Introduction to Transcriptomics

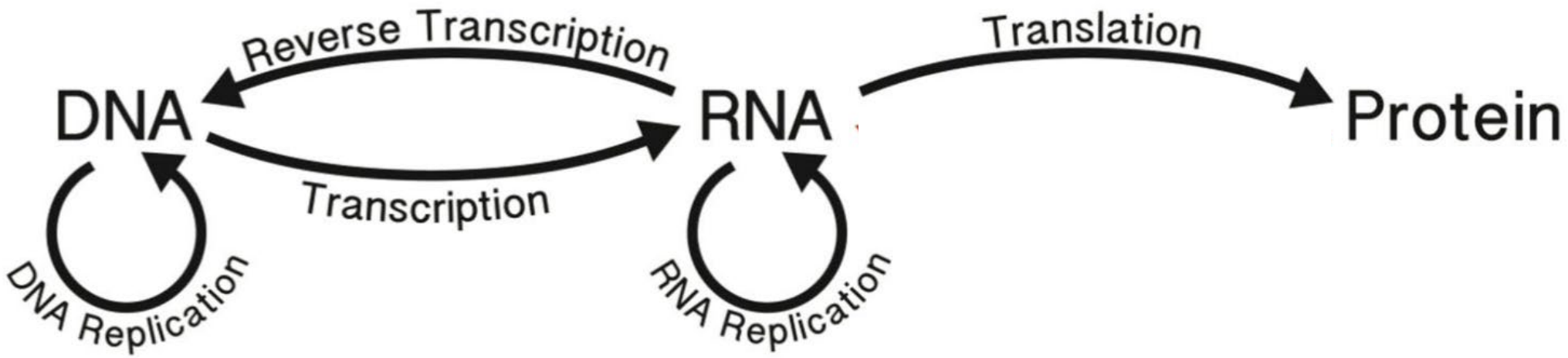
HackBio

Melyssa Minto

West Lab, Duke Neurobiology

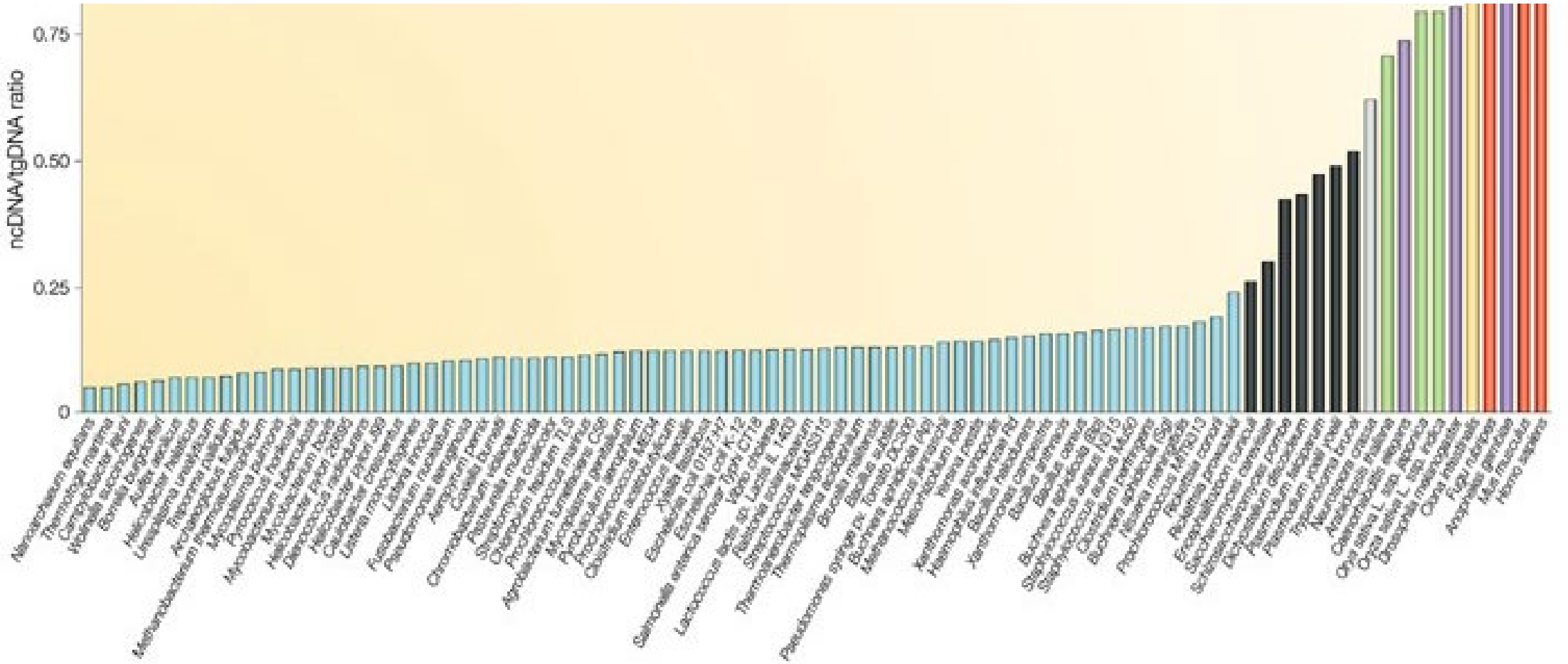
Computational Biology and Bioinformatics

