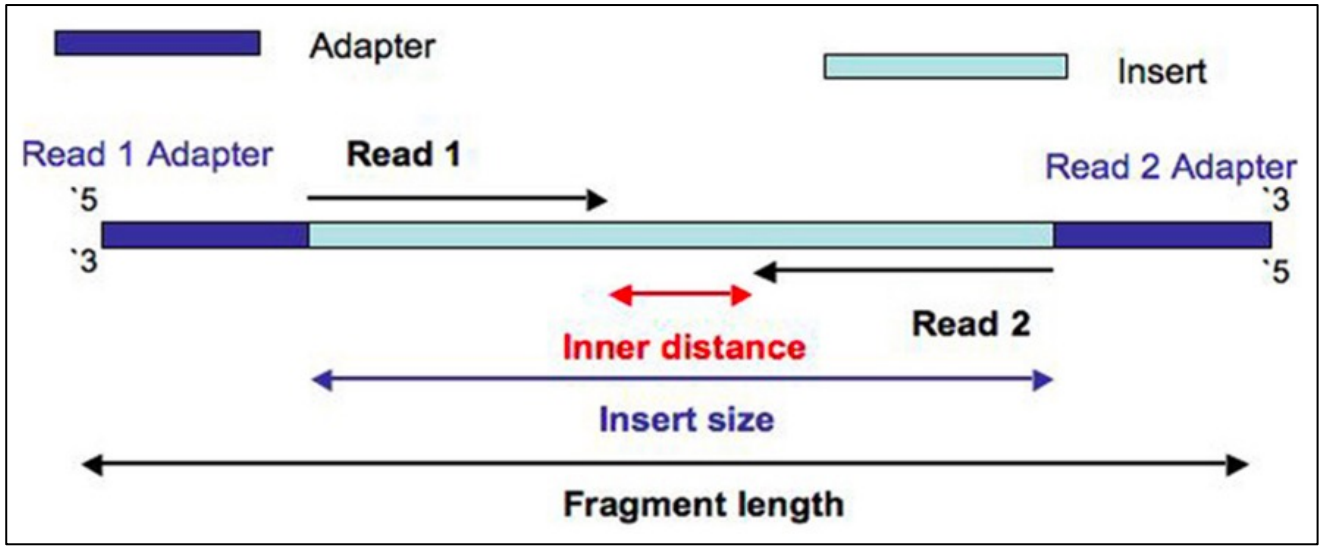
RNA-Seq Workflow

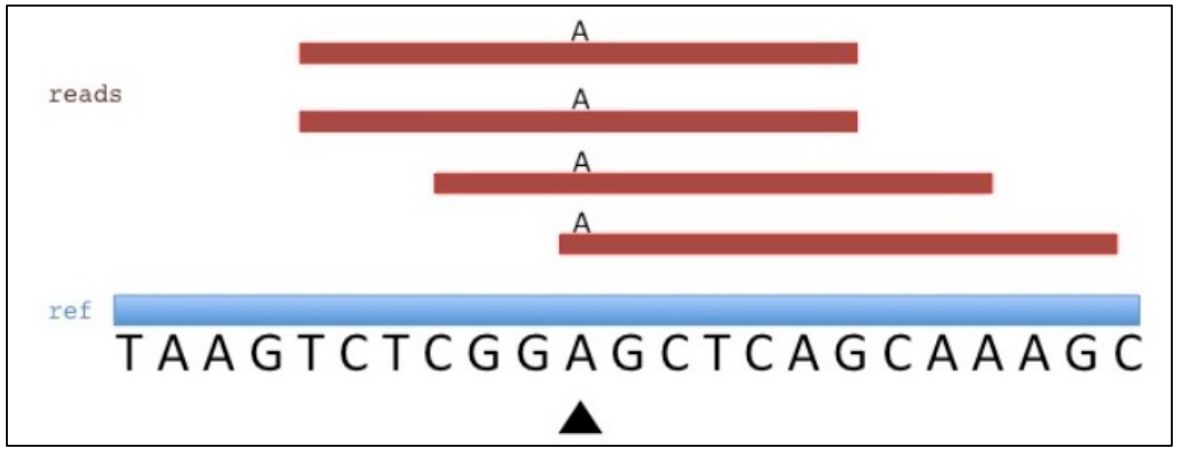


Terminology

Insert size?
Fragment size?
Pair-end?
Single-end?

Replicates
RNA
libraries





Sequencing depth?
Coverage depth?
Library size?

What is a read?

@SIM:1:FCX:1:15:6329:1045 1:N:0:2
AGCTCAGAAAGCTTCGCTCCAAG
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=

@SIM:1:FCX:1:15:6329:1045 1:N:0:3
TCGCACTCAACGCCCTGCATATGA
+
<>;##=><9=AAAFI01AAAA9#:<#<;<<<????#=

AGCTCAGAAAGCTTCGCTCCAAG

map/align

chr1 2345 AGCTCAGAAAGCTTCGCTCCAAG

FASTA and FASTQ formats

FastA

```
>SeqID HEADER
TAATTTGGTAACGGCTGATGGTGGACCGCA
AGAAGGTTATCCATATCGTG
```

It only contains sequence information.

Qual

```
>SeqID HEADER
33 33 33 37 37 37 37 37 37 37 37 37 37 40 40 40 40 37 37 40
40 33 37 37 40 40 40 37 40 40 40 40 40 40 37 40 37 37 37 40
40 37 40 37 33 06 15 27 15 22
```

It only contains quality information.
Heavy file: 3 bytes / base.

