Introduction to RNA-Seq

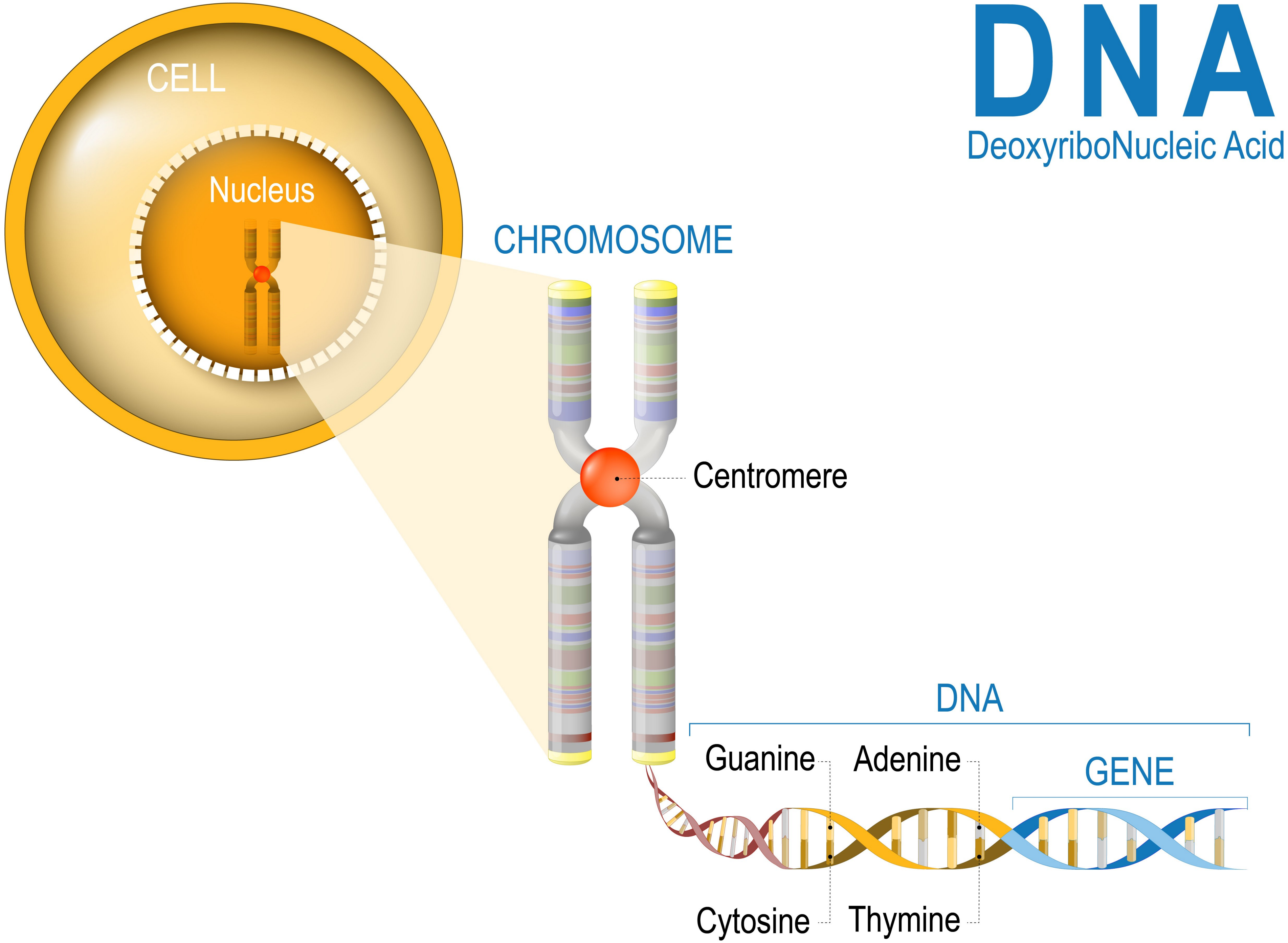
RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