Central Dogma of Biology



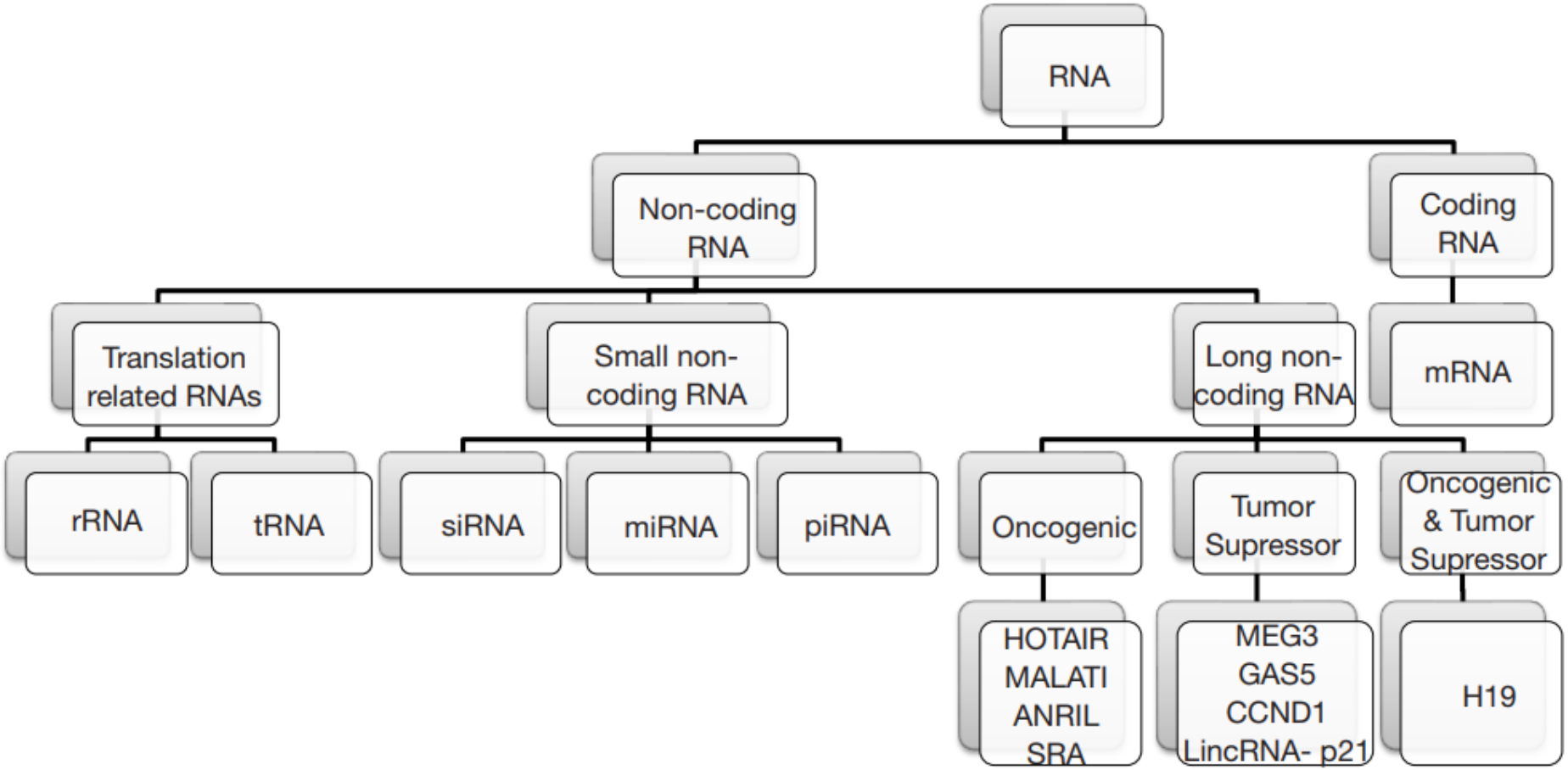
Adapted From: Lanier et. Al. J Mol Evol (2017)

Most of our genome is non-protein coding



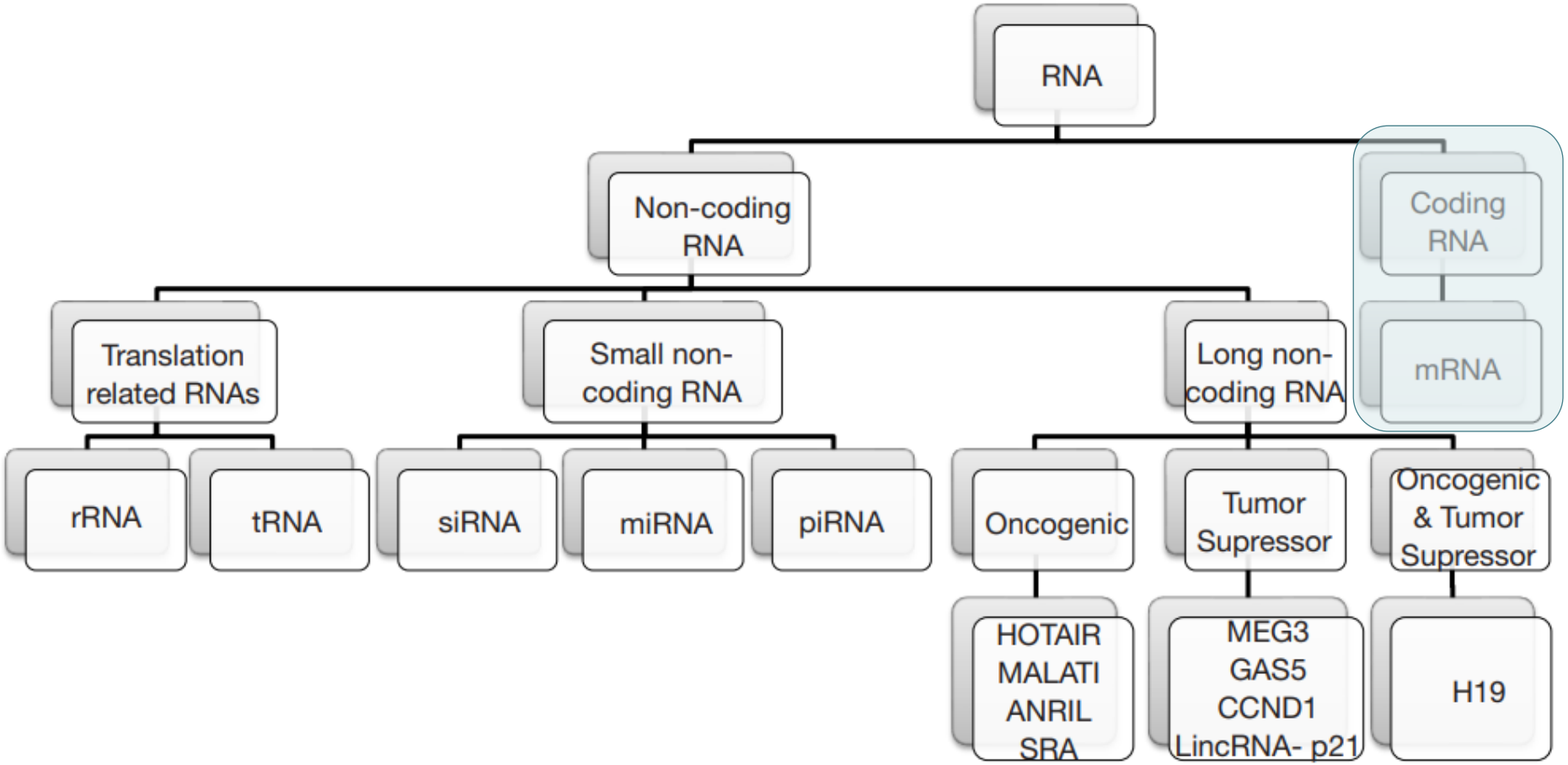
Mattick. Nat Rev Genetics (2004)