FastQ

```
@SeqID HEADER
TAATTTGGTAACGGCTGATGGTGGACCGCA
AGAAGGTTATCCATATCGTG
+
BBBFFFFFFFFFFIIIIFFIIBFFIIIFII
IIIIFIFFFIIFIFB'0<07
```

Contains both sequence *and* quality information.
Quality: 1 byte / base.

ASCII values: 33 to 126 → Quality values: 0 to 93.

Symbol	Q-score
B	33
F	37
I	40
‘	6
0	15
<	27
7	22

Phred Quality

$$Q_{\text{phred}} = -10 \log_{10} e$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy	ASCII	Character
20	1 in 100	99%	53	5
30	1 in 1,000	99.9%	63	?
40	1 in 10,000	99.99%	73	I

https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm

Quality Control

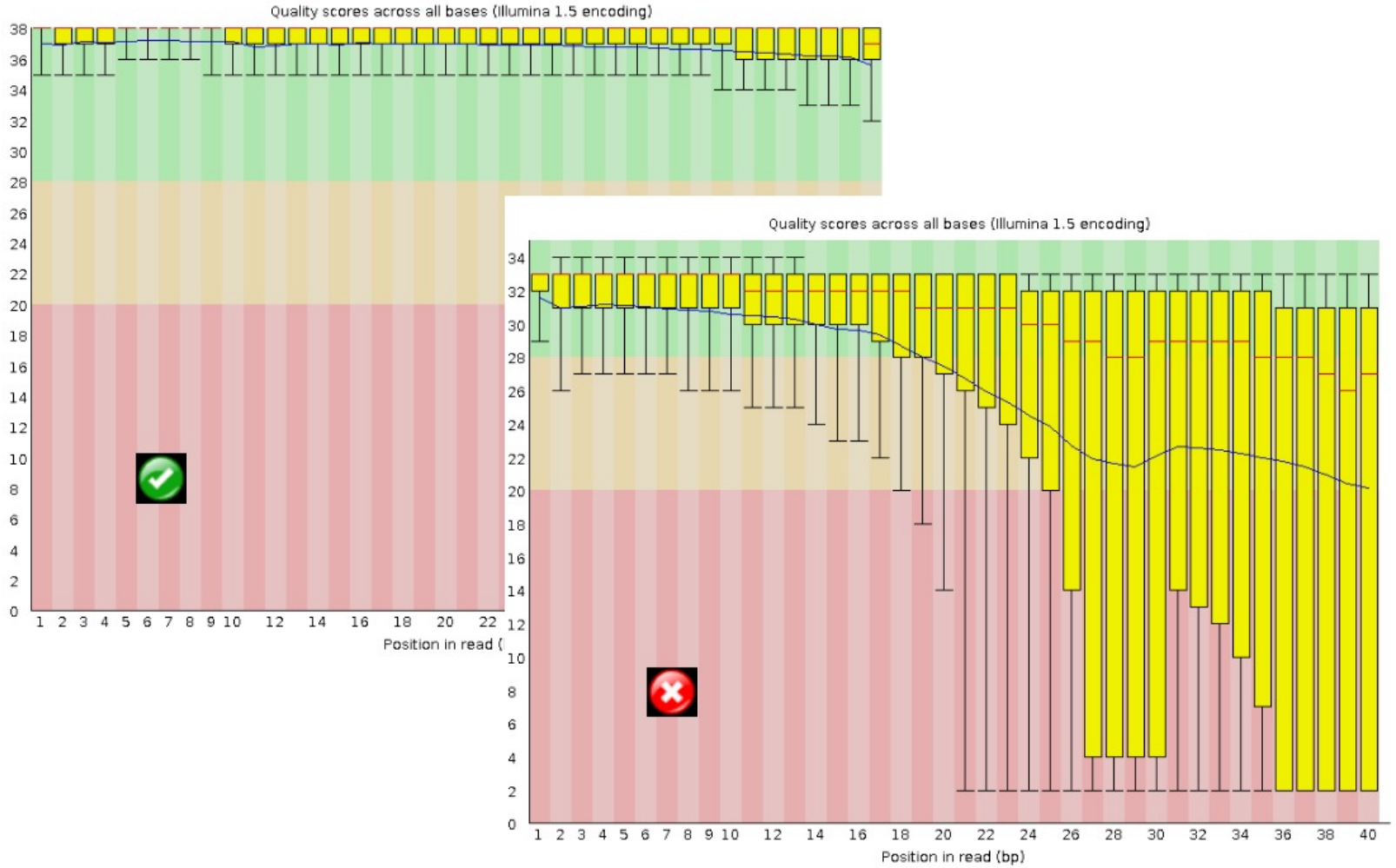
FastQC
Widely used for Illumina data because it's fast. It works on a subset of reads.
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

PrinseQ
Used for smaller datasets because it computes every sequence.
http://prinseq.sourceforge.net/