datacamp

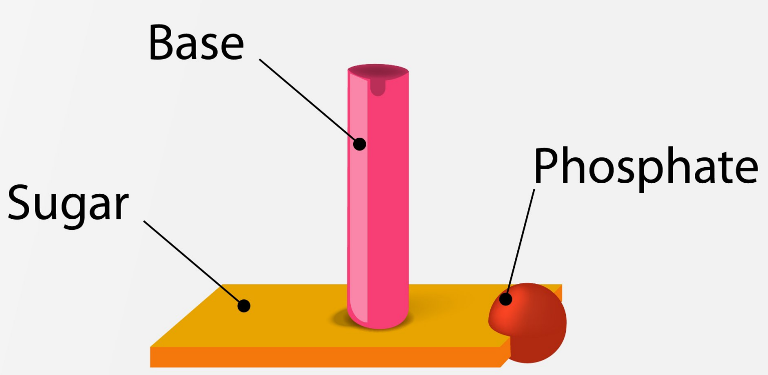
DNA

DeoxyriboNucleic Acid



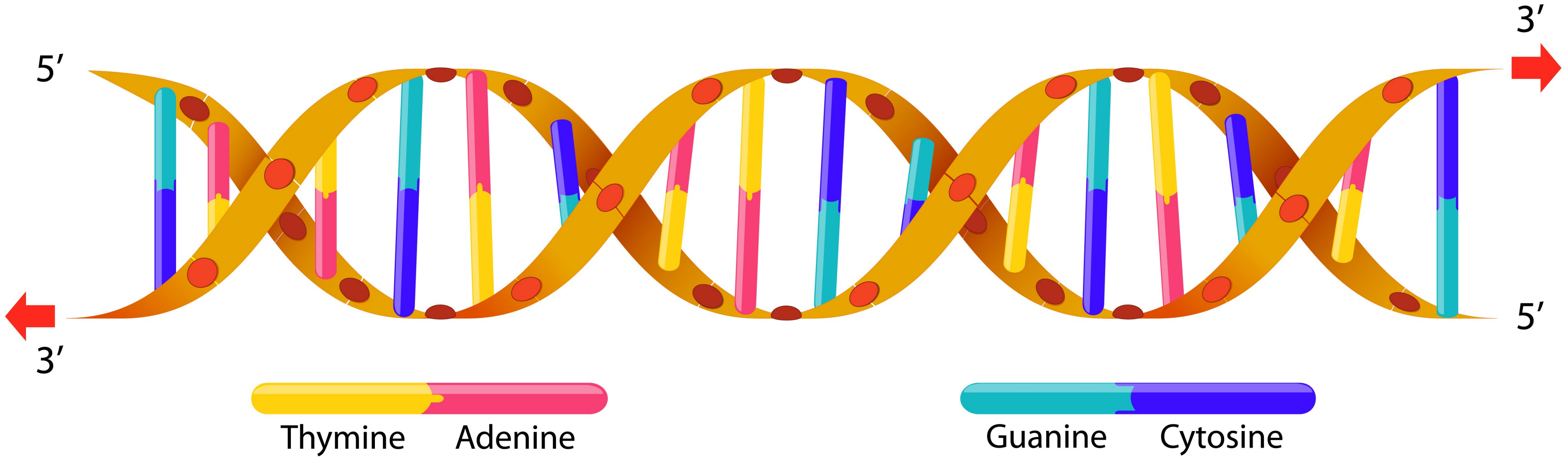
DNA structure

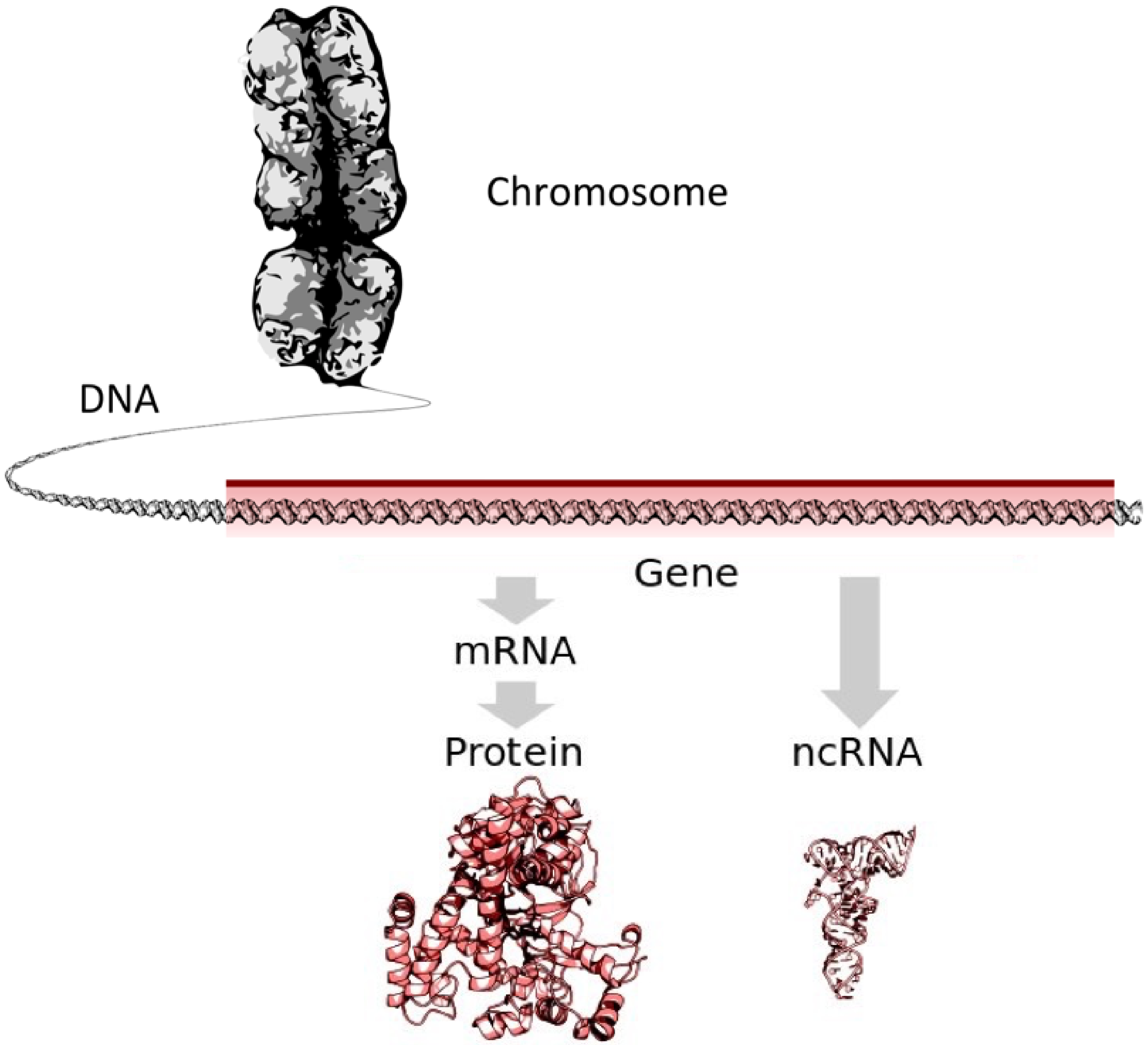
Nucleotide



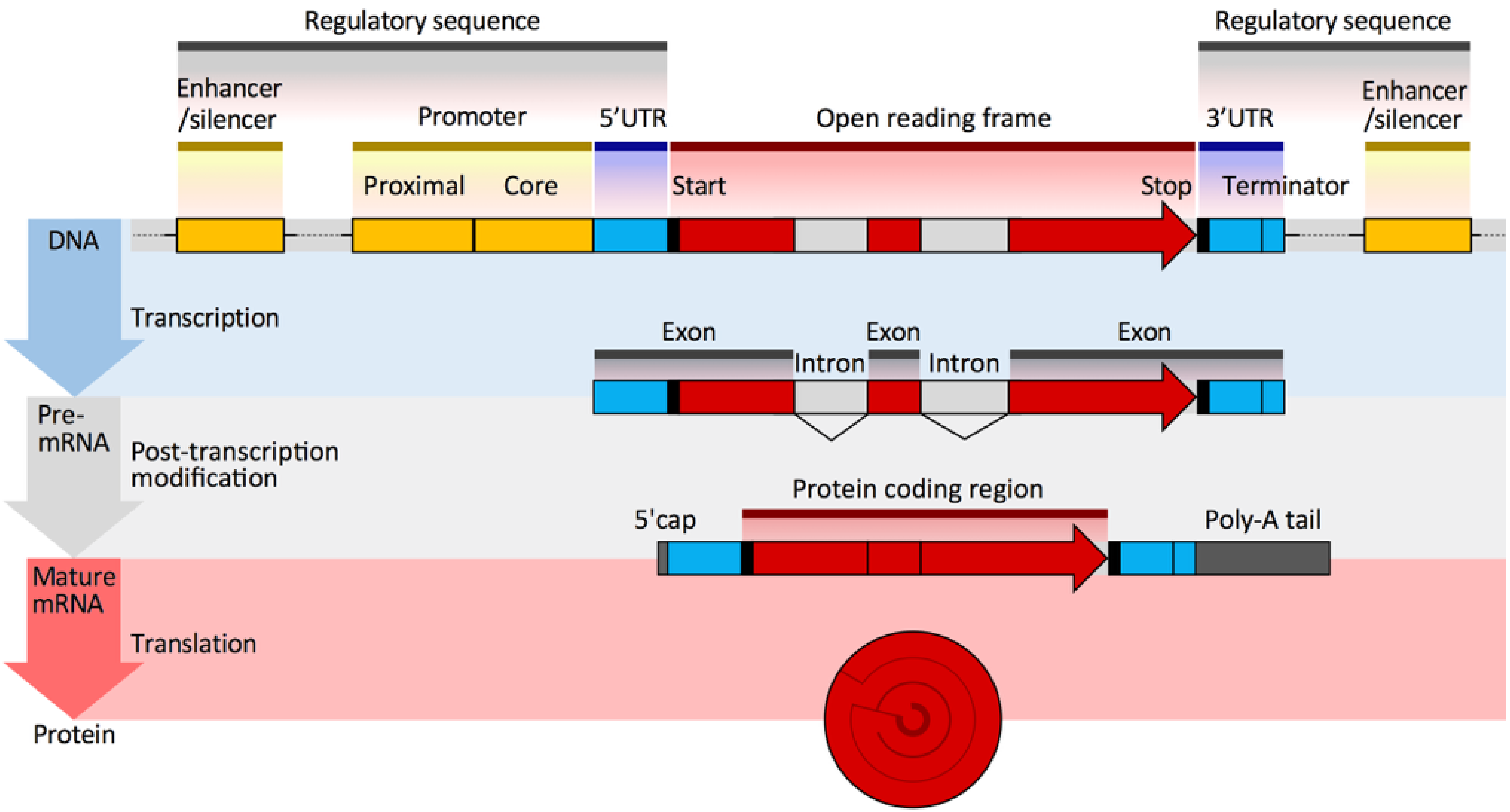
Bases

- 🟥 Adenine
- 🟦 Cytosine
- 🟩 Guanine
- 🟨 Thymine

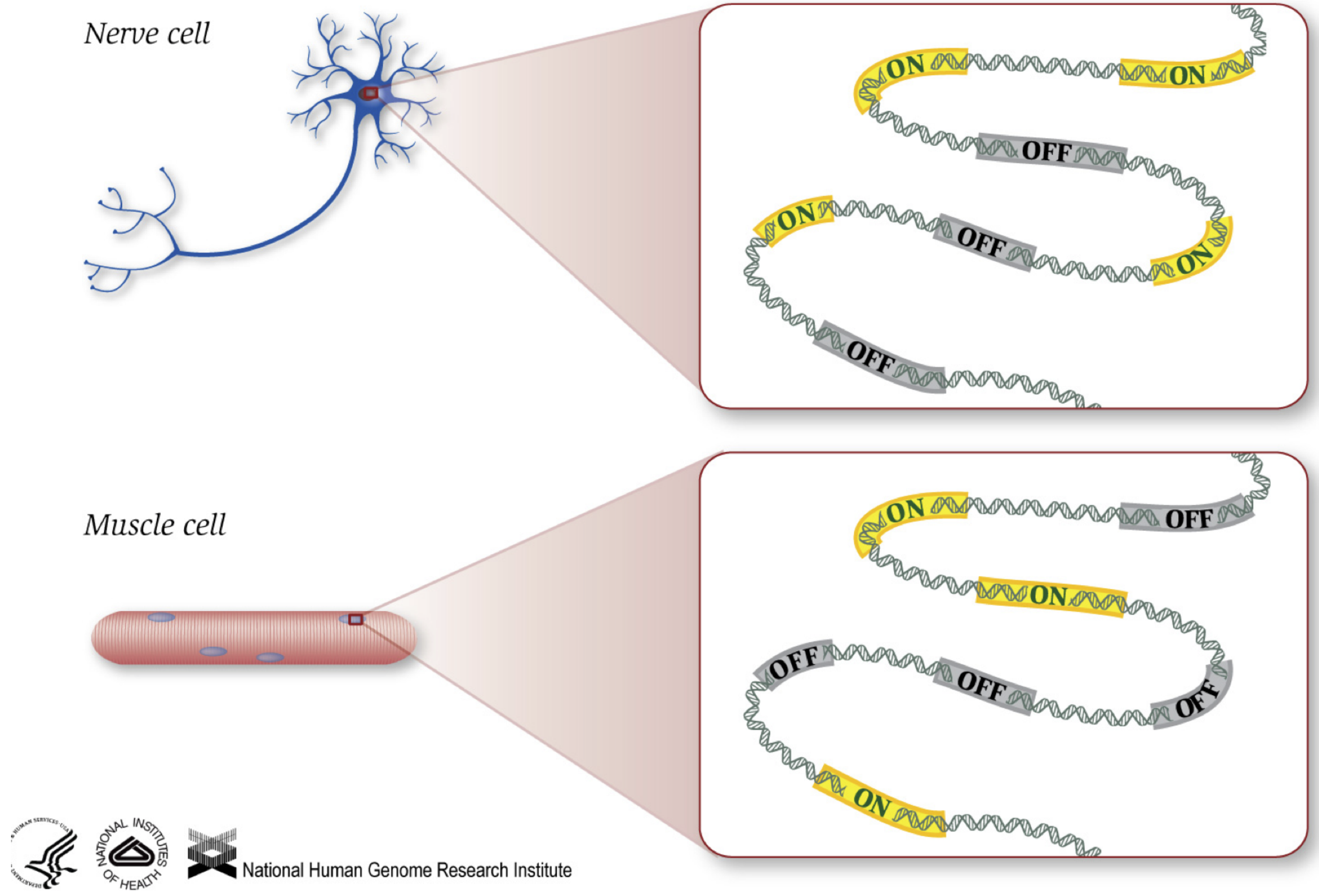


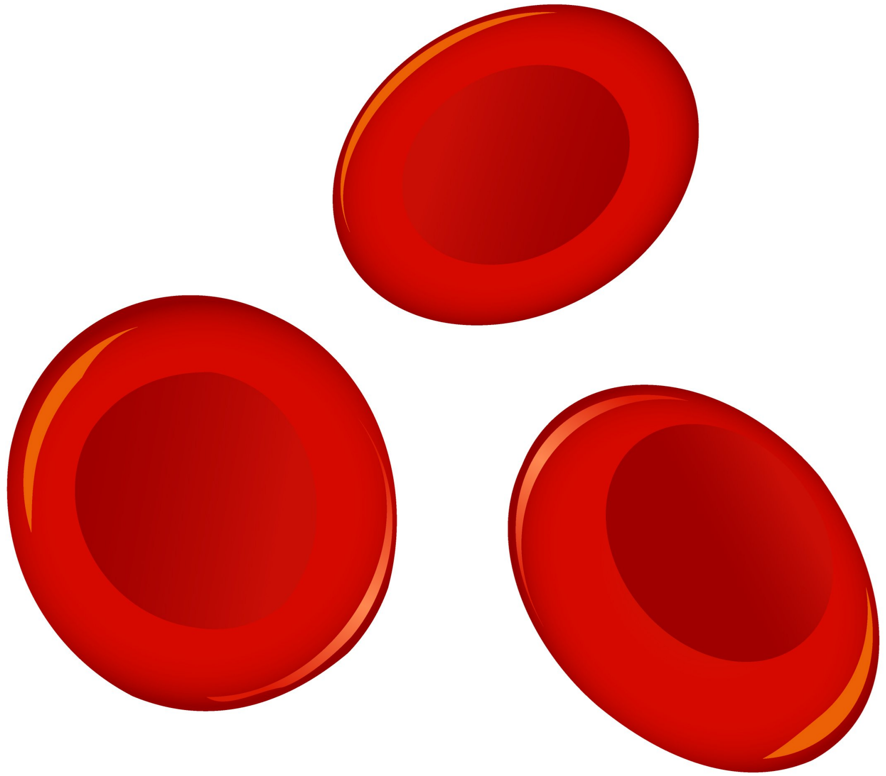


Wikimedia Commons Chromosome DNA Gene.svg and DNA to protein or ncRNA.svg" by Thomas Shafee, used under Creative Commons Attribution 4.0 International /
Combined originals and added highlights



Wikimedia Commons Gene structure eukaryote 2 annotated.svg by Thomas Shafee, used under Creative Commons Attribution 4.0 International


Nerve cell
ON
ON
OFF
ON
OFF
ON
OFF
Muscle cell
ON
OFF
ON
OFF
OFF
OFF
ON
National Human Genome Research Institute



Normal
Red Blood Cell

Sickled
Red Blood Cell

RNA-Seq questions

- What genes are differentially expressed between sample groups?
- Are there any trends in gene expression over time or across conditions.
- Which groups of genes change similarly over time or across conditions.
- What processes or pathways are important for my condition of interest?

Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R

RNA-Seq Workflow

RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

RNA-Seq Workflow: RNA-Seq Experimental Design

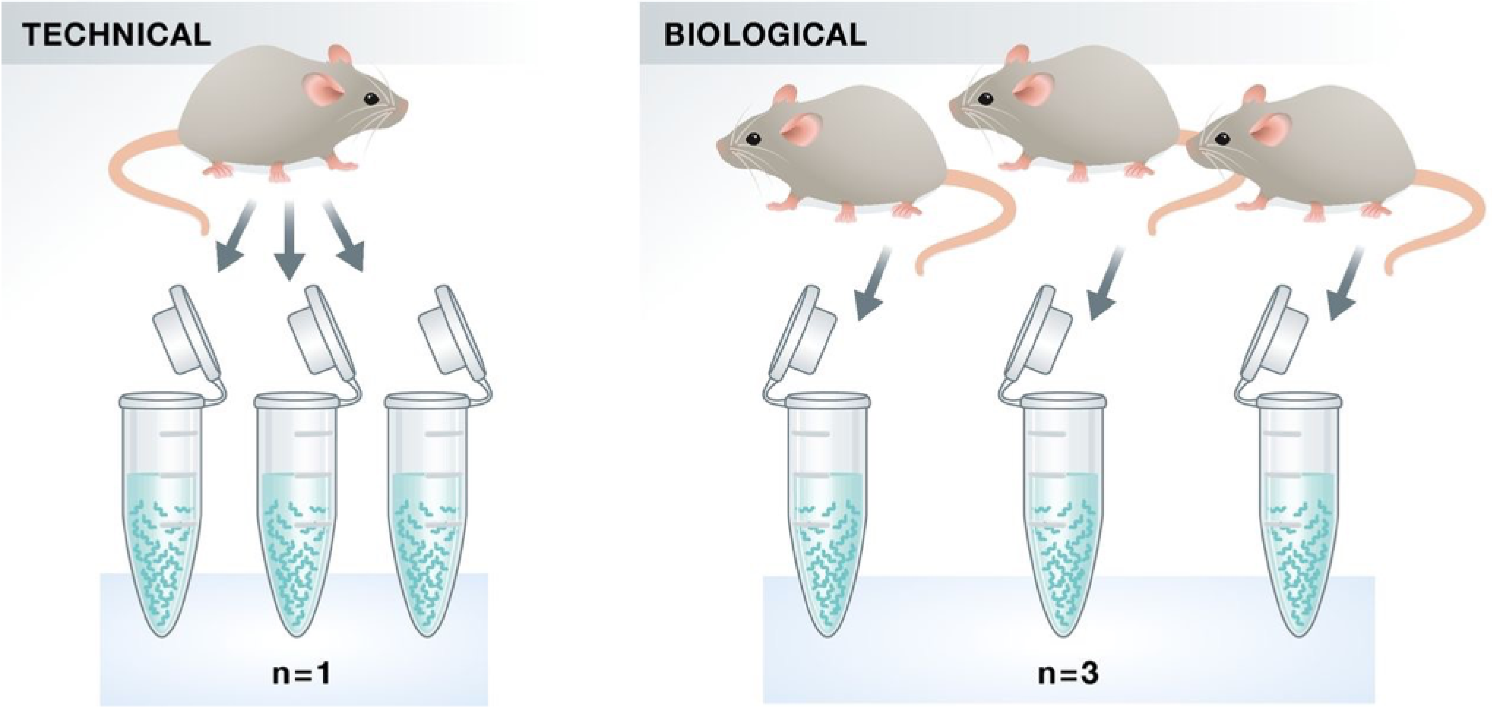
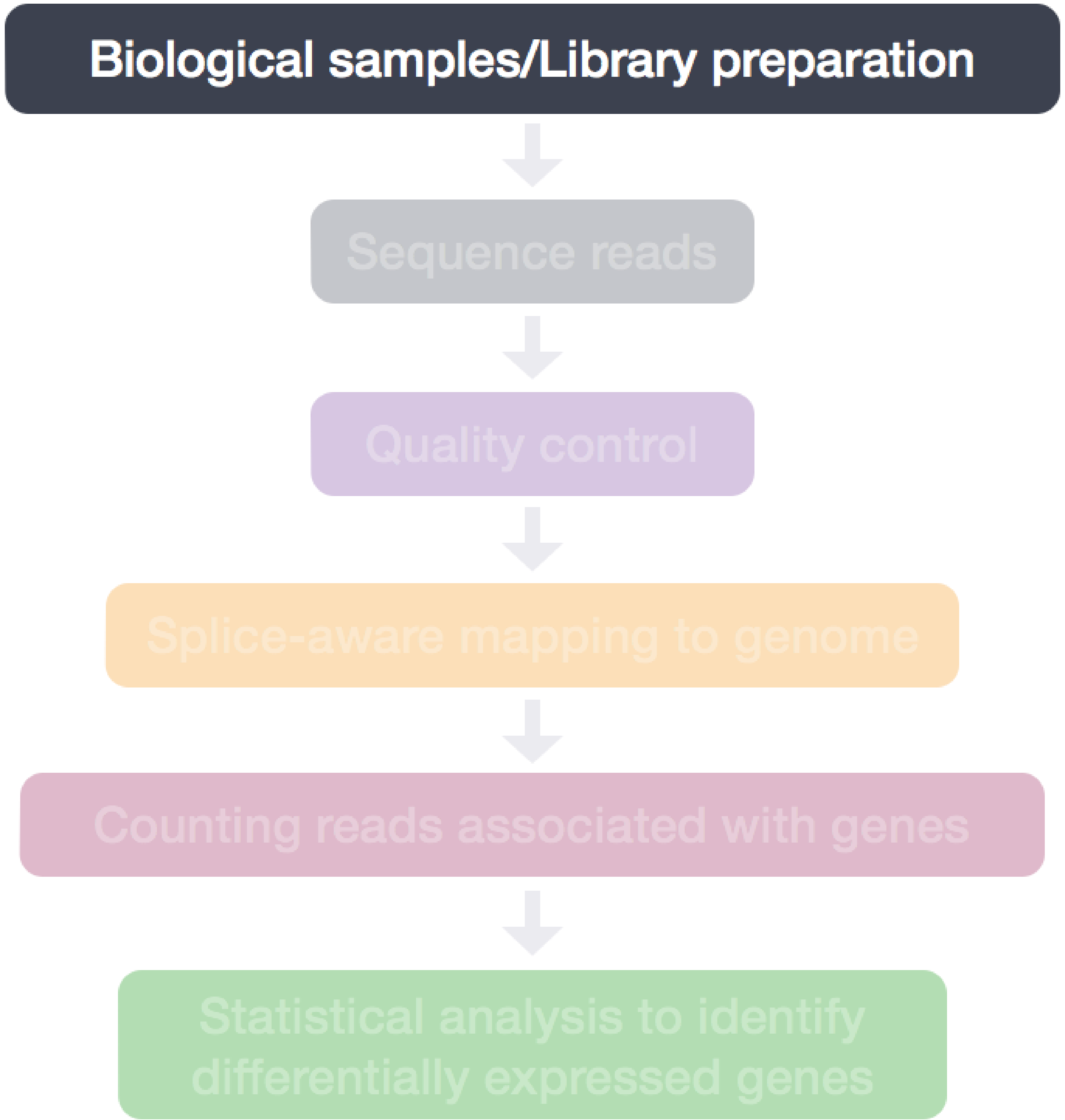