Diversity of RNA in the genome



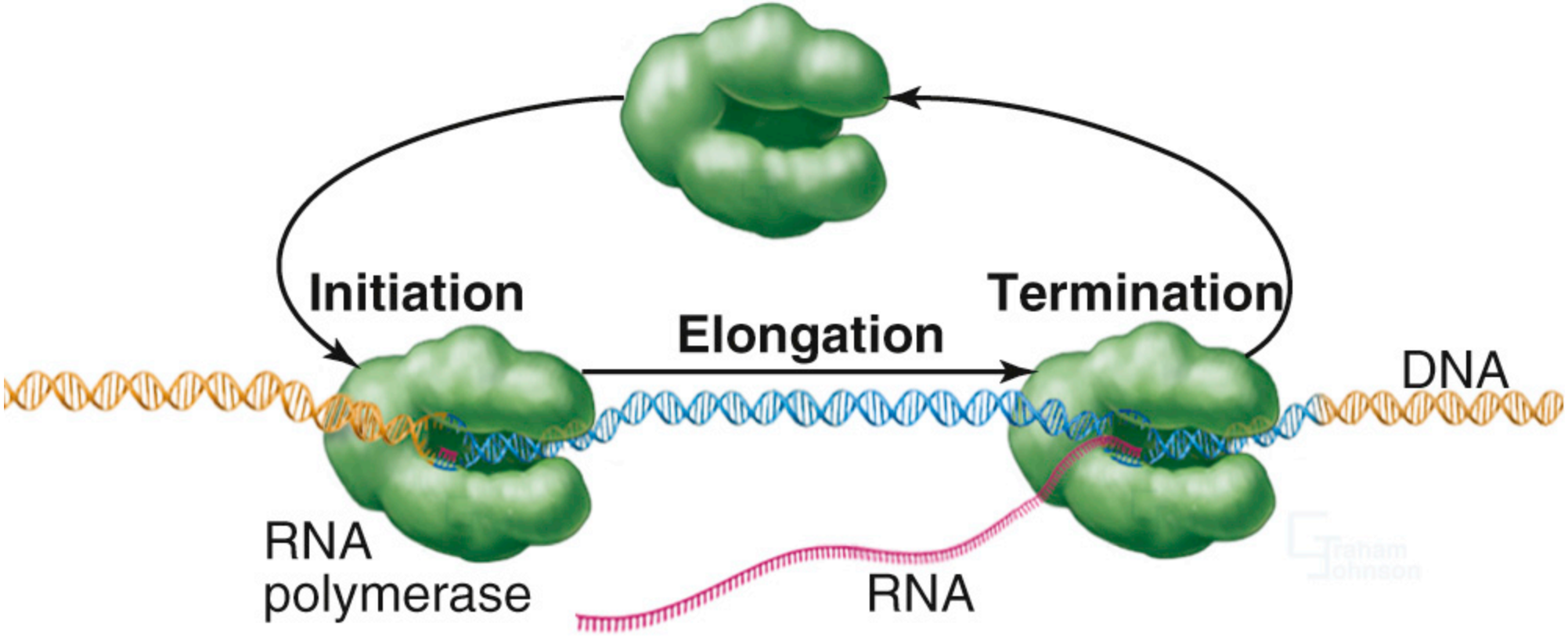
Diamantopoulos et. Al. Ann trans Med (2018)

Diversity of RNA in the genome

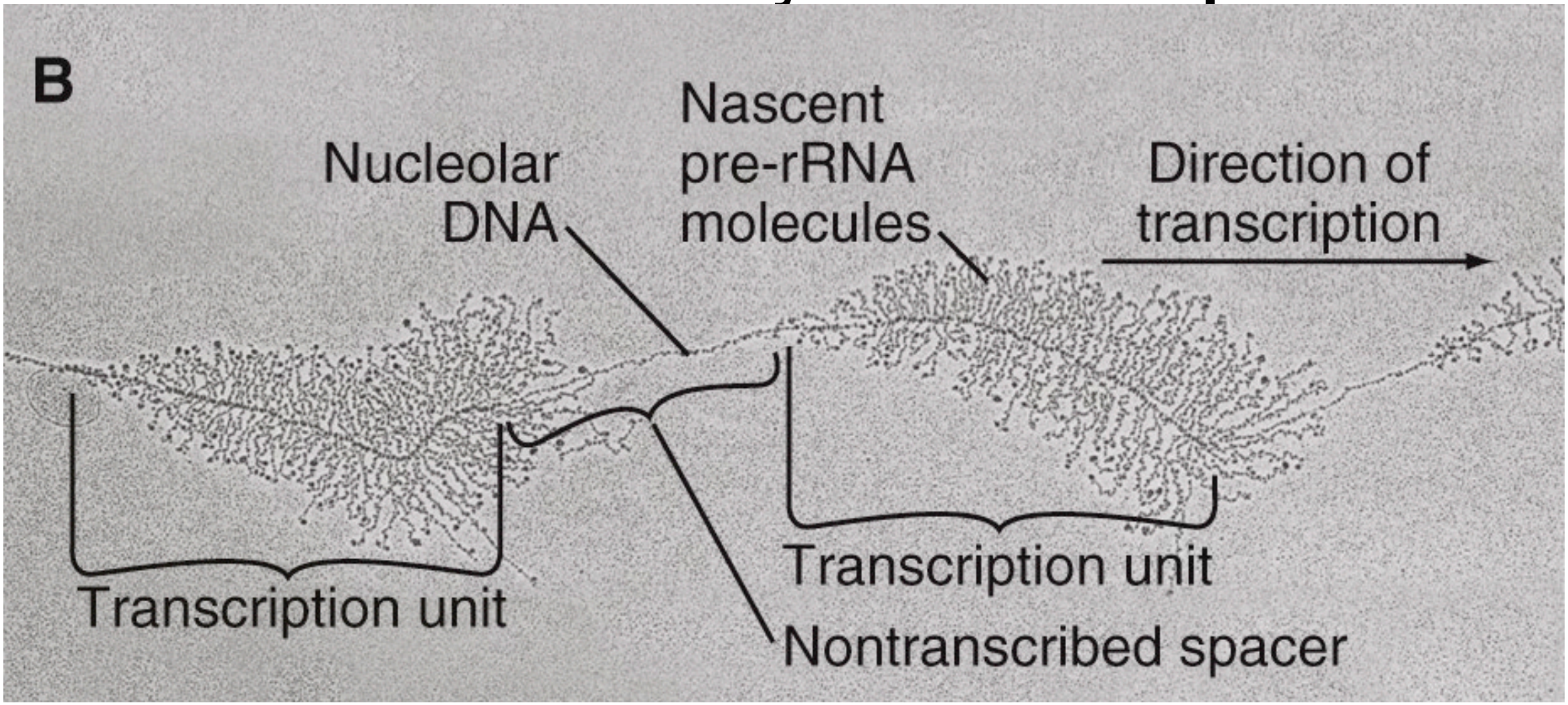


Diamantopoulos et. Al. Ann trans Med (2018)