FastQC

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



Alignment

Bowtie

- Ultrafast
- Memory-efficient

Tophat

- Fast splice junction mapper
- Identify splice junctions btw exons

STAR - Spliced Transcripts Alignment to a Reference

- Ultrafast universal RNA-seq aligner
- Outperform other aligners

Alignment and sam/bam format

SAM | Sequence Alignment/Map

BAM | Binary Alignment/Map

Reference sequence with aligned reads	CIGAR string	Explanation
C T G C A T G T T A G A T A A * * G A T A G C T G T G C T A		
A A G G A T A * C T G	1M2I4M1D3M	Insertion & Deletion
G A T A A * G G A T A	5M1P1I4M	Padding & Insertion
T G T T A ▬▬▬▬▬▬▬▬▬▬ T G C T A	5M15N5M	Spliced read
a a a C A T G T T A G	3S8M	Soft clipping
A A A C A T G T T A G	3H8M	Hard clipping

```
@HD     VN:1.5  SO:coordinate
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
…
K00198:242:HLGYVBBXX:8:1119:6137:36112   163   1   12636   1   75M   =   13406   845
CCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCA
AAFFFJJJJJJJJJJJJJJJJJJJAJJJJJJJJFJJJJ7FJJJ<FFJJFJFJJJJJJJJJJJJJJJJJFFJF NH:i:3   HI:i:1   AS:i:148   nM:i:0

K00198:242:HLGYVBBXX:8:1119:6137:36112   83   1   13406   1   75M   =   12636   -845
CTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCCAGTTTCACCAGAAGTAGGCCTCTTCCTGAC
JJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA   NH:i:3   HI:i:1   AS:i:148   nM:i:0
```

Flag 83
- read paired (0x1)
- read mapped in proper pair (0x2)
- read reverse strand (0x10)
- first in pair (0x40)

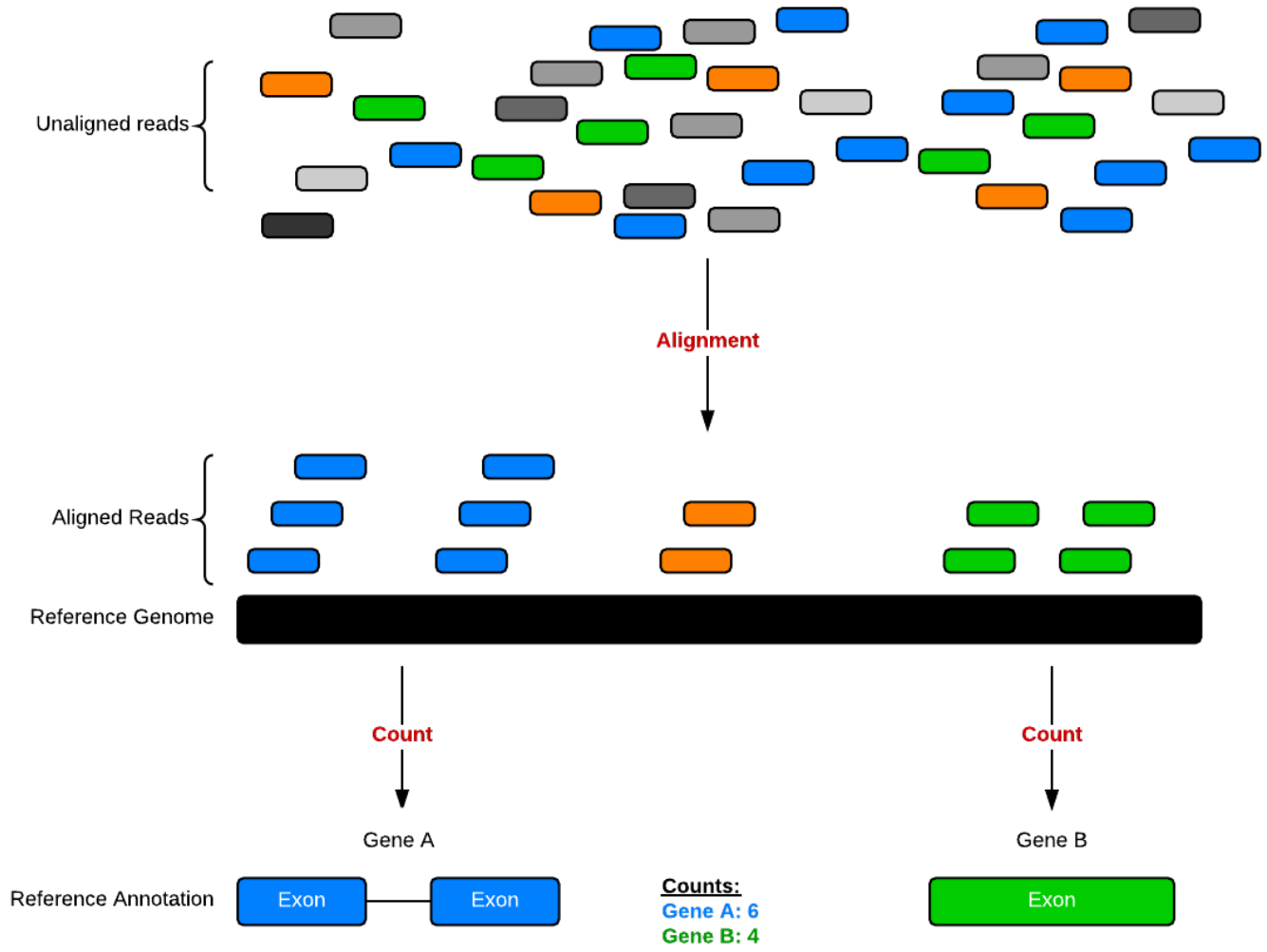
Flag 163
- read paired (0x1)
- read mapped in proper pair (0x2)
- mate reverse strand (0x20)
- second in pair (0x80)