Image adapted from: Klaus B., EMBO J (2015) 34: 2727–2730

- **Technical replicates:** Generally low technical variation, so unnecessary.
- **Biological replicates:** Crucial to the success of RNA-Seq differential expression analyses. The more replicates the better, but at the very least have 3.
- **Batch effects:** Avoid as much as possible and note down all experimental variables.



Biological samples/Library preparation

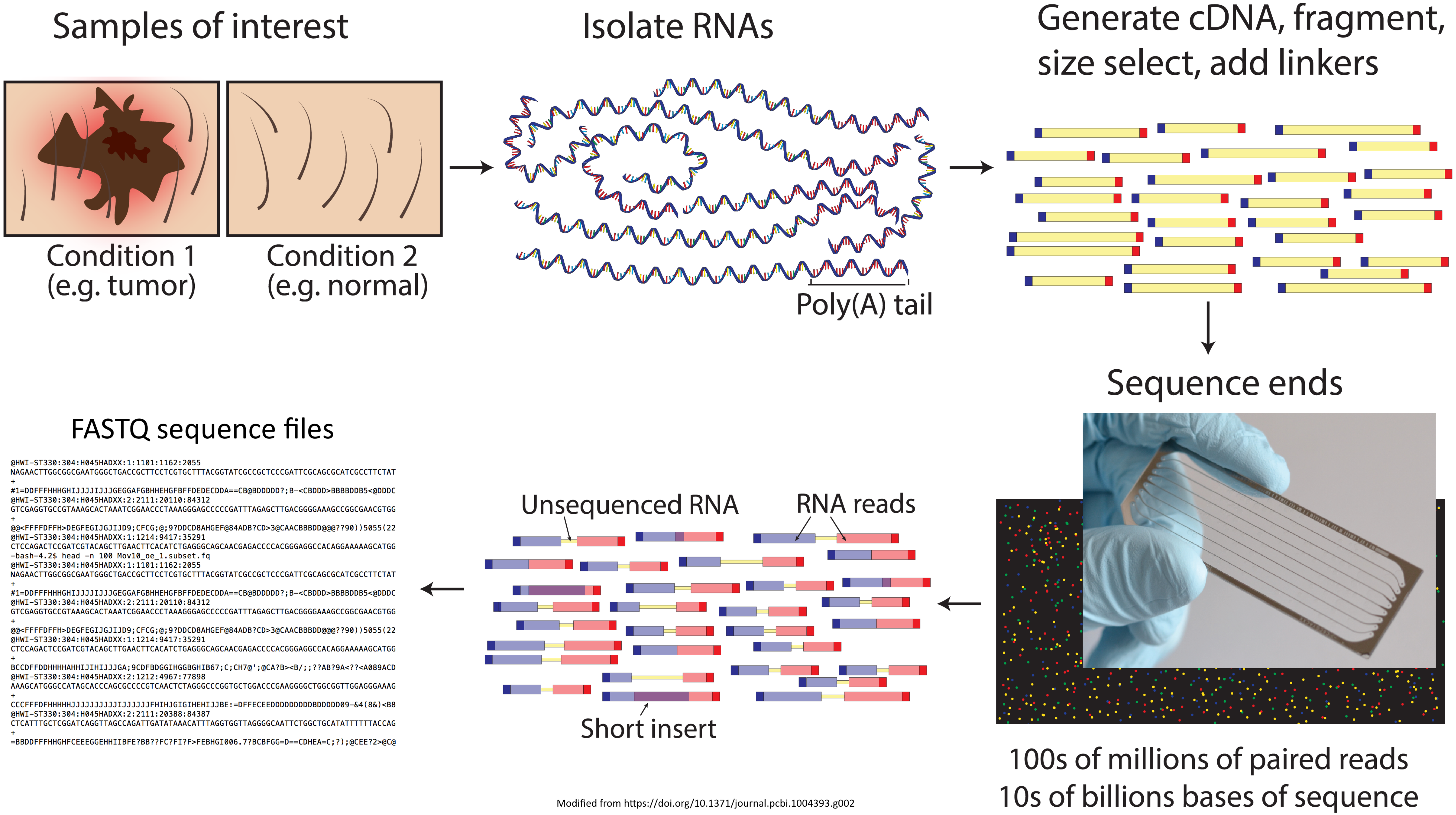
Sequence reads

Quality control

Splice-aware mapping to genome

Counting reads associated with genes

Statistical analysis to identify differentially expressed genes



Samples of interest

Condition 1
(e.g. tumor)

Condition 2
(e.g. normal)