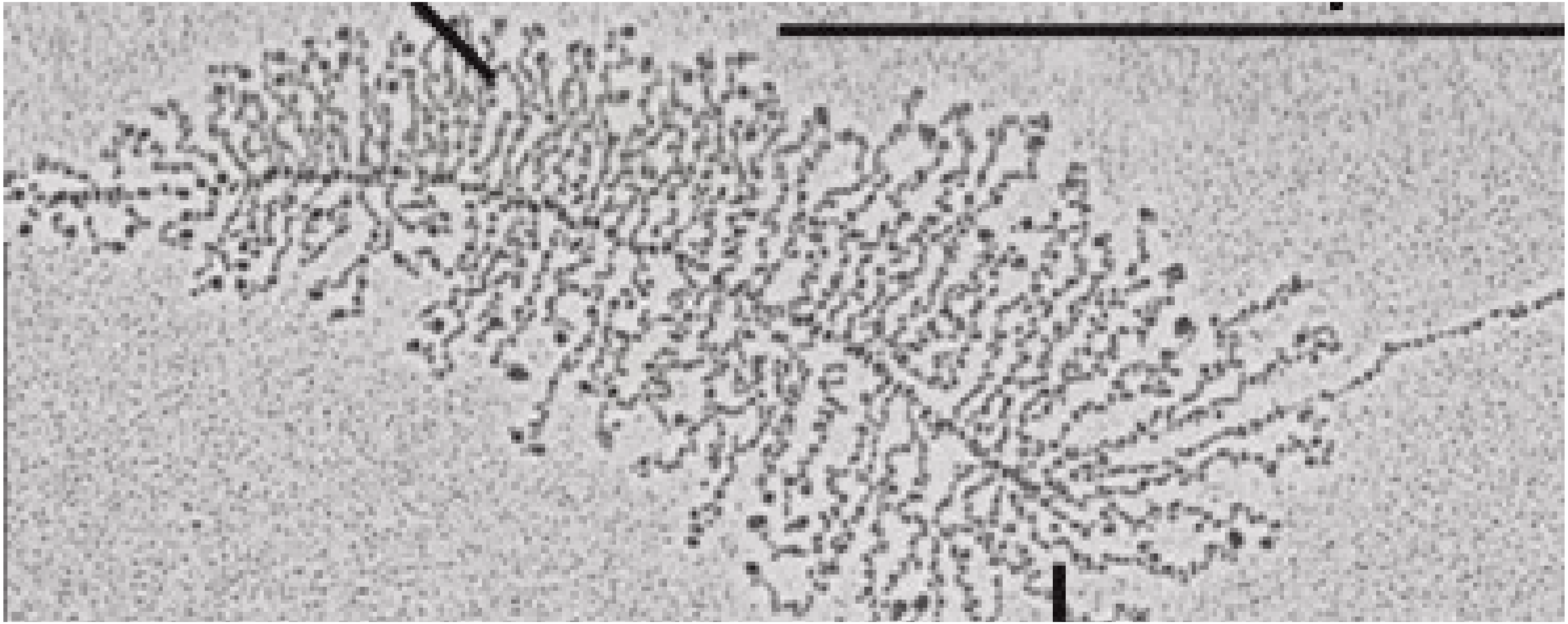
Overview of eukaryotic transcription



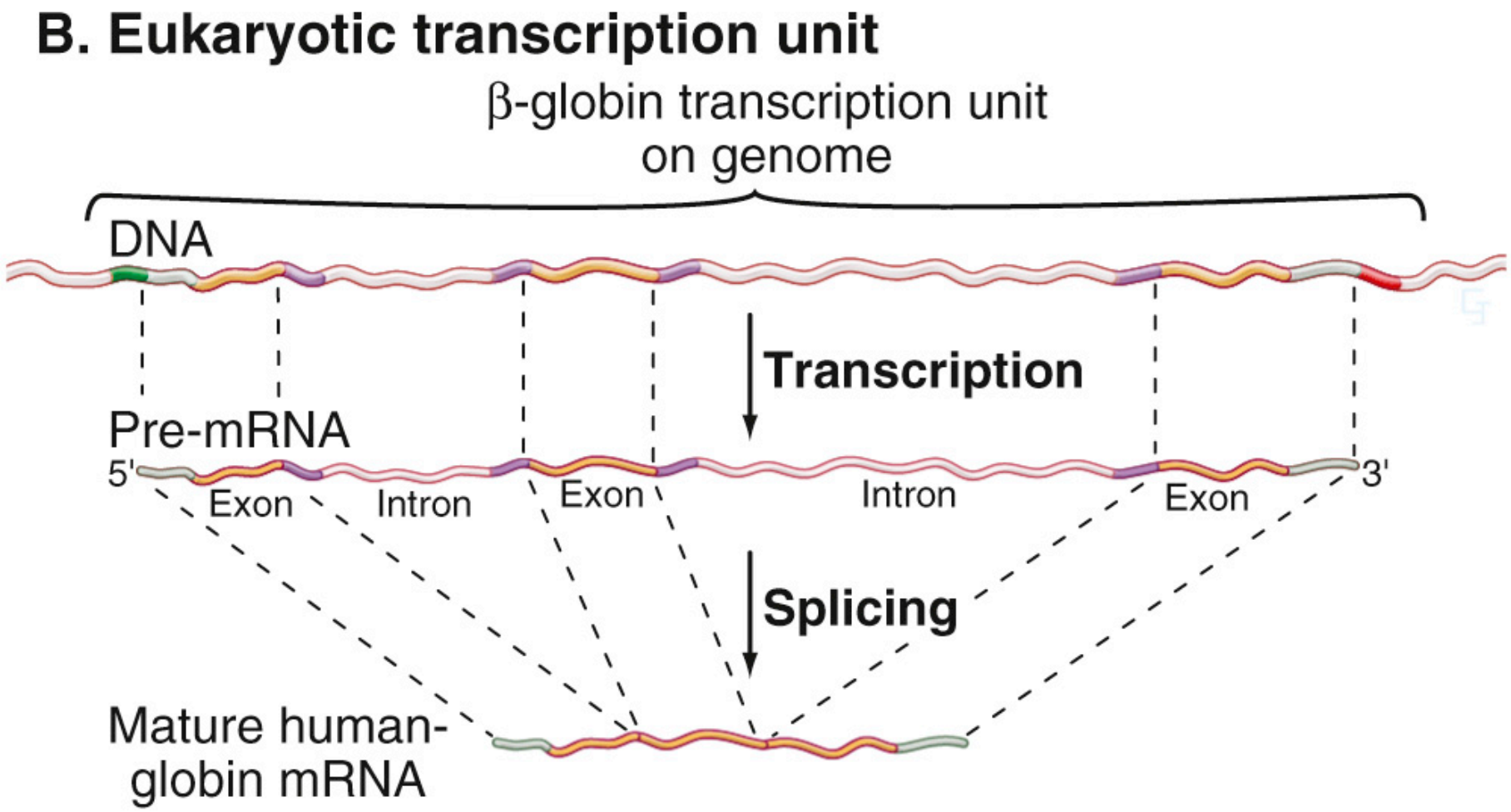
Overview of eukaryotic transcription



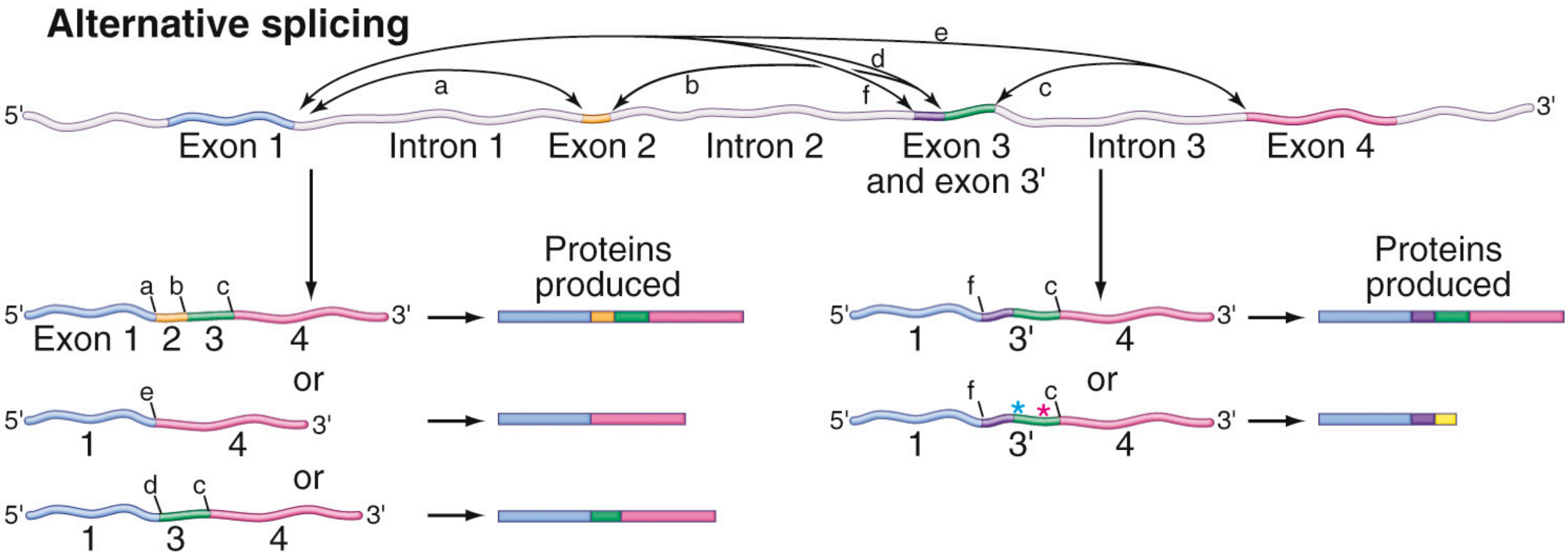
Overview of eukaryotic transcription



mRNA splicing



mRNA splicing



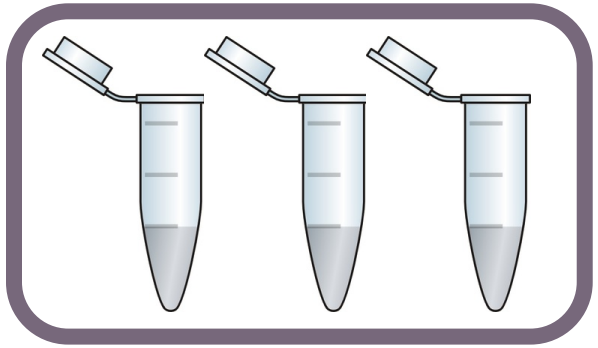
How is transcription captured?

- Extract RNA from cell
- RNA preparation
- Sequencing library preparation
- Sequencing
- Data capture
- Data analysis

RNA Library Preparation

Isolate and purify RNA

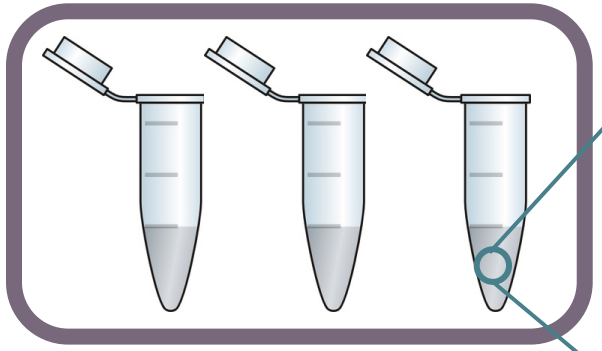
1. Extraction
2. Amplification regions of interest
3. Fragment RNA