Pos.	Field	Example entry	Description	NA value
1	QNAME	Read1	Query template (= read) name (PE: read pair name)	required
2	FLAG	83	Information about the read's mapping properties encoded as bit-wise flags (see next section and Table 4).	required
3	RNAME	chrI	Reference sequence name. This should match a `@SQ` line in the header.	*
4	POS	15364	1-based leftmost mapping position of the first matching base. Set as 0 for an unmapped read without coordinates.	0
5	MAPQ	30	Mapping quality of the alignment. Should be a Phred-scaled posterior probability that the position of the read is incorrect, but the value is completely dependent on the alignment program. Some tools set this to 0 if multiple alignments are found for one read.	0
6	CIGAR	51M	Detailed information about the alignment (see below).	*
7	RNEXT	=	PE reads: reference sequence name of the next read. Set to "=" if both mates are mapped to the same chromosome.	*
8	PNEXT	15535	PE reads: leftmost mapping position of the next read.	0
9	TLEN	232	PE reads: inferred template length (fragment size).	0
10	SEQ	CCA...GGC	The sequence of the aligned read on the forward strand (not including indels).	*
11	QUAL	BBH...1+B	Base quality (same as the quality string in the `FASTQ` format, but always in Sanger format [ASCII+33]).	*
12ff	OPT	NM:i:0	Optional fields (format: `<TAG>:<TYPE>:<VALUE>`; see below).	

Transcript quantification

HTSeq-count

featureCounts



Count data

samples: want to see if differences across condition are significant (w.r.t. biological and technical variation)

features (e.g. genes)

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

Count data

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

These are the “raw” counts and will be used in statistical programs downstream for differential gene expression.

Counting reads with featureCounts

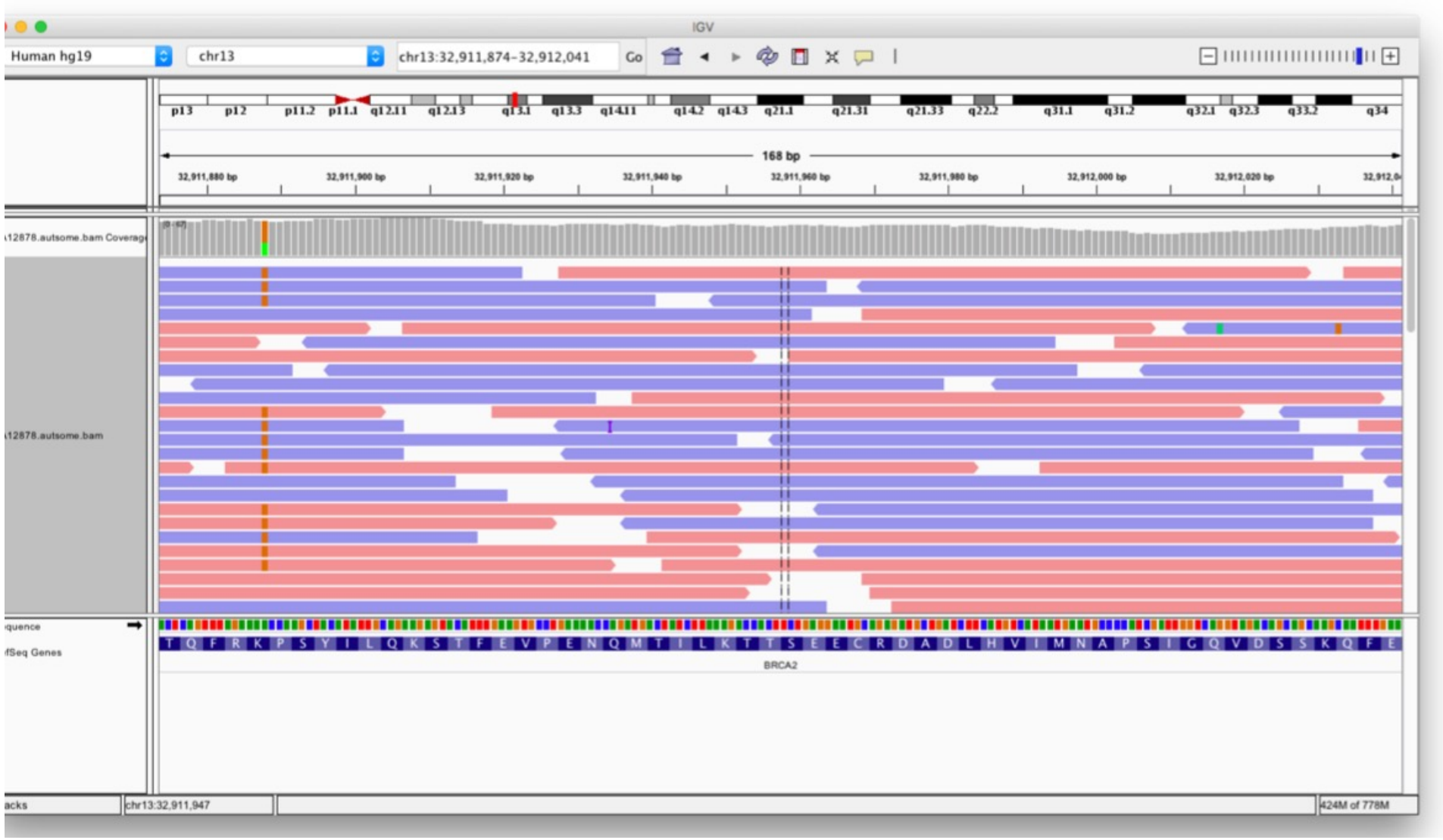
featureCounts takes as input SAM/BAM files and an annotation file.