Isolate RNAs

Poly(A) tail

Generate cDNA, fragment, size select, add linkers

Sequence ends

100s of millions of paired reads
10s of billions bases of sequence

Unsequenced RNA

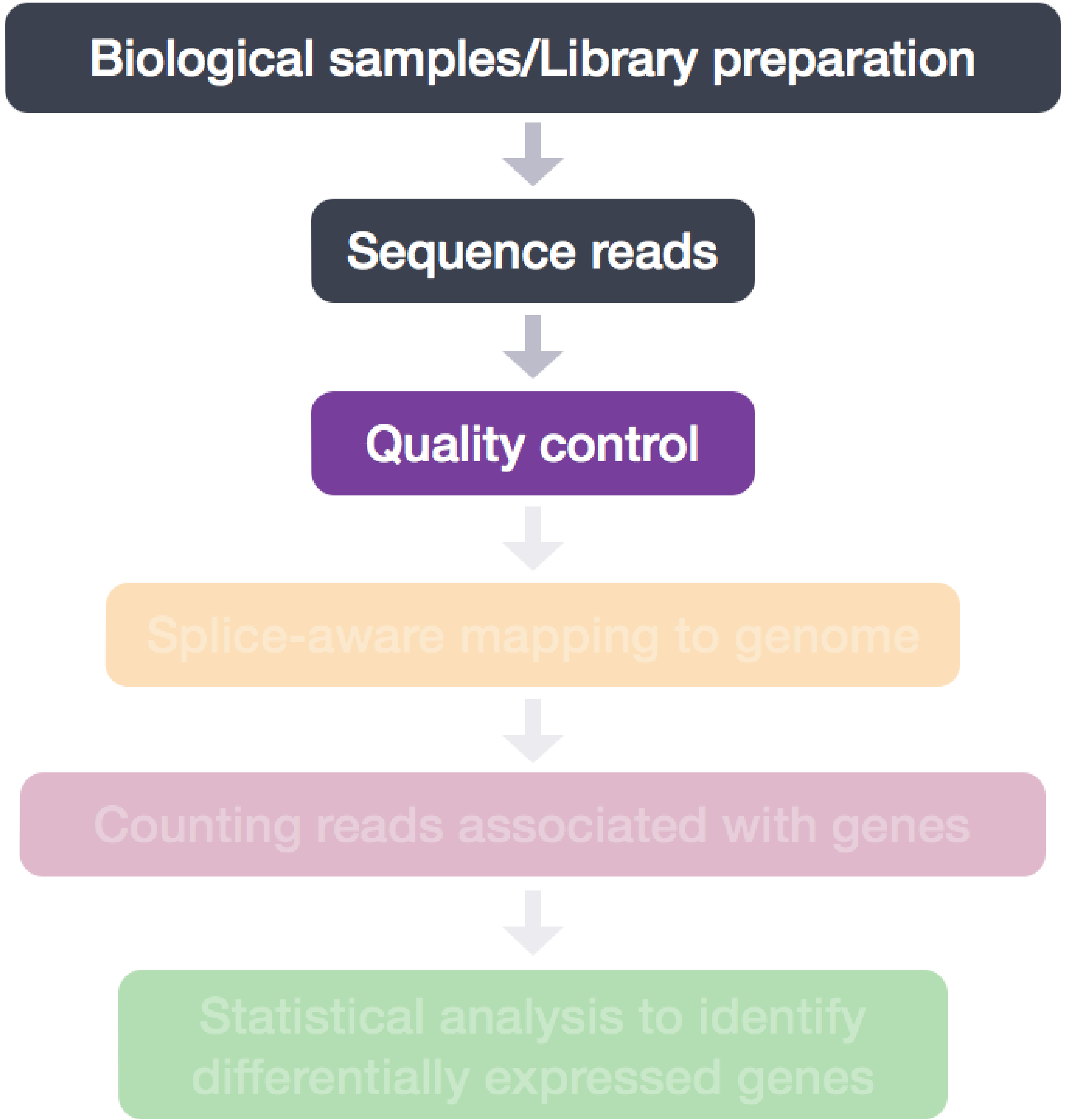
RNA reads

Short insert

FASTQ sequence files

```
@HWI-ST330:304:H045HADXX:1:1101:1162:2055
NAGAACTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTAT
+
#1=DDFFFHHHGHIJJJJIJJJGEGGAFGBHHEHGFBFFDEDECDDA==CB@BDDDDD?;B-<CBDDD>BBBBDDB5<@DDDC
@HWI-ST330:304:H045HADXX:2:2111:20110:84312
GTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGG
+
@@<FFFFDFFH>DEGFEGIJGJIJD9;CFCG;@;9?DDCD8AHGEF@84ADB?CD>3@CAACBBBDD@@@??90))5055(22
@HWI-ST330:304:H045HADXX:1:1214:9417:35291
CTCCAGACTCCGATCGTACAGCTTGAACTTCACATCTGAGGGCAGCAACGAGACCCCACGGGAGGCCACAGGAAAAAGCATGG
-bash-4.2$ head -n 100 Mov10_oe_1.subset.fq
@HWI-ST330:304:H045HADXX:1:1101:1162:2055
NAGAACTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTAT
+
#1=DDFFFHHHGHIJJJJIJJJGEGGAFGBHHEHGFBFFDEDECDDA==CB@BDDDDD?;B-<CBDDD>BBBBDDB5<@DDDC
@HWI-ST330:304:H045HADXX:2:2111:20110:84312
GTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGG
+
@@<FFFFDFFH>DEGFEGIJGJIJD9;CFCG;@;9?DDCD8AHGEF@84ADB?CD>3@CAACBBBDD@@@??90))5055(22
@HWI-ST330:304:H045HADXX:1:1214:9417:35291
CTCCAGACTCCGATCGTACAGCTTGAACTTCACATCTGAGGGCAGCAACGAGACCCCACGGGAGGCCACAGGAAAAAGCATGG
+
BCCDFFDDHHHHAHHIJIHIJJJGA;9CDFBDGGIHGGBGHIB67;C;CH7@';@CA?B><B/;;??AB?9A<??<A089ACD
@HWI-ST330:304:H045HADXX:2:1212:4967:77898
AAAGCATGGGCCATAGCACCCAGCGCCCCGTCAACTCTAGGGCCCGGTGCTGGACCCGAAGGGGCTGGCGGTTGGAGGGAAAG
+
CCCFFFDFHHHHHJJJJJJJJJJIJJJJJFHIHJGIGIHEHIJJBE:=DFFECEEDDDDDDDDDBDDDDD09-&4(8&)<B8
@HWI-ST330:304:H045HADXX:2:2111:20388:84387
CTCATTTGCTCGGATCAGGTTAGCCAGATTGATATAAACATTTAGGTGGTTAGGGGCAATTCTGGCTGCATATTTTTTACCAG
+
=BBDDFFFHHGHFCEEEGGEHHIIBFE?BB??FC?FI?F>FEBHGI006.7?BCBFGG=D==CDHEA=C;?);@CEE?2>@C@
```

Modified from https://doi.org/10.1371/journal.pcbi.1004393.g002



Biological samples/Library preparation

Sequence reads

Quality control

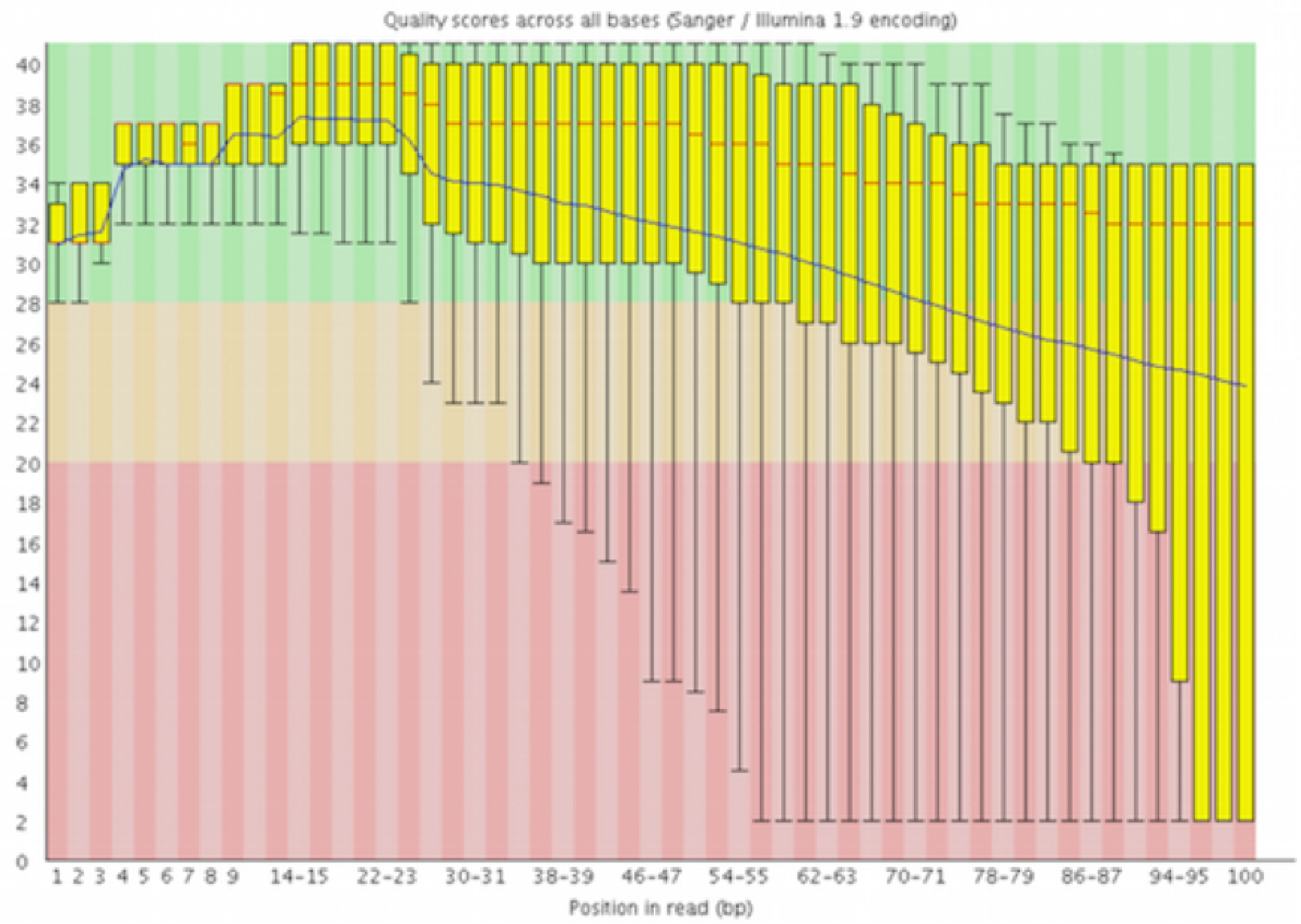
Splice-aware mapping to genome

Counting reads associated with genes

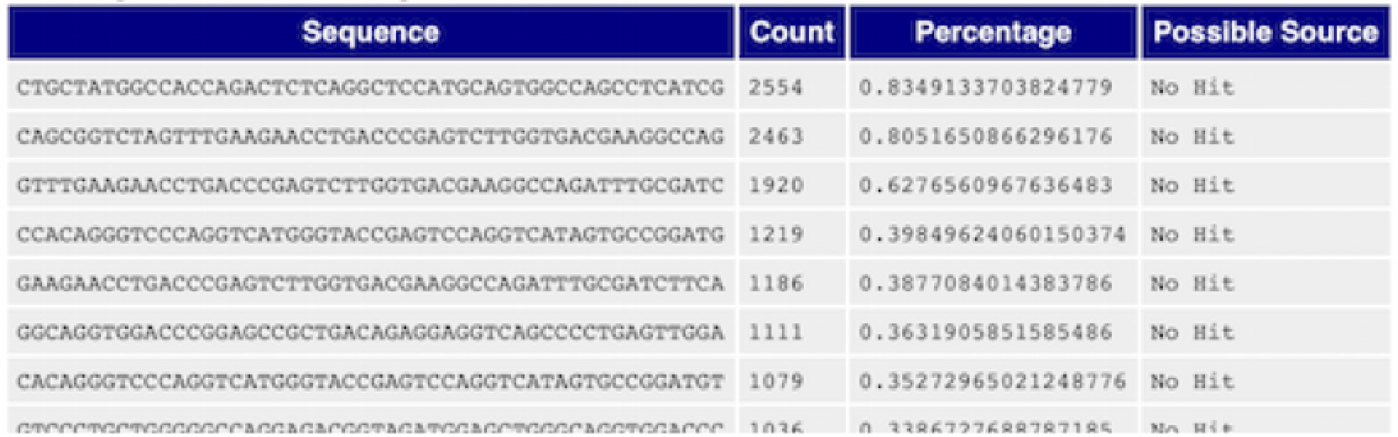
Statistical analysis to identify differentially expressed genes

RNA-Seq Workflow: Quality control

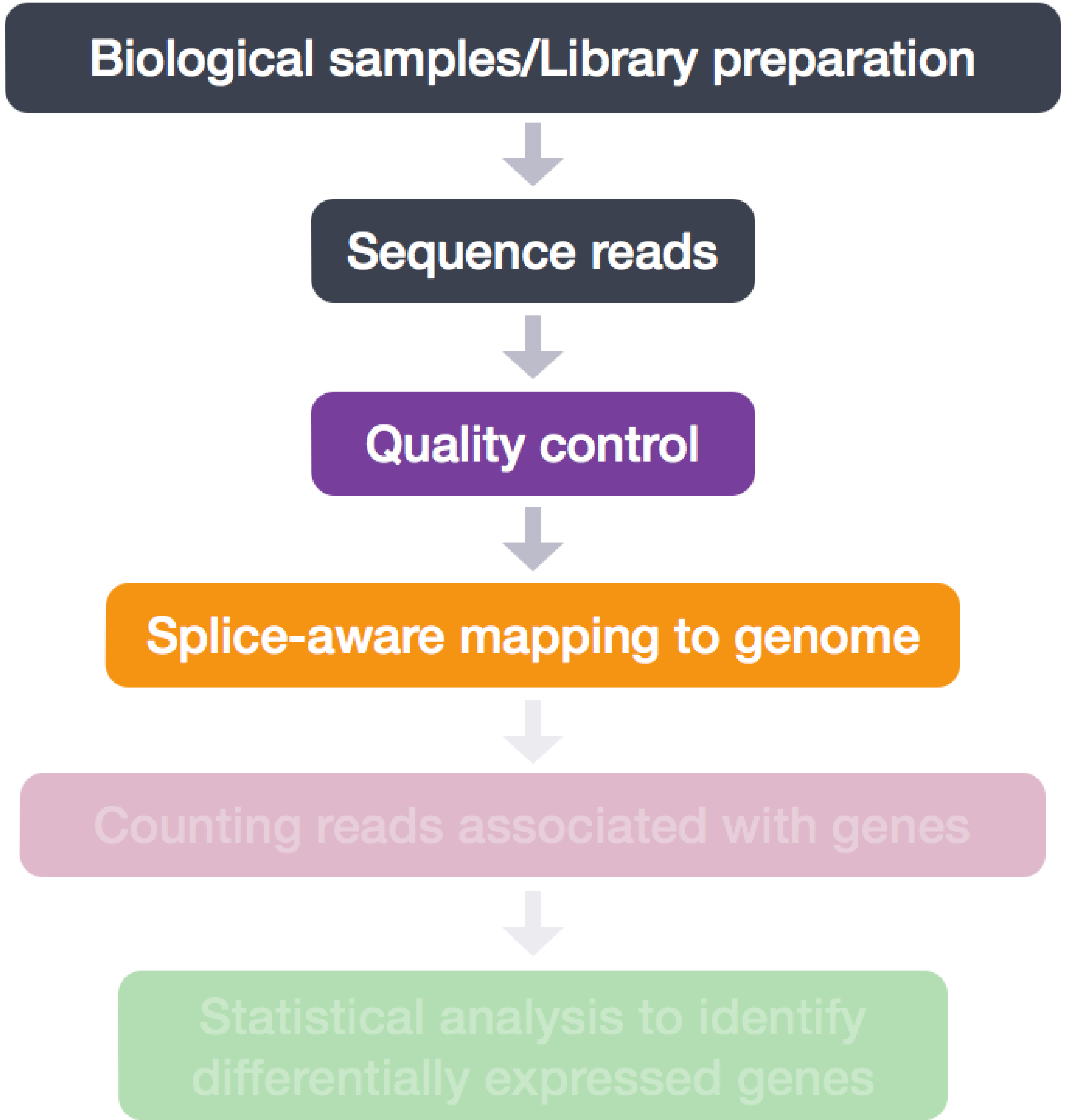
Per base sequence quality



Overrepresented sequences



Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGGAGCCGCTGACAGAGGAGGTCAGCCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATGT	1079	0.35272965021248776	No Hit
GTCCCTGCTGGGGGCCAGGAGACGGTAGATGGAGCTGGGCAGGTGGACCC	1036	0.3386727688787185	No Hit


Biological samples/Library preparation
Sequence reads
Quality control
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