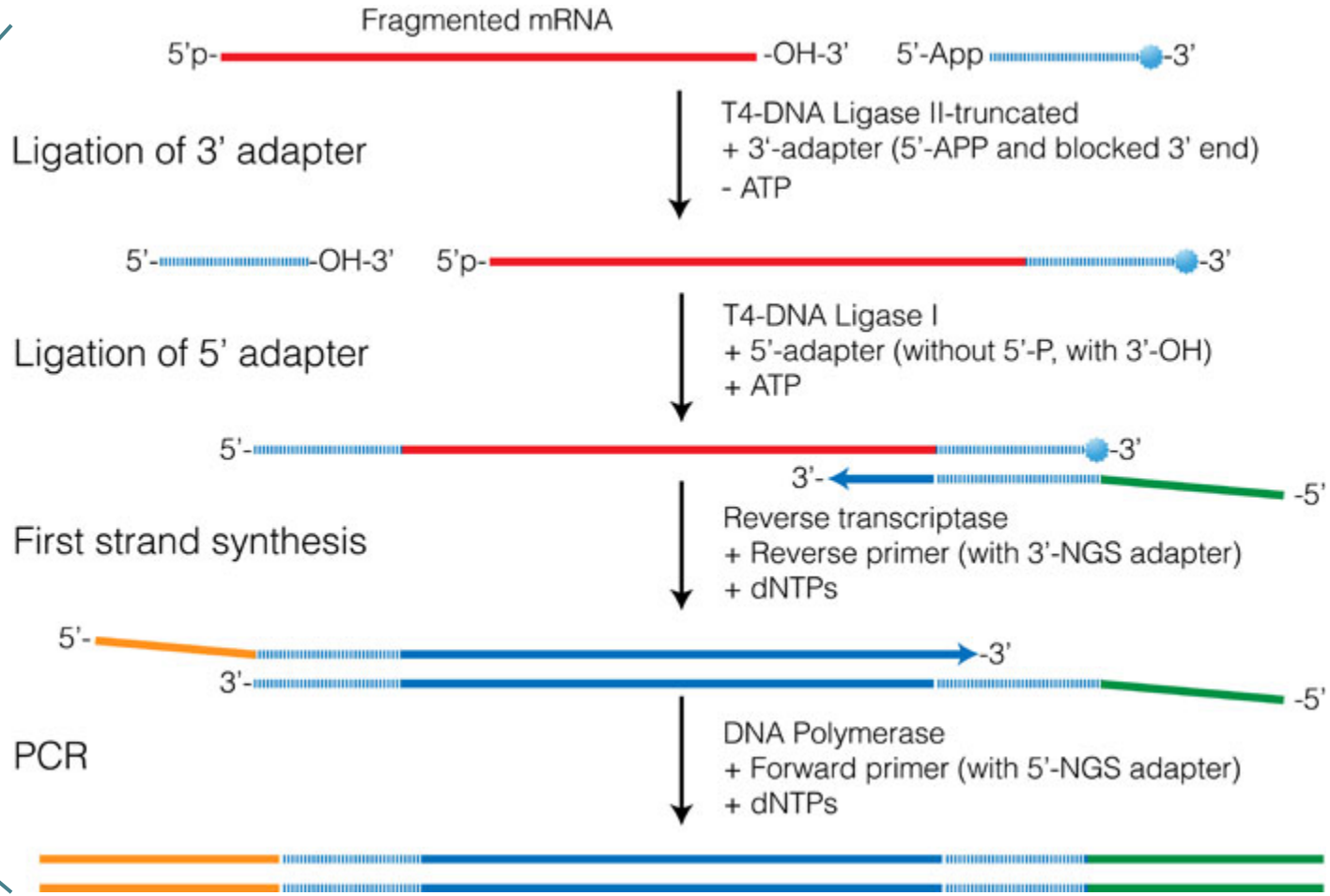


RNA Library Preparation

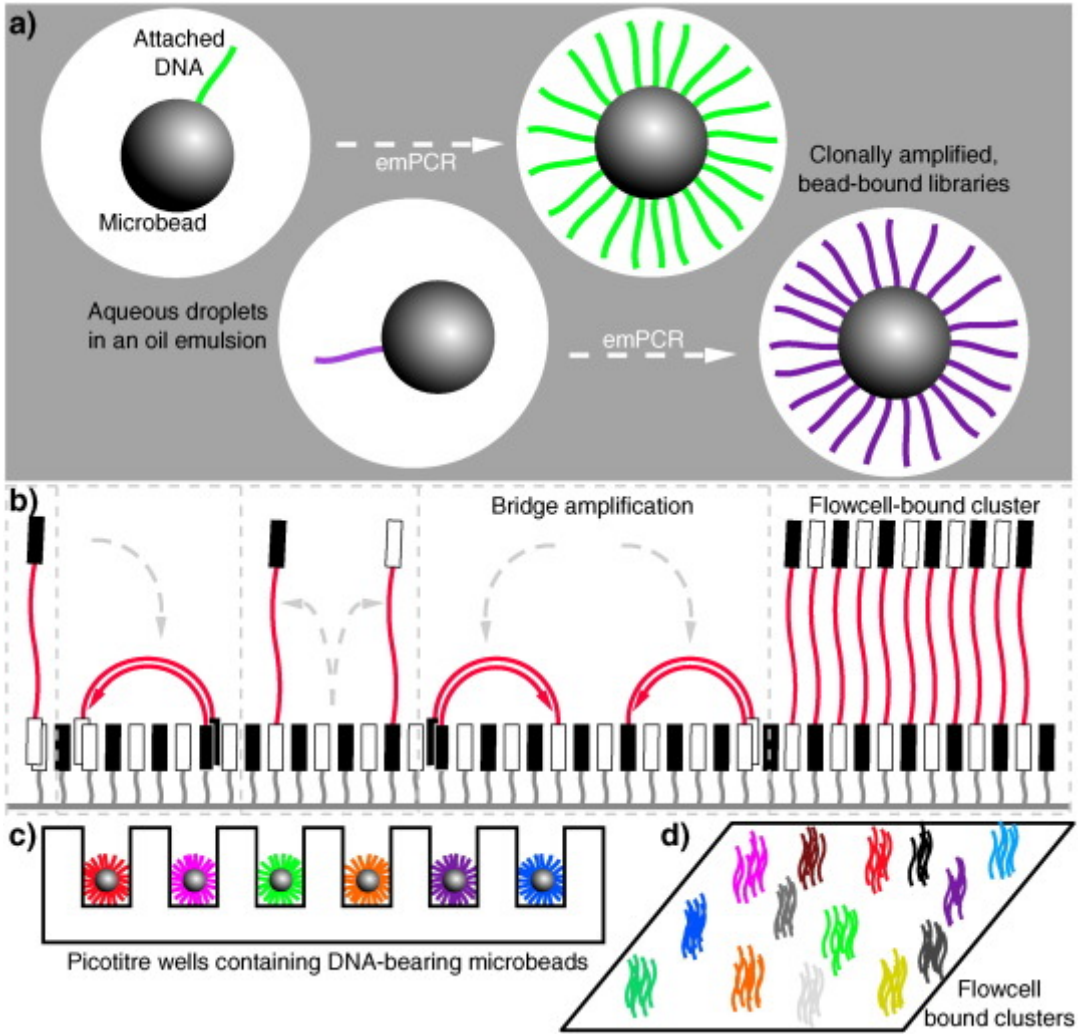
Isolate and purify RNA

1. Extraction
2. Amplification regions of interest
3. Fragment RNA



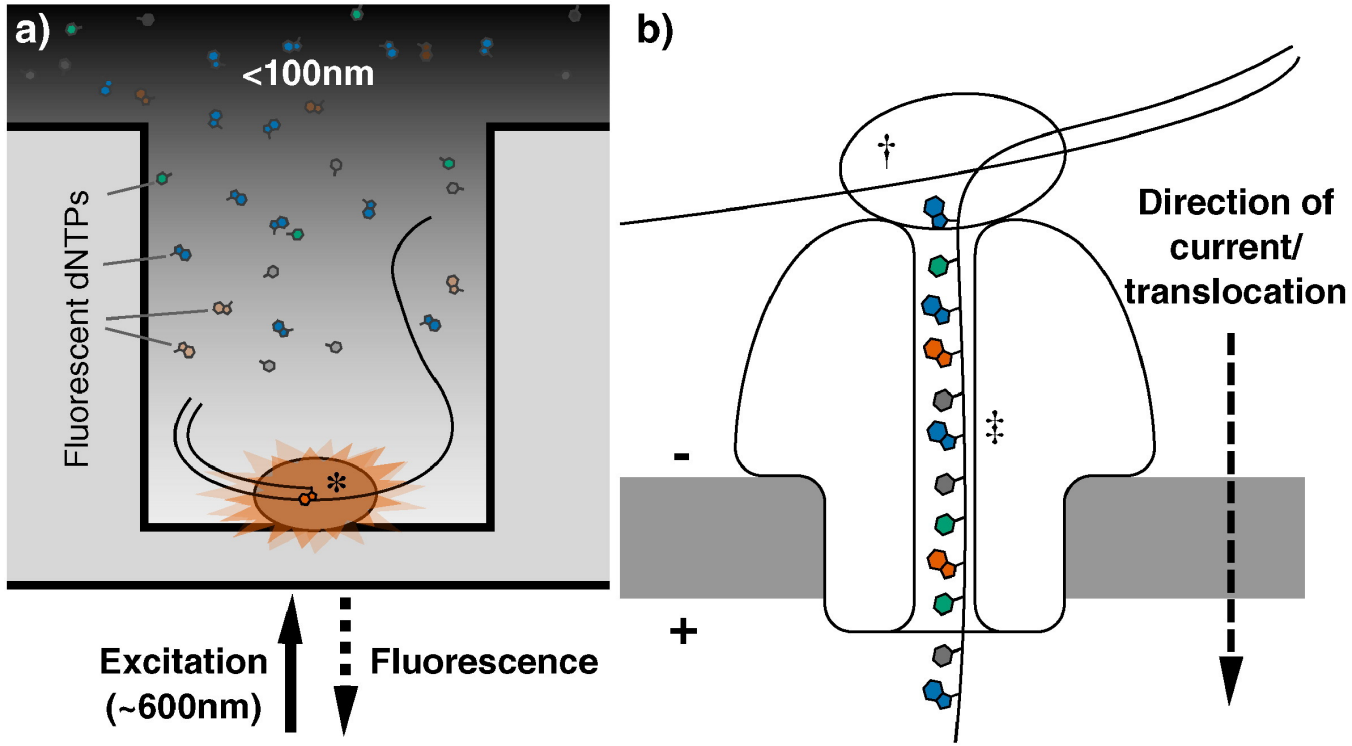


Sequencing tools



Healther & Chain. Genomics (2016)