The annotation file should be in either GTF format or a simplified annotation format (SAF) as shown below (columns are tab-delimited):

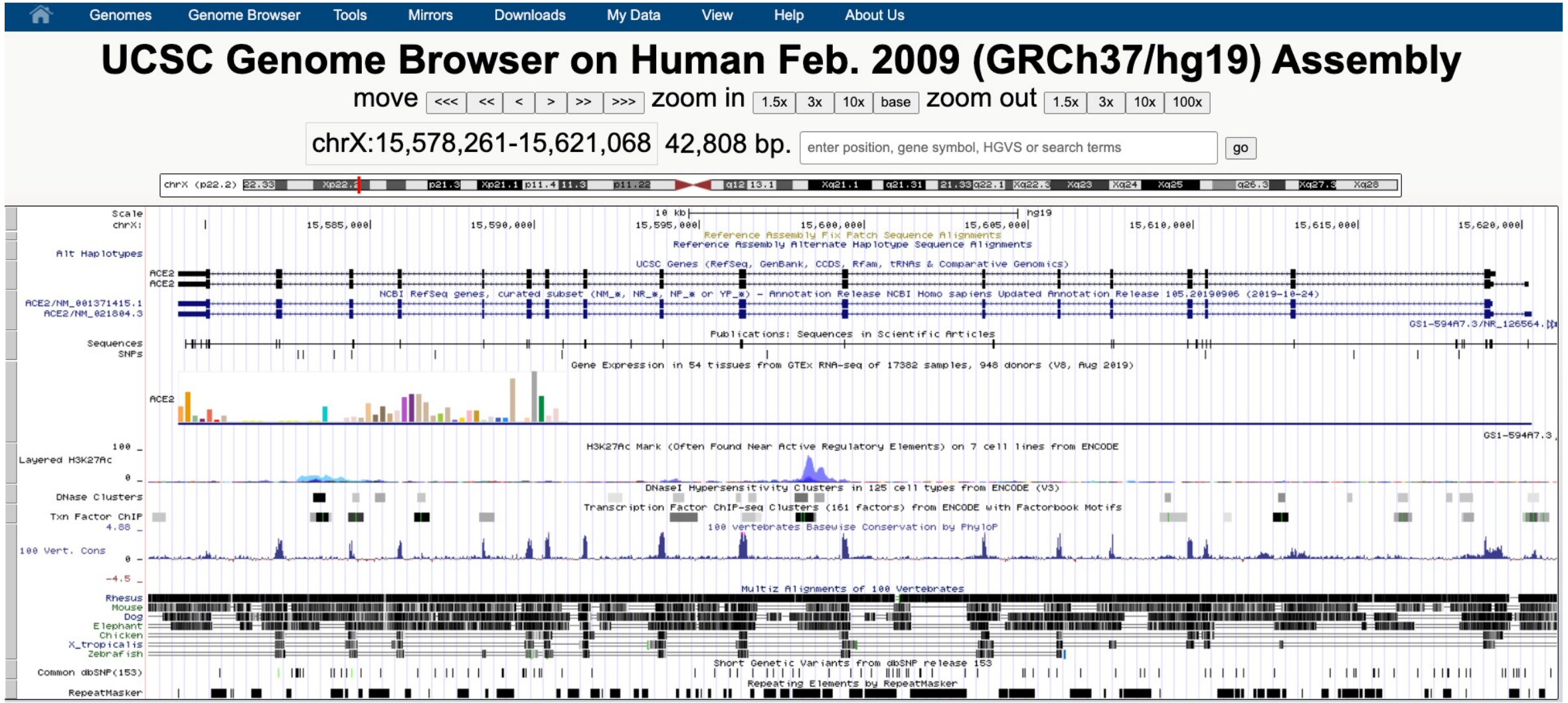
GeneID	Chr	Start	End	Strand
497097	chr1	3204563	3207049	–
497097	chr1	3411783	3411982	–
497097	chr1	3660633	3661579	–

…

The Integrative Genomics Viewer (IGV)



UCSC Genome Browser



Let's practice

Practical session II
Introduction to RNA-seq Analysis

ISCB SC RSG Turkey 8th Student Symposium

12-13 September 2021

Practice with data

Tutorial

- https://github.com/griffithlab/rnaseq_tutorial
- https://rnabio.org/
- 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.

Publicly Available Dataset

- RNA samples: Universal Human Reference (UHR) and Human Brain Reference (HBR)
- The UHR is total RNA isolated from a diverse set of 10 cancer cell lines.
- *The HBR is total RNA isolated from the brains* of 23 Caucasians (male and female, of varying age but mostly 60-80 years old)

Practice with data

Publicly Available Dataset we will use today

- *The HBR is total RNA isolated from the brains* of 23 Caucasians (male and female, of varying age but mostly 60-80 years old)
- wget http://genomedata.org/rnaseq-tutorial/HBR_UHR_ERCC_ds_5pc.tar

```
(base) RavzaOzturksMBP:data ravza$ ls -lhrt
total 569800
-rw-r--r--  1 ravza  staff   6.6M  6 Nov  2014 HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2.fastq.gz
-rw-r--r--  1 ravza  staff   6.3M  6 Nov  2014 HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz
-rw-r--r--  1 ravza  staff   7.6M  6 Nov  2014 HBR_Rep2_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz
-rw-r--r--  1 ravza  staff   8.1M  6 Nov  2014 HBR_Rep2_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2.fastq.gz
-rw-r--r--  1 ravza  staff   6.9M  6 Nov  2014 HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz
-rw-r--r--  1 ravza  staff   7.3M  6 Nov  2014 HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2.fastq.gz
-rw-r--r--  1 ravza  staff    14M  6 Nov  2014 UHR_Rep1_ERCC-Mix1_Build37-ErccTranscripts-chr22.read2.fastq.gz
-rw-r--r--  1 ravza  staff    13M  6 Nov  2014 UHR_Rep1_ERCC-Mix1_Build37-ErccTranscripts-chr22.read1.fastq.gz
-rw-r--r--  1 ravza  staff   9.7M  6 Nov  2014 UHR_Rep2_ERCC-Mix1_Build37-ErccTranscripts-chr22.read1.fastq.gz
-rw-r--r--  1 ravza  staff    10M  6 Nov  2014 UHR_Rep2_ERCC-Mix1_Build37-ErccTranscripts-chr22.read2.fastq.gz
-rw-r--r--  1 ravza  staff    11M  6 Nov  2014 UHR_Rep3_ERCC-Mix1_Build37-ErccTranscripts-chr22.read2.fastq.gz
-rw-r--r--  1 ravza  staff    11M  6 Nov  2014 UHR_Rep3_ERCC-Mix1_Build37-ErccTranscripts-chr22.read1.fastq.gz
-rw-r--r--  1 ravza  staff   111M 23 Oct  2018 HBR_UHR_ERCC_ds_5pc.tar
```