RNA-Seq Workflow: Alignment

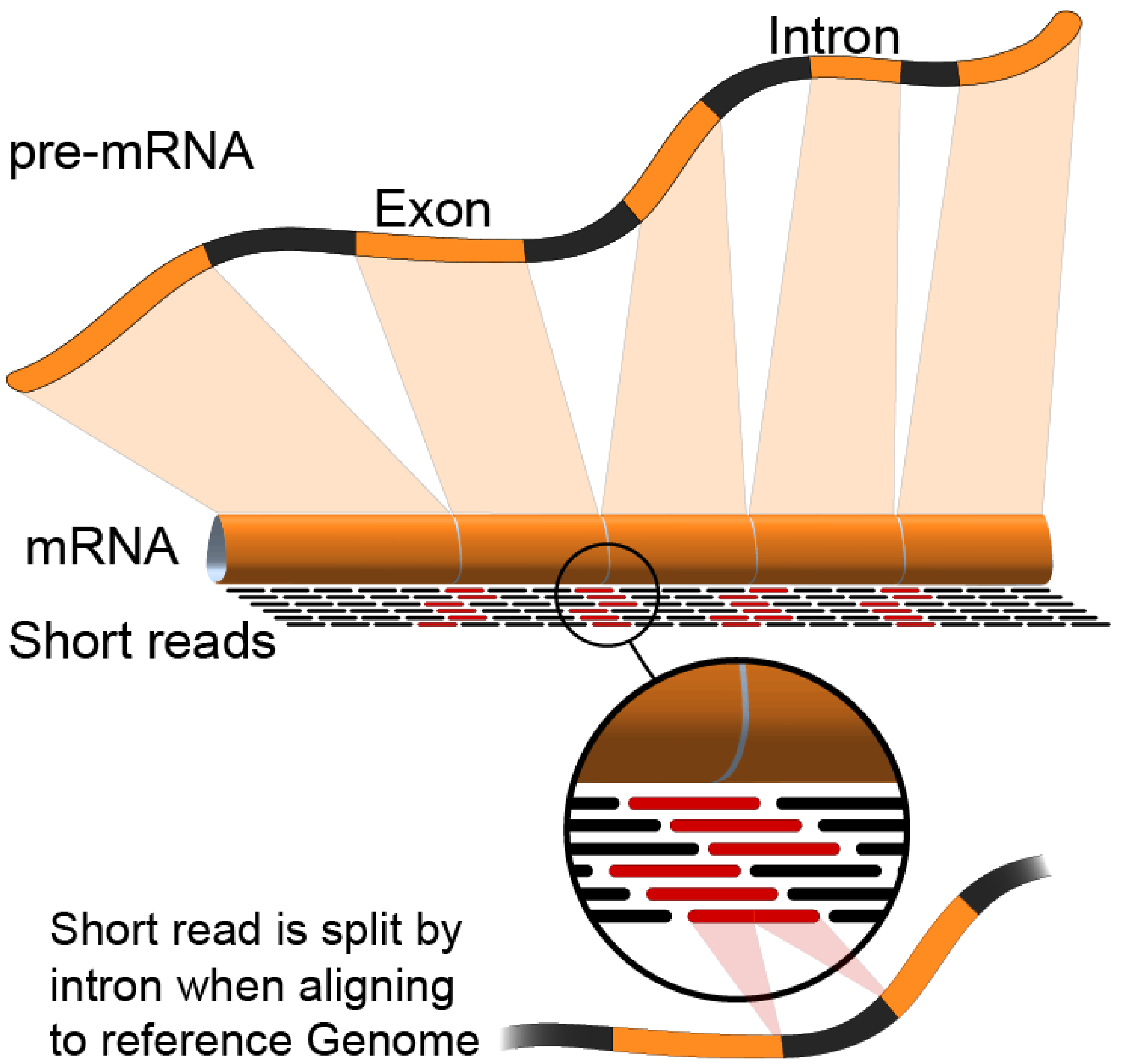
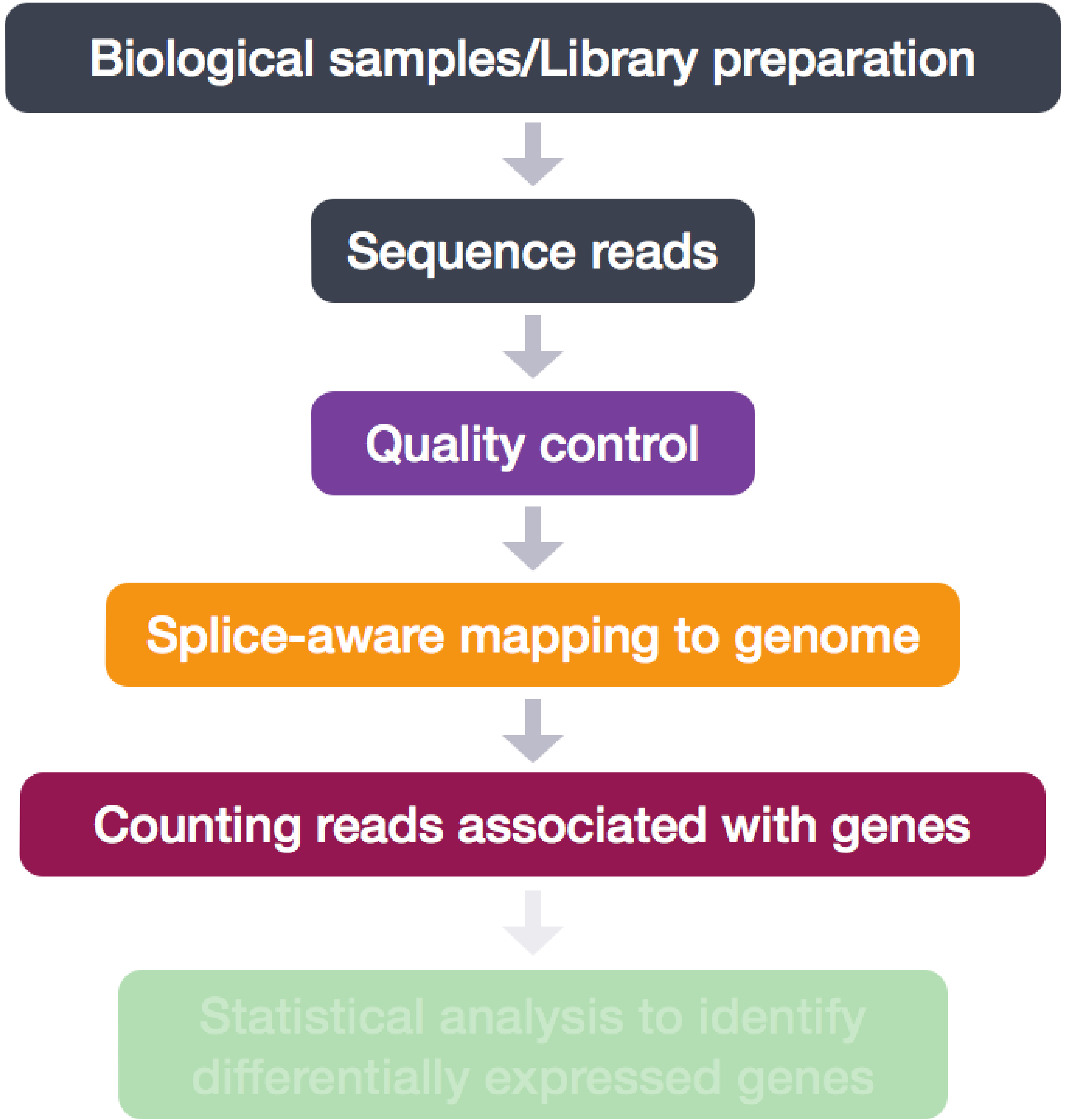

Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Image credit: Wikimedia commons RNA-Seq-alignment.png



Biological samples/Library preparation

Sequence reads

Quality control

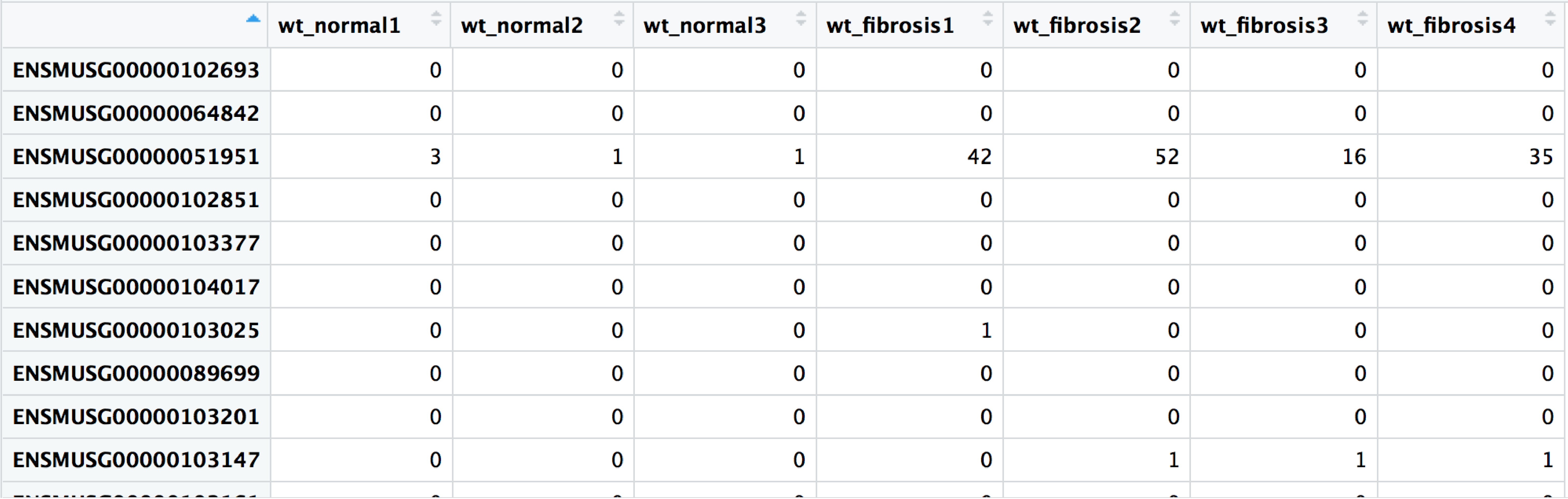
Splice-aware mapping to genome

Counting reads associated with genes

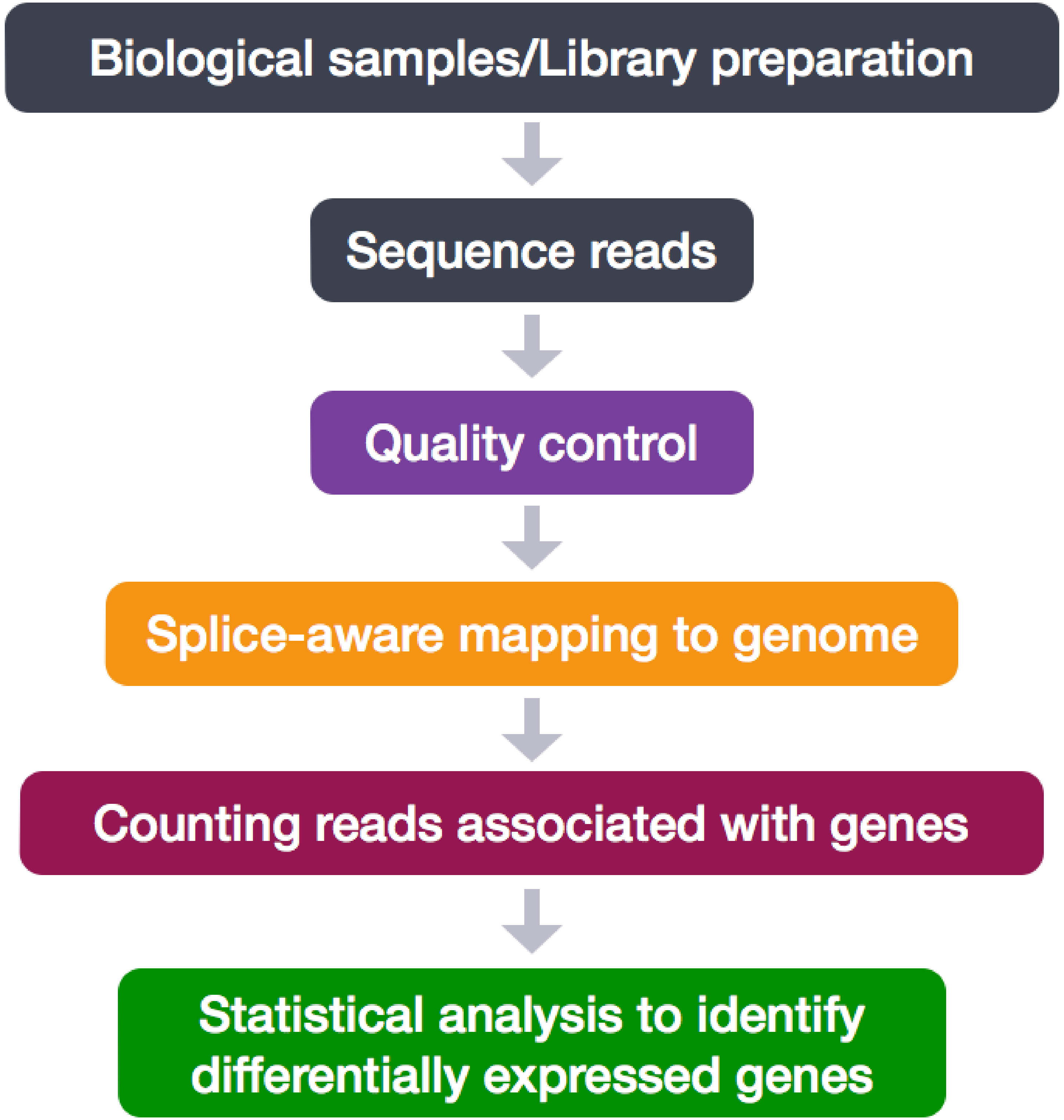
Statistical analysis to identify differentially expressed genes