Sequencing tools



1. RNA Pol is loaded at the bottom of each assay
2. DNA is isolated & incorporated on assay
3. As DNA passes through channel and RNAPol reads DNA – the nucleotides produce a fluorescent flash

Raw sequencing data

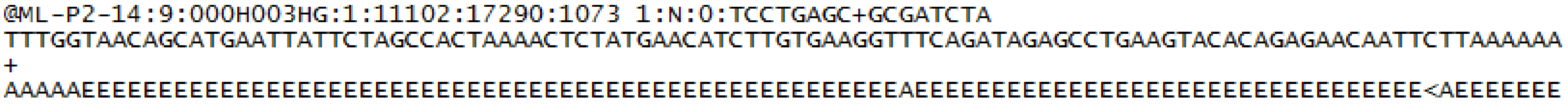
```
@A00257:355:HK7CTDRXX:1:2101:3522:1204 1:N:0:GACTACGA
CNCTTGAATGCTGAGATTACAGATGTGCTCATAGACAACAGTAGCCACATC
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3577:1204 1:N:0:GACTACGA
CNGGGAGAACCAGGTTAAAATTGAAGGTAGAAAACACTATAAGATGGAGGA
+
F#FFFFFFFFFFFFF:FFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3703:1204 1:N:0:GACTACGA
CNTATCCATATAAGAATTCAACAGAGAAACGGCAGGAAGACCCTTACCACT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Raw sequencing data

```
@A00257:355:HK7CTDRXX:1:2101:3522:1204 1:N:0:GACTACGA
CNCTTGAATGCTGAGATTACAGATGTGCTCATAGACAACAGTAGCCACATC
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3577:1204 1:N:0:GACTACGA
CNGGGAGAACCAGGTTAAAATTGAAGGTAGAAAACACTATAAGATGGAGGA
+
F#FFFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3703:1204 1:N:0:GACTACGA
CNTATCCATATAAGAATTCAACAGAGAAACGGCAGGAAGACCCTTACCACT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Raw sequencing data

Sequence identifier



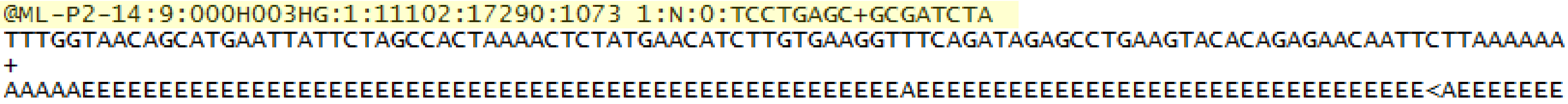
```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Actual sequence

Base quality score

Raw sequencing data

Sequence identifier



```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Actual sequence

Base quality score

https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html

Raw sequencing data

Sequence identifier

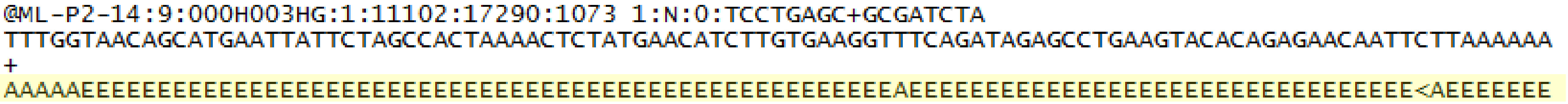
```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Actual sequence

Base quality score

Raw sequencing data

Sequence identifier



```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Actual sequence

Base quality score

Raw sequencing data

Sequence ider

Actual sequence

```
@ML-P2-14:9:000H
TTTGGTAACAGCATGA
+
AAAAAEEEEEEEEEEE
```

```
AACAATTCTTAAAAAA
EEEEEEE<AEEEEEEE
```

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

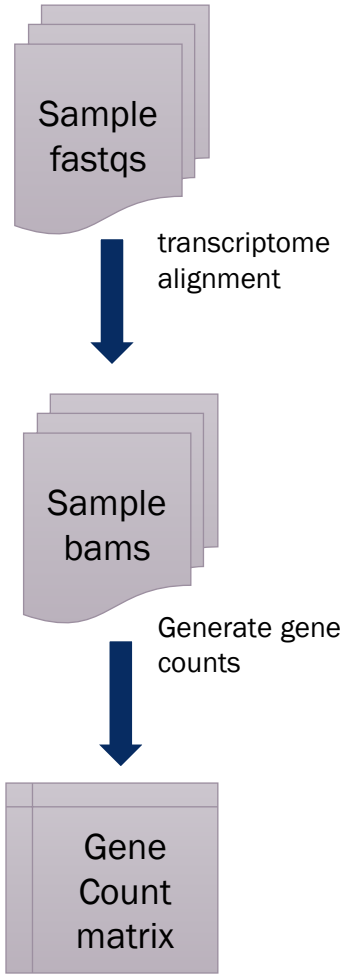
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