Checking reads' quality with Fastqc

- ☞ `conda install -c bioconda fastqc`
- ☞ `which fastqc`
- ☞ `fastqc –help`

- ☞ `fastqc -o fastqc_report/ \`

`data/HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz \`

`data/HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2.fastq.gz`

☞ (open) HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1_fastqc.html

(This command might be problematic for some of you)

Aligning read to human genome hg19

- `conda install -c bioconda star`
- http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/chr22.fa.gz

```
STAR --runThreadN 2 \
     --runMode genomeGenerate \
     --genomeDir chr22_hg19_index \
     --genomeFastaFiles chr22.fa
```

~ 1-2 mins

Aligning read to human genome hg19

☞ Usage: STAR [options]... --genomeDir /path/to/genome/index/ \
--readFilesIn R1.fq R2.fq

☞ STAR --genomeDir ~/Documents/RSG-Turkey/Sempozyum/RNA_demo/alignment/chr22_hg19_index \
--readFilesIn ~/Documents/RSG-Turkey/Sempozyum/RNA_demo/data/HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz ~/Documents/RSG-Turkey/Sempozyum/RNA_demo/data/HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read2.fastq.gz \
--outSAMtype BAM SortedByCoordinate \
--outSAMattributes Standard \
--runThreadN 2 \
--outFileNamePrefix HBR_Rep1_ERCC-Mix2_hg19_ErccTranscripts-chr22

~ 7-8 mins

Data filtering with samtools

Samtools installation via conda

☞ `conda install -c bioconda samtools`

View binary alignment file

☞ `samtools view –h {name}.bam | less –S`

Sorting bam file

☞ `samtools sort {name}.bam -o {name}.sorted.bam`

Data filtering with samtools

Counting the number of reads which are paired and mapped in proper pair

☞ samtools view -c -f 3 {name}.bam

Filtering the number of reads which are paired and mapped

☞ samtools view -b -h -f 3 {name}.bam > {name}_properpairs.bam

Filter bad quality reads

☞ samtools view -bShuF 4 -f 2 -q 30 {name}_properpairs.bam > {name}_properpairs_q30.bam

Data filtering with samtools

Removing PCR duplicates

☞ samtools rmdup {name}_properpairs_q30.bam >
{name}_properpairs_q30_rmvdup.bam

Sort bam file by read name

☞ samtools sort -n {name}_properpairs_q30_rmvdup_srtByreadName.bam

Indexing bam file

☞ samtools index {name}_properpairs_q30_rmvdup_srtByreadName.bam

Counting reads that mapped to genes

To be able to use featureCounts, we need to install **subread** package using conda.

☞ `conda install -c bioconda subread`

We can access Information about how we can run featureCount by typing the below command into the terminal.

☞ `featureCounts`

Usage

☞ `featureCounts [options] -a <annotation_file> -o <output_file> inputFile1 [inputFile2] ...`

featureCounts

Usage

☞ featureCounts [options] -a <annotation_file> -o <output_file> inputFile1 [inputFile2] ...

Download gtf file for human genome 19 (37)

☞ wget http://ftp.ensembl.org/pub/grch37/release-104/gtf/homo_sapiens/Homo_sapiens.GRCh37.87.chr.gtf.gz

Run featureCounts

☞ featureCounts -T 2 -a Homo_sapiens.GRCh37.87.chr.gtf -t exon -g gene_id -o feature_counted.txt -p -s 0 HBR_Rep1_chr22_sortByName.bam

featureCounts

Run featureCounts

☞ featureCounts -T 2 -a Homo_sapiens.GRCh37.87.chr.gtf -t exon -g gene_id -o feature_counted.txt -p -s 0 HBR_Rep1_chr22_sortByName.bam

☞ -T number of threads
☞ -a specifies annotation file
☞ -t specify feature type to count
☞ -g Specify attribute type in GTF annotation
☞ -o set output file name
☞ -p specify if data is paired
☞ -s set the strand specificity (0=unstranded, 1=stranded)