RNA-Seq Workflow: Count matrix

```
wt_rawcounts <- read.csv("fibrosis_wt_rawcounts.csv")
```



	wt_normal1	wt_normal2	wt_normal3	wt_fibrosis1	wt_fibrosis2	wt_fibrosis3	wt_fibrosis4
ENSMUSG00000102693	0	0	0	0	0	0	0
ENSMUSG00000064842	0	0	0	0	0	0	0
ENSMUSG00000051951	3	1	1	42	52	16	35
ENSMUSG00000102851	0	0	0	0	0	0	0
ENSMUSG00000103377	0	0	0	0	0	0	0
ENSMUSG00000104017	0	0	0	0	0	0	0
ENSMUSG00000103025	0	0	0	1	0	0	0
ENSMUSG00000089699	0	0	0	0	0	0	0
ENSMUSG00000103201	0	0	0	0	0	0	0
ENSMUSG00000103147	0	0	0	0	1	1	1



Biological samples/Library preparation

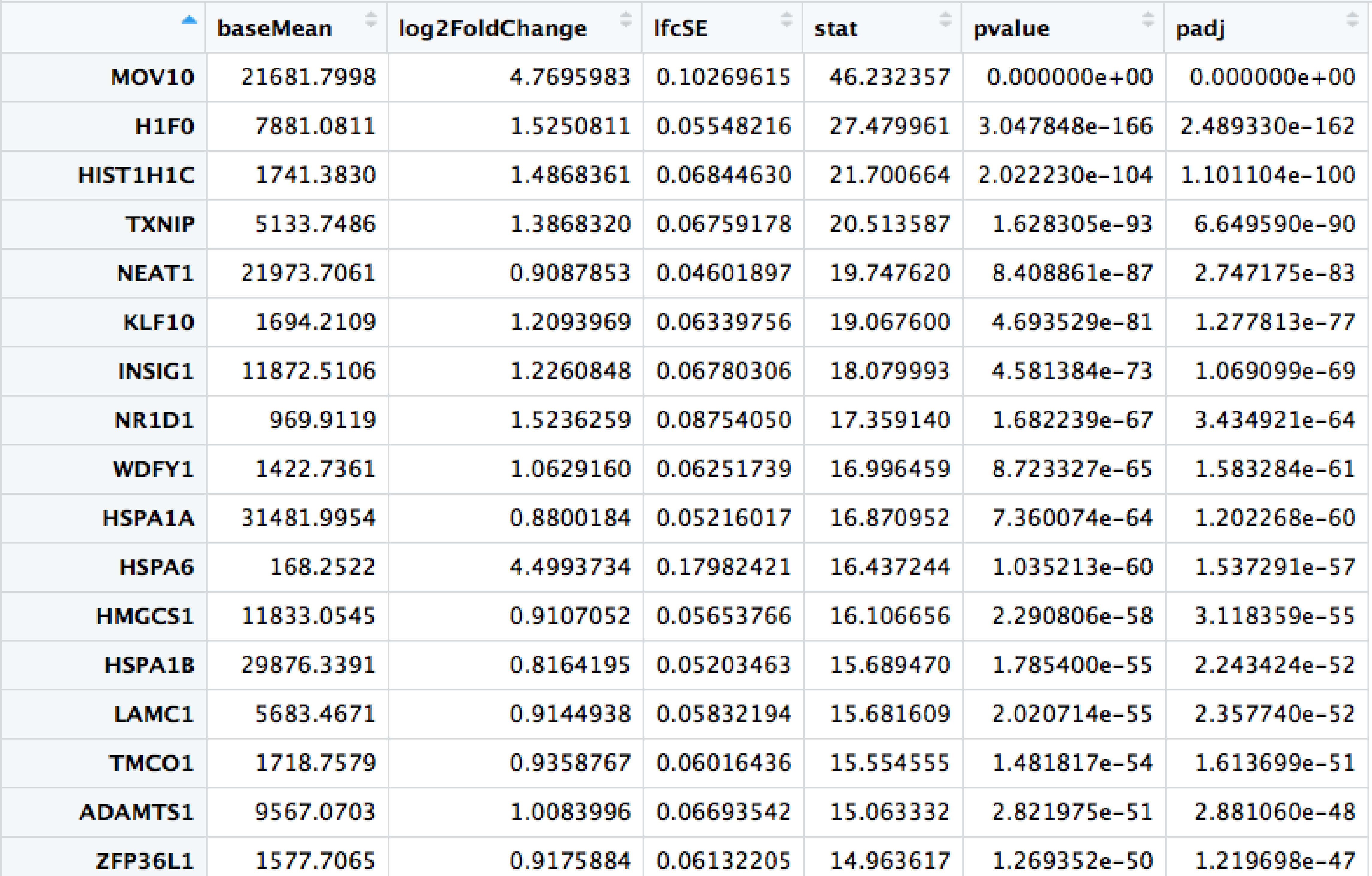
Sequence reads

Quality control

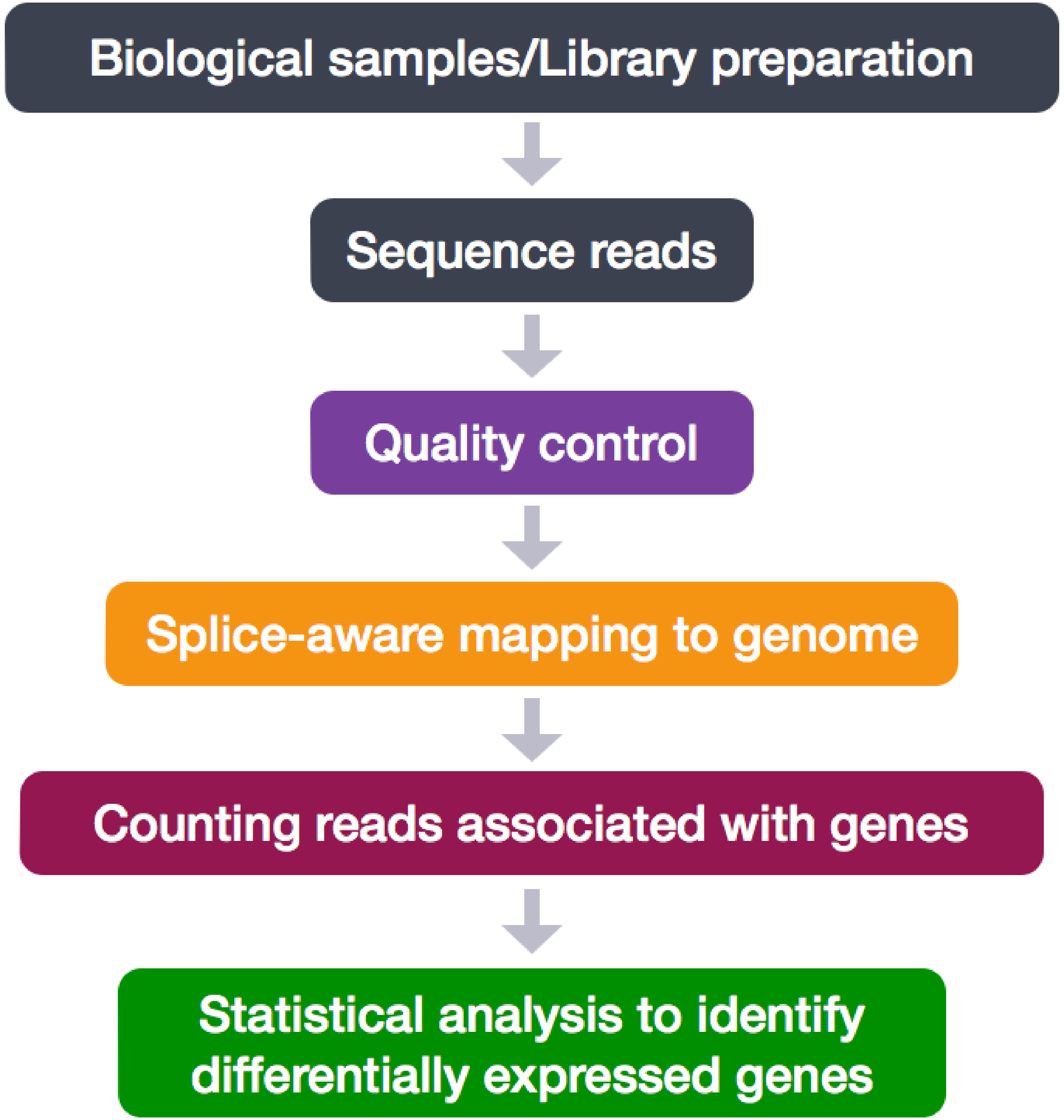
Splice-aware mapping to genome

Counting reads associated with genes

Statistical analysis to identify differentially expressed genes



	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
MOV10	21681.7998	4.7695983	0.10269615	46.232357	0.000000e+00	0.000000e+00
H1F0	7881.0811	1.5250811	0.05548216	27.479961	3.047848e-166	2.489330e-162
HIST1H1C	1741.3830	1.4868361	0.06844630	21.700664	2.022230e-104	1.101104e-100
TXNIP	5133.7486	1.3868320	0.06759178	20.513587	1.628305e-93	6.649590e-90
NEAT1	21973.7061	0.9087853	0.04601897	19.747620	8.408861e-87	2.747175e-83
KLF10	1694.2109	1.2093969	0.06339756	19.067600	4.693529e-81	1.277813e-77
INSIG1	11872.5106	1.2260848	0.06780306	18.079993	4.581384e-73	1.069099e-69
NR1D1	969.9119	1.5236259	0.08754050	17.359140	1.682239e-67	3.434921e-64
WDFY1	1422.7361	1.0629160	0.06251739	16.996459	8.723327e-65	1.583284e-61
HSPA1A	31481.9954	0.8800184	0.05216017	16.870952	7.360074e-64	1.202268e-60
HSPA6	168.2522	4.4993734	0.17982421	16.437244	1.035213e-60	1.537291e-57
HMGCS1	11833.0545	0.9107052	0.05653766	16.106656	2.290806e-58	3.118359e-55
HSPA1B	29876.3391	0.8164195	0.05203463	15.689470	1.785400e-55	2.243424e-52
LAMC1	5683.4671	0.9144938	0.05832194	15.681609	2.020714e-55	2.357740e-52
TMCO1	1718.7579	0.9358767	0.06016436	15.554555	1.481817e-54	1.613699e-51
ADAMTS1	9567.0703	1.0083996	0.06693542	15.063332	2.821975e-51	2.881060e-48
ZFP36L1	1577.7065	0.9175884	0.06132205	14.963617	1.269352e-50	1.219698e-47



Biological samples/Library preparation

Sequence reads

Quality control

Splice-aware mapping to genome

Counting reads associated with genes

Statistical analysis to identify differentially expressed genes

Back to you!