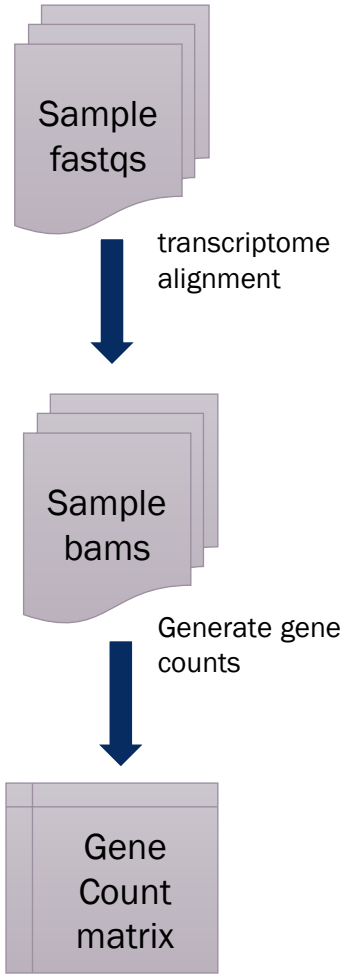
Transcriptomics pre-proccesing workflow

Preprocessing



Transcriptomics pre-proccesing workflow

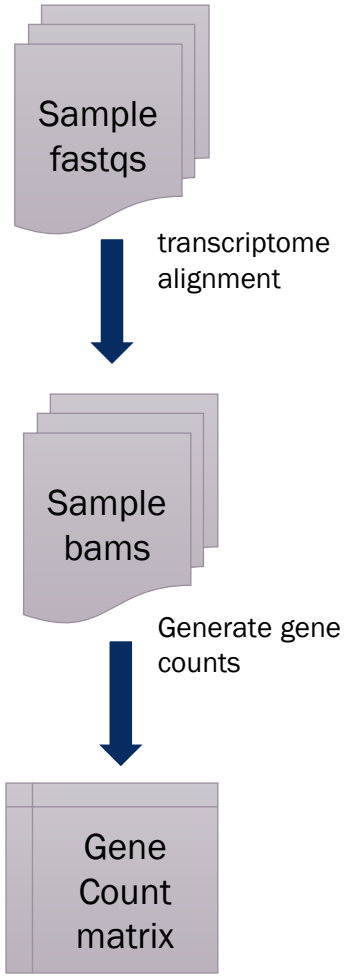
Preprocessing

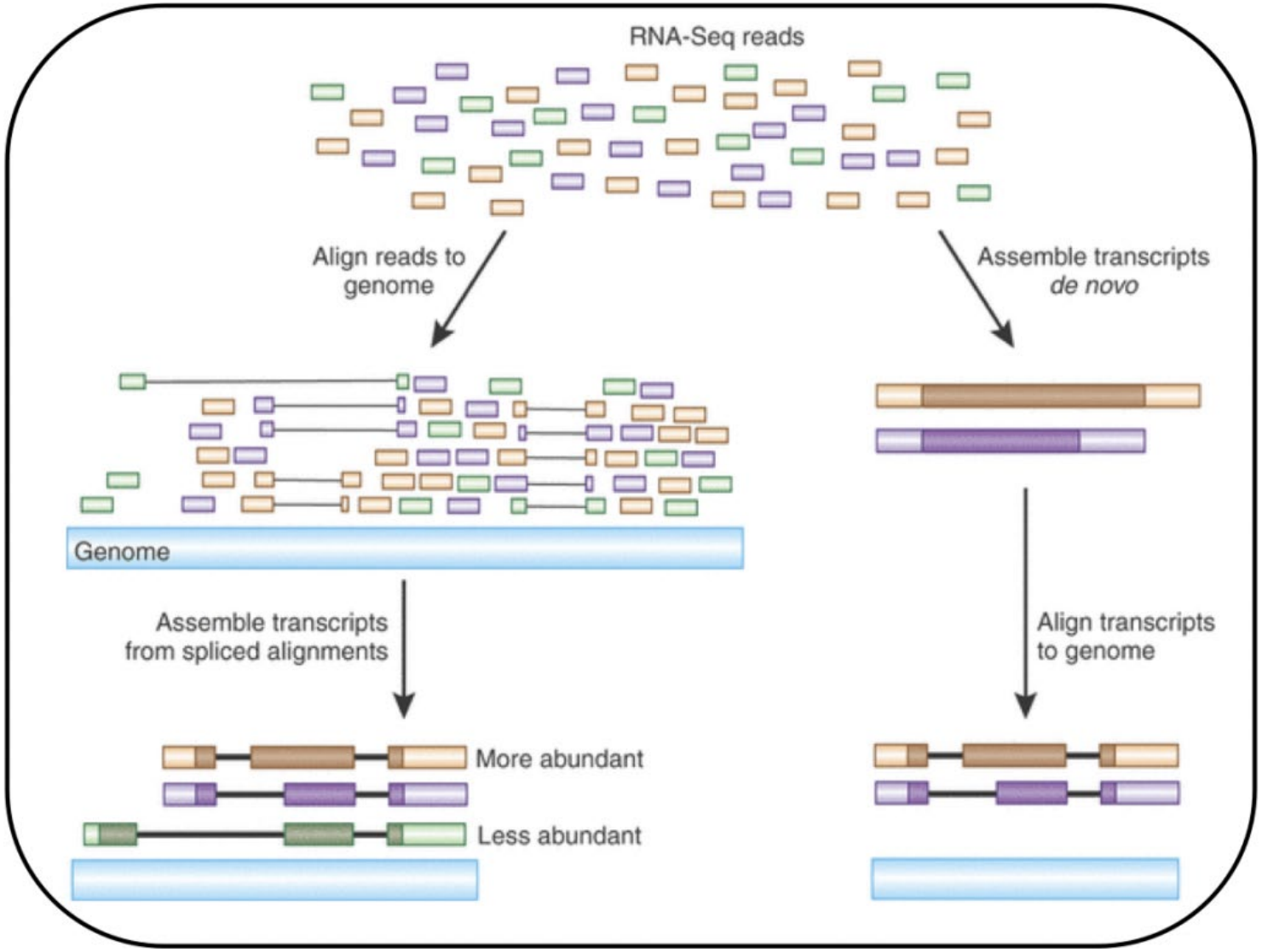



```
@A00257:355:HK7CTDRXX:1:2101:3522:1204 1:N:0:GACTACGA
CNCTTGAATGCTGAGATTACAGATGTGCTCATAGACAACAGTAGCCACATC
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3577:1204 1:N:0:GACTACGA
CNGGGAGAACCAGGTTAAAATTGAAGGTAGAAAACACTATAAGATGGAGGA
+
F#FFFFFFFFFFFF:FFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFF
@A00257:355:HK7CTDRXX:1:2101:3703:1204 1:N:0:GACTACGA
CNTATCCATATAAGAATTCAACAGAGAAACGGCAGGAAGACCCTTACCACT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Transcriptomics pre-proccesing workflow

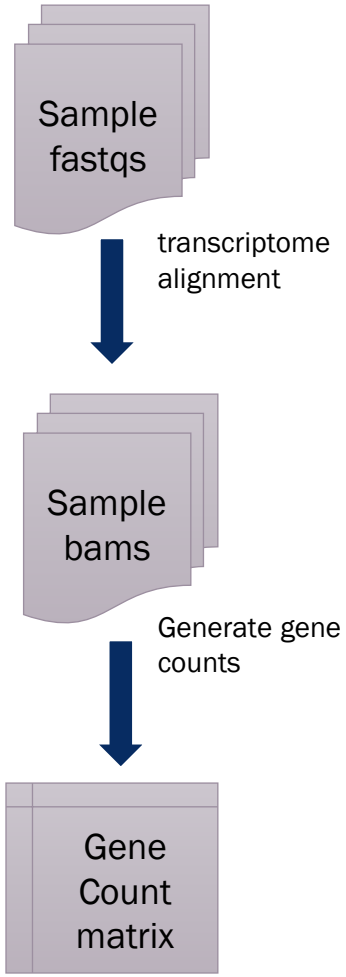
Preprocessing