featureCounts in R

☞ `fc <- featureCounts(bam.files, annot.inbuilt="mm9", isPairedEnd = T)`

☞ `count_fc <- fc$counts`

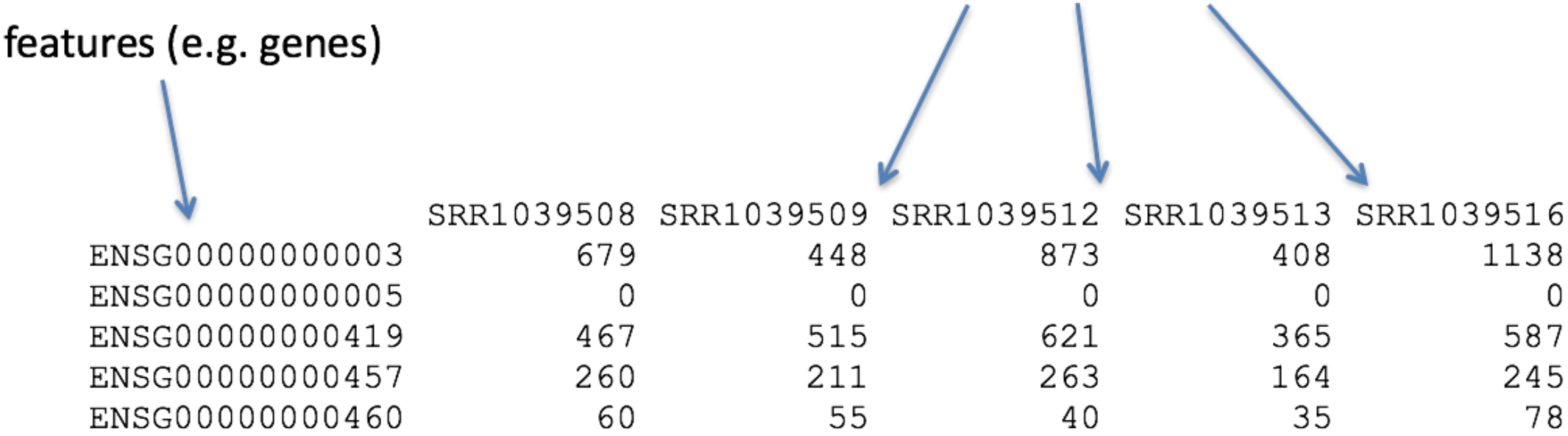
☞ `data.frame(count_fc)`

	HBR_Rep1_chr22_sortByName.bam
653635	0
100422834	0
645520	0
79501	0
729737	0
387590	1
27439	26
27443	85
529	398

Count data

samples: want to see if differences across condition are significant (w.r.t. biological and technical variation)

features (e.g. genes)



	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

Visualize result using IGV

- Download IGV in your computer
- https://software.broadinstitute.org/software/igv/download

Visualize result using IGV

- **Data:**

http://ftp.ensembl.org/pub/grch37/release-104/data_files/homo_sapiens/GRCh37/rnaseq/

- Create subfolder in the RNA-seq analysis directory for IGV_data and download the above dataset to visualize in IGV.

- Open IGV app and load from file.

Advanced reading

☞ https://www.bioconductor.org/

☞ https://www.ncbi.nlm.nih.gov/gds
☞ GSE38823 example

Advanced reading

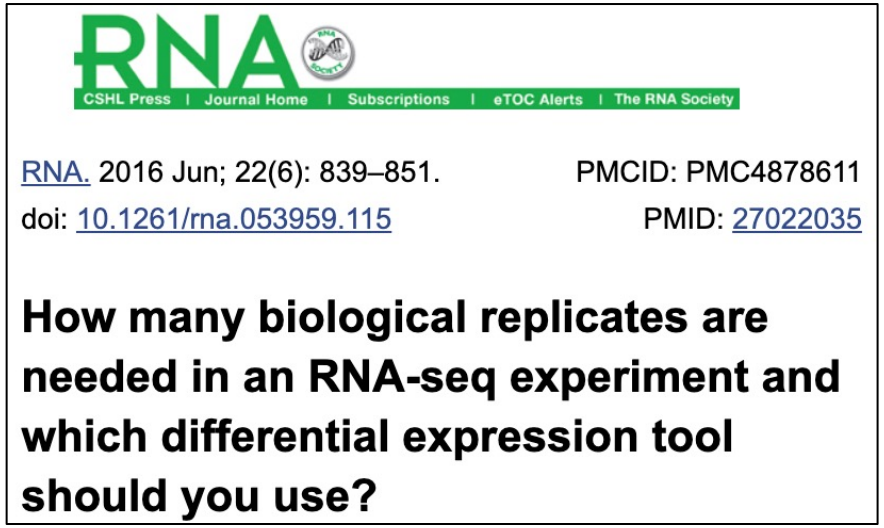
Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