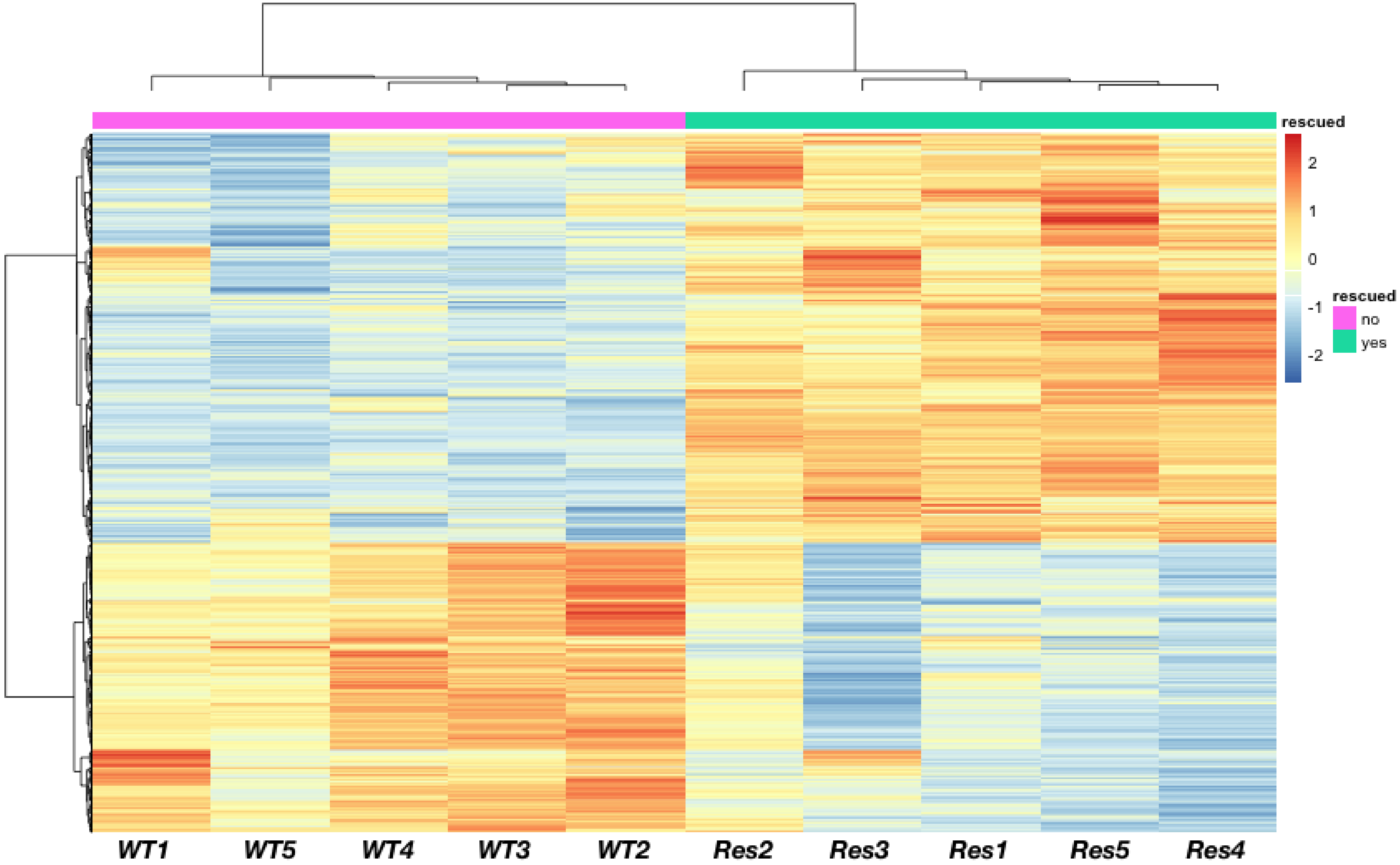
RNA-SEQ WITH BIOCONDUCTOR IN R

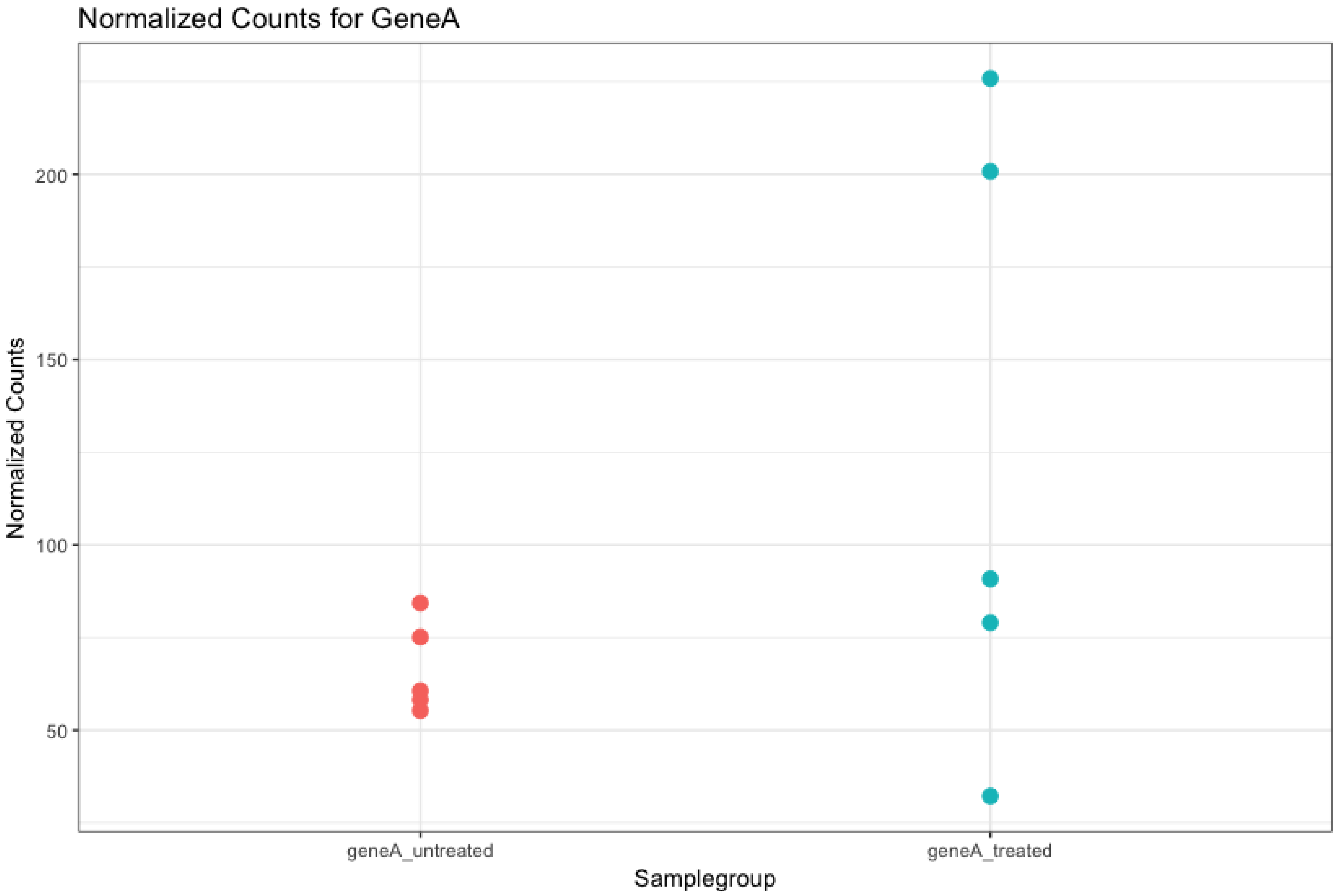
Differential gene expression overview

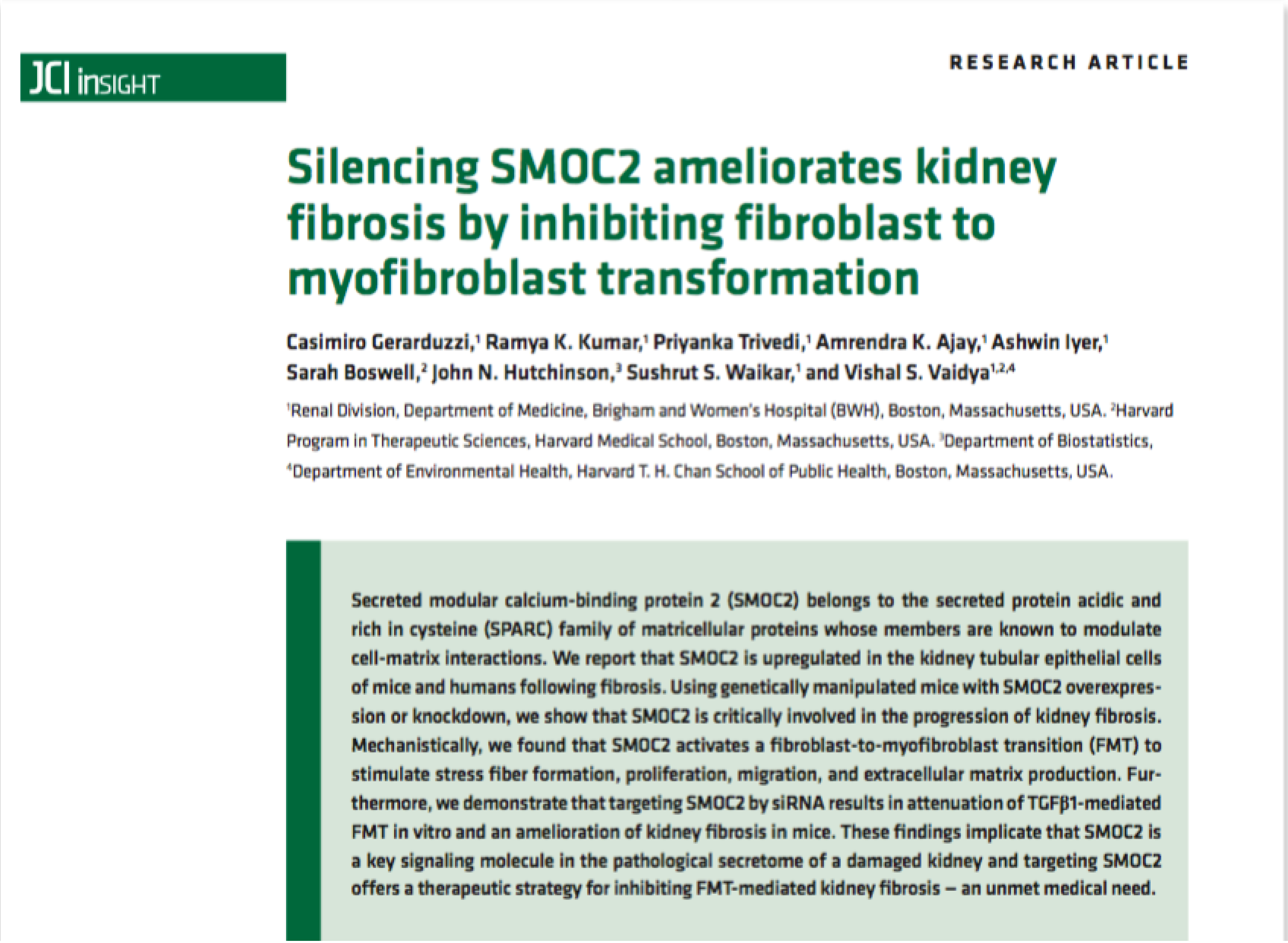
RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper

Bioinformatics Consultant and Trainer


rescued
2
1
0
-1
-2
rescued
🟪 no
🟩 yes
WT1
WT5
WT4
WT3
WT2
Res2
Res3
Res1
Res5
Res4


Normalized Counts for GeneA
Normalized Counts
200
150
100
50
geneA_untreated
geneA_treated
Samplegroup



JCI insight

RESEARCH ARTICLE

Silencing SMOC2 ameliorates kidney fibrosis by inhibiting fibroblast to myofibroblast transformation

Casimiro Gerarduzzi,[1] Ramya K. Kumar,[1] Priyanka Trivedi,[1] Amrendra K. Ajay,[1] Ashwin Iyer,[1] Sarah Boswell,[2] John N. Hutchinson,[3] Sushrut S. Waikar,[1] and Vishal S. Vaidya[1,2,4]

[1]Renal Division, Department of Medicine, Brigham and Women's Hospital (BWH), Boston, Massachusetts, USA. [2]Harvard Program in Therapeutic Sciences, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Biostatistics, [4]Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA.

Secreted modular calcium-binding protein 2 (SMOC2) belongs to the secreted protein acidic and rich in cysteine (SPARC) family of matricellular proteins whose members are known to modulate cell-matrix interactions. We report that SMOC2 is upregulated in the kidney tubular epithelial cells of mice and humans following fibrosis. Using genetically manipulated mice with SMOC2 overexpression or knockdown, we show that SMOC2 is critically involved in the progression of kidney fibrosis. Mechanistically, we found that SMOC2 activates a fibroblast-to-myofibroblast transition (FMT) to stimulate stress fiber formation, proliferation, migration, and extracellular matrix production. Furthermore, we demonstrate that targeting SMOC2 by siRNA results in attenuation of TGFβ1-mediated FMT in vitro and an amelioration of kidney fibrosis in mice. These findings implicate that SMOC2 is a key signaling molecule in the pathological secretome of a damaged kidney and targeting SMOC2 offers a therapeutic strategy for inhibiting FMT-mediated kidney fibrosis – an unmet medical need.

Introduction to dataset: Smoc2

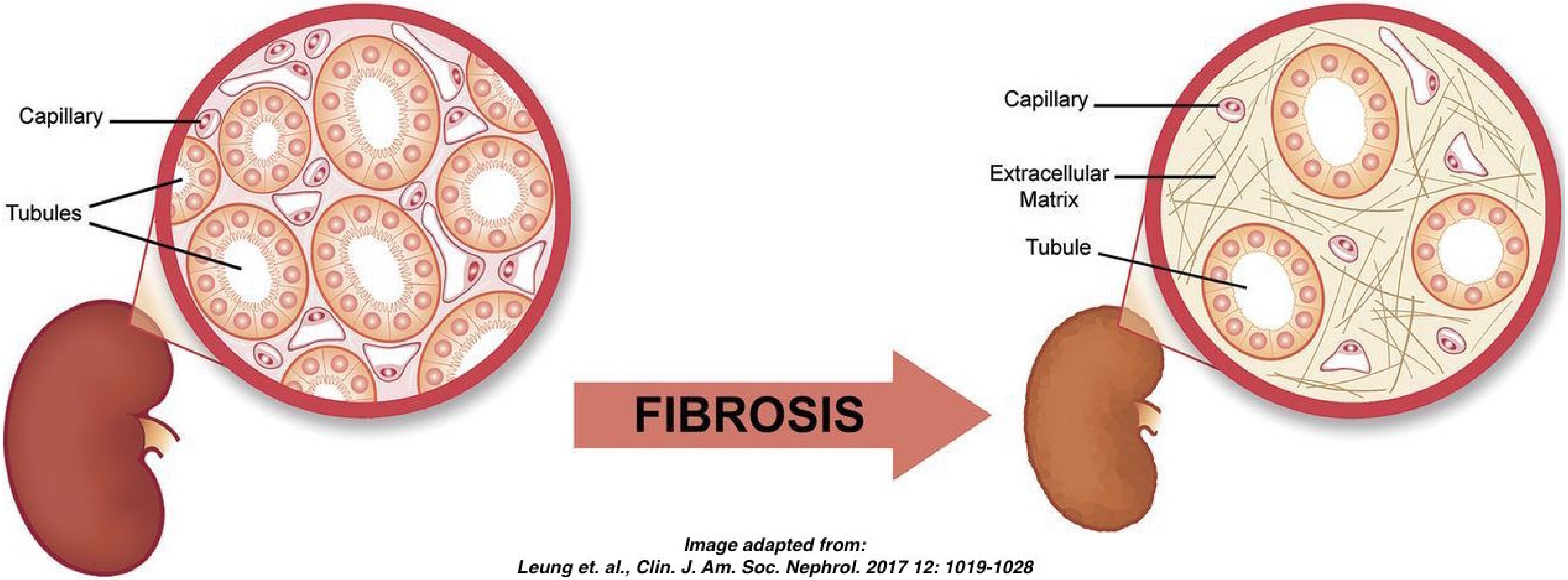


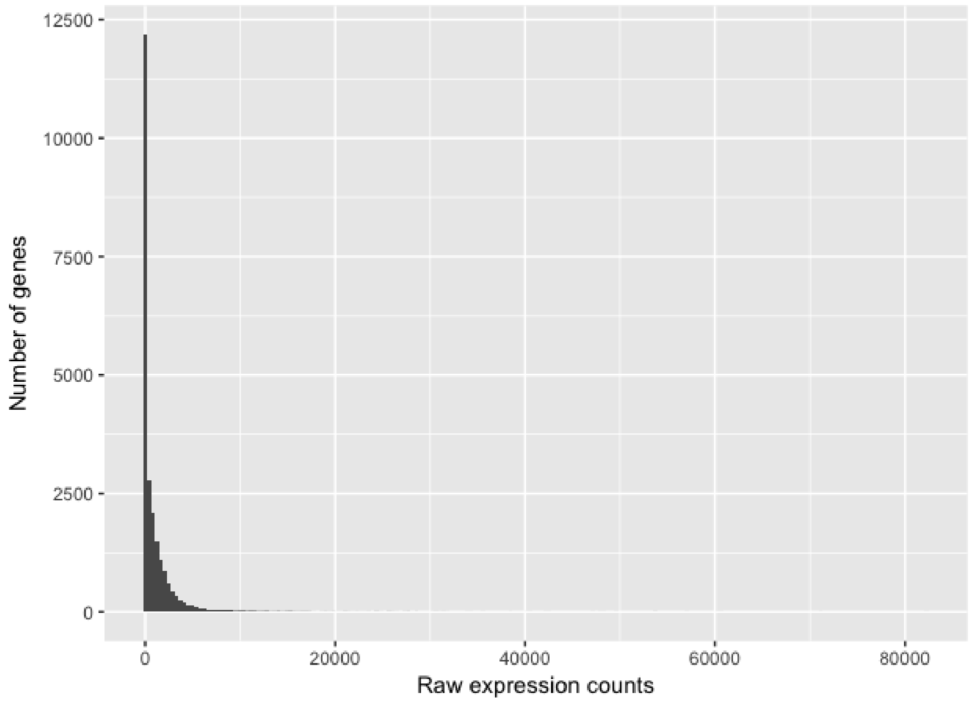
Image adapted from:
Leung et. al., Clin. J. Am. Soc. Nephrol. 2017 12: 1019-1028

Smoc2 Expression

Fibrosis Condition		*Wild type*	*Smoc2 over-expression*
	Normal	WT normal1	Smoc2 normal1
		WT normal2	Smoc2 normal2
		WT normal3	Smoc2 normal4
	Fibrosis	WT fibrosis1	Smoc2 fibrosis1
		WT fibrosis2	Smoc2 fibrosis2
		WT fibrosis3	Smoc2 fibrosis3
		WT fibrosis4	Smoc2 fibrosis4

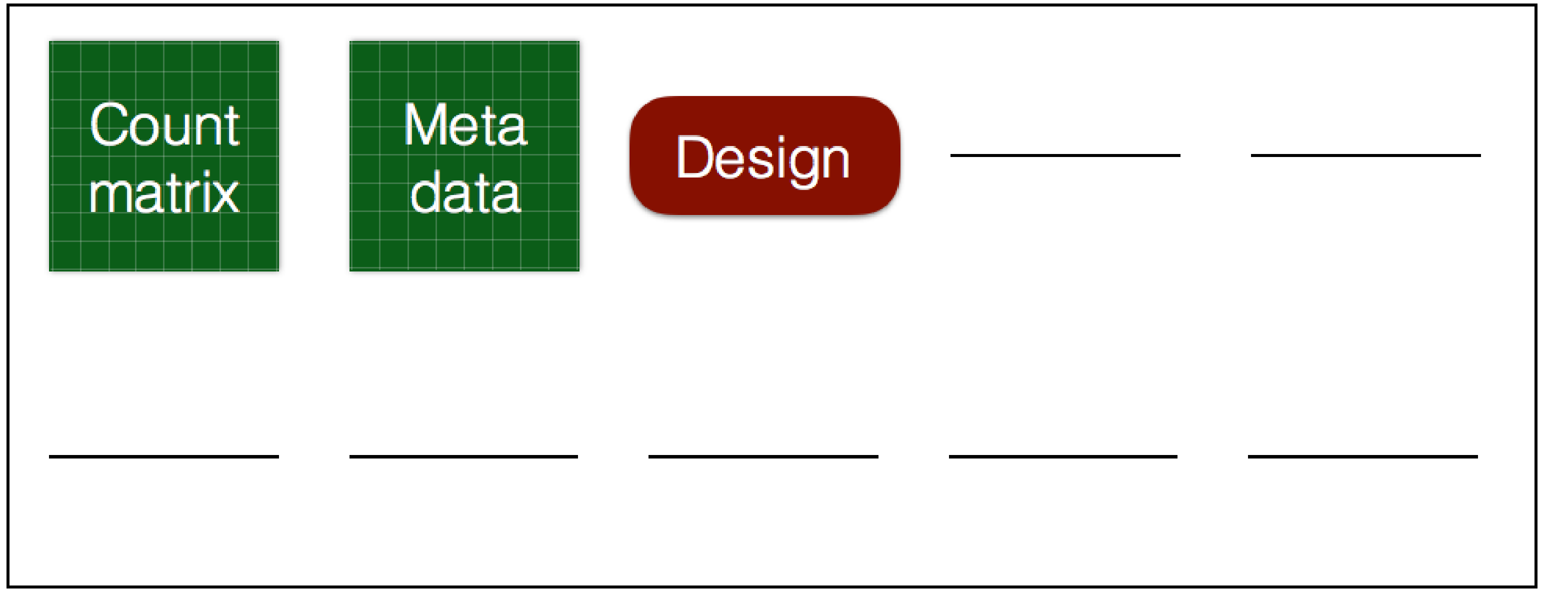
RNA-Seq count distribution

```
ggplot(raw_counts) +
  geom_histogram(aes(x = wt_normal1), stat = "bin", bins = 200) +
  xlab("Raw expression counts") +
  ylab("Number of genes")
```



Preparation for differential expression analysis: DESeq2 object

```
dds <- DESeqDataSetFromMatrix(countData = rawcounts,
                              colData = metadata,
                              design = ~ condition)
```



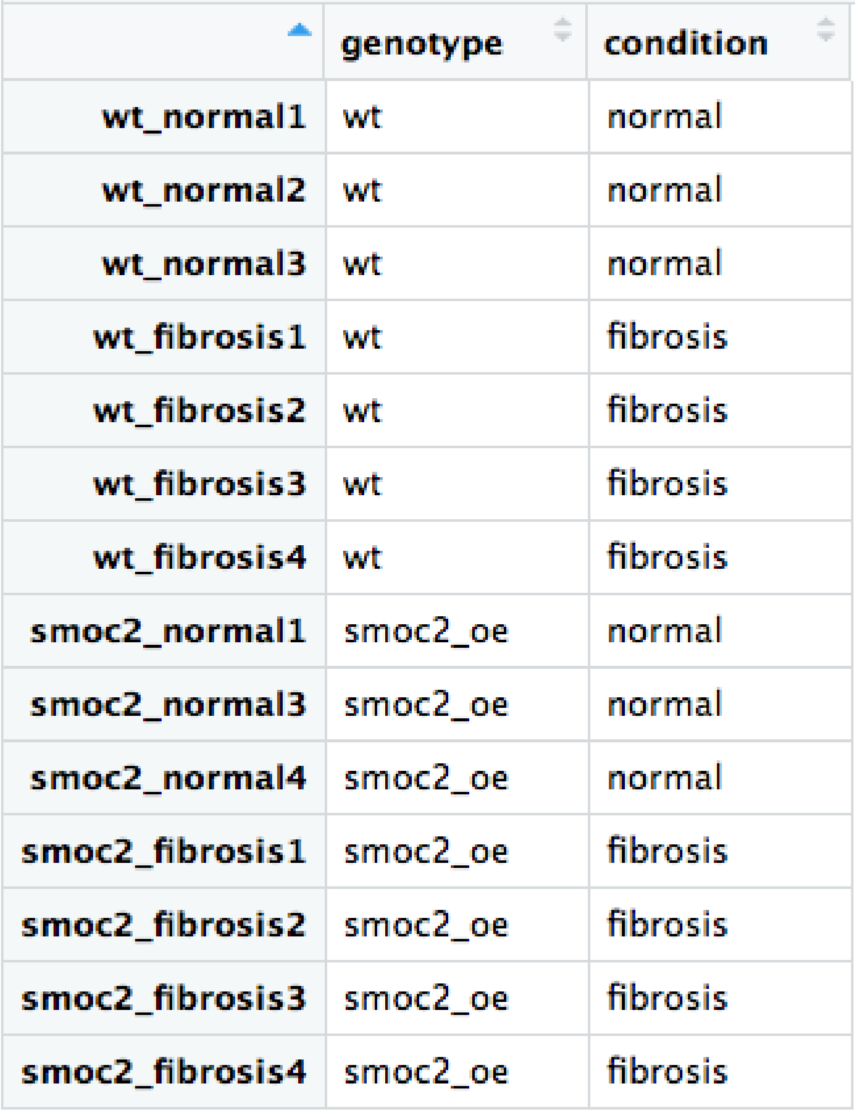
Preparation for differential expression analysis: metadata

```r
# Create vectors containing metadata for the samples
genotype <- c("wt", "wt", "wt", "wt", "wt", "wt", "wt")
condition <- c("normal", "fibrosis", "normal",
               "fibrosis", "normal", "fibrosis", "fibrosis")

# Combine vectors into a data frame
wt_metadata <- data.frame(genotype, wildtype)

# Create the row names with the associated sample names
rownames(wt_metadata) <- c("wt_normal3", "wt_fibrosis3", "wt_normal1",
                "wt_fibrosis2", "wt_normal2", "wt_fibrosis4", "wt_fibrosis1")
```

Preparation for differential expression analysis: metadata



	genotype	condition
wt_normal1	wt	normal
wt_normal2	wt	normal
wt_normal3	wt	normal
wt_fibrosis1	wt	fibrosis
wt_fibrosis2	wt	fibrosis
wt_fibrosis3	wt	fibrosis
wt_fibrosis4	wt	fibrosis
smoc2_normal1	smoc2_oe	normal
smoc2_normal3	smoc2_oe	normal
smoc2_normal4	smoc2_oe	normal
smoc2_fibrosis1	smoc2_oe	fibrosis
smoc2_fibrosis2	smoc2_oe	fibrosis
smoc2_fibrosis3	smoc2_oe	fibrosis
smoc2_fibrosis4	smoc2_oe	fibrosis

Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R