Genome Biology

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

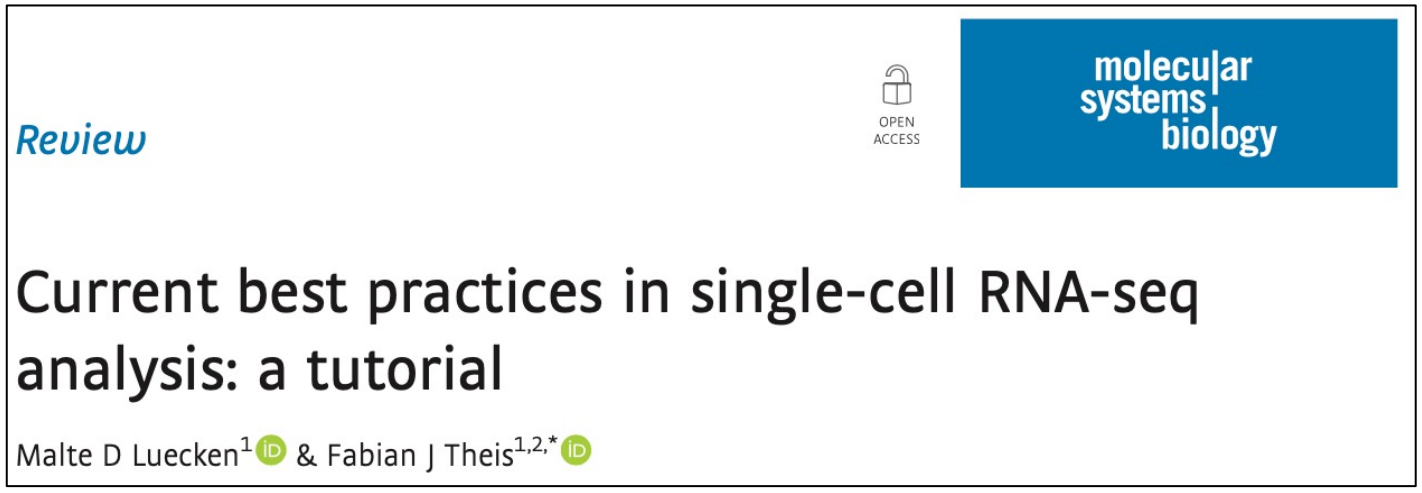


RNA

CSHL Press | Journal Home | Subscriptions | eTOC Alerts | The RNA Society

RNA. 2016 Jun; 22(6): 839–851. PMCID: PMC4878611
doi: 10.1261/rna.053959.115 PMID: 27022035

How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?



Review

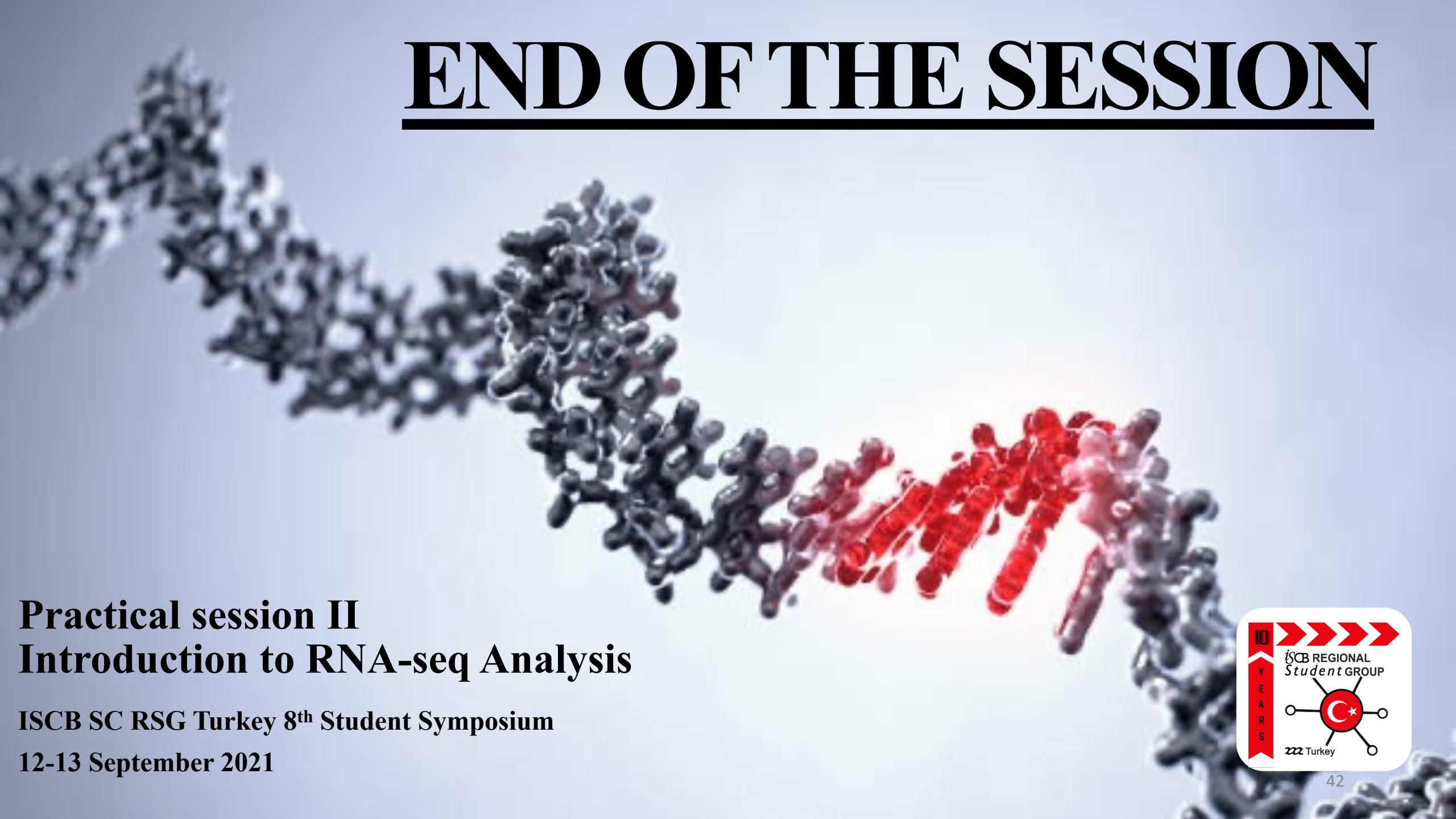
OPEN ACCESS

molecular systems biology

Current best practices in single-cell RNA-seq analysis: a tutorial

Malte D Luecken[1] & Fabian J Theis[1,2,*]

END OF THE SESSION



Practical session II
Introduction to RNA-seq Analysis

ISCB SC RSG Turkey 8th Student Symposium

12-13 September 2021

JOURNAL CLUB

ISCB RSG TURKEY

Our journal club initiative has been designed to encourage our participants to improve their presentation skills in front of the public and to create an academic friendly environment to discuss science together. It is open for anyone and free!

You can join now for 2022!.

For more information, please contact:
eravza@hotmail.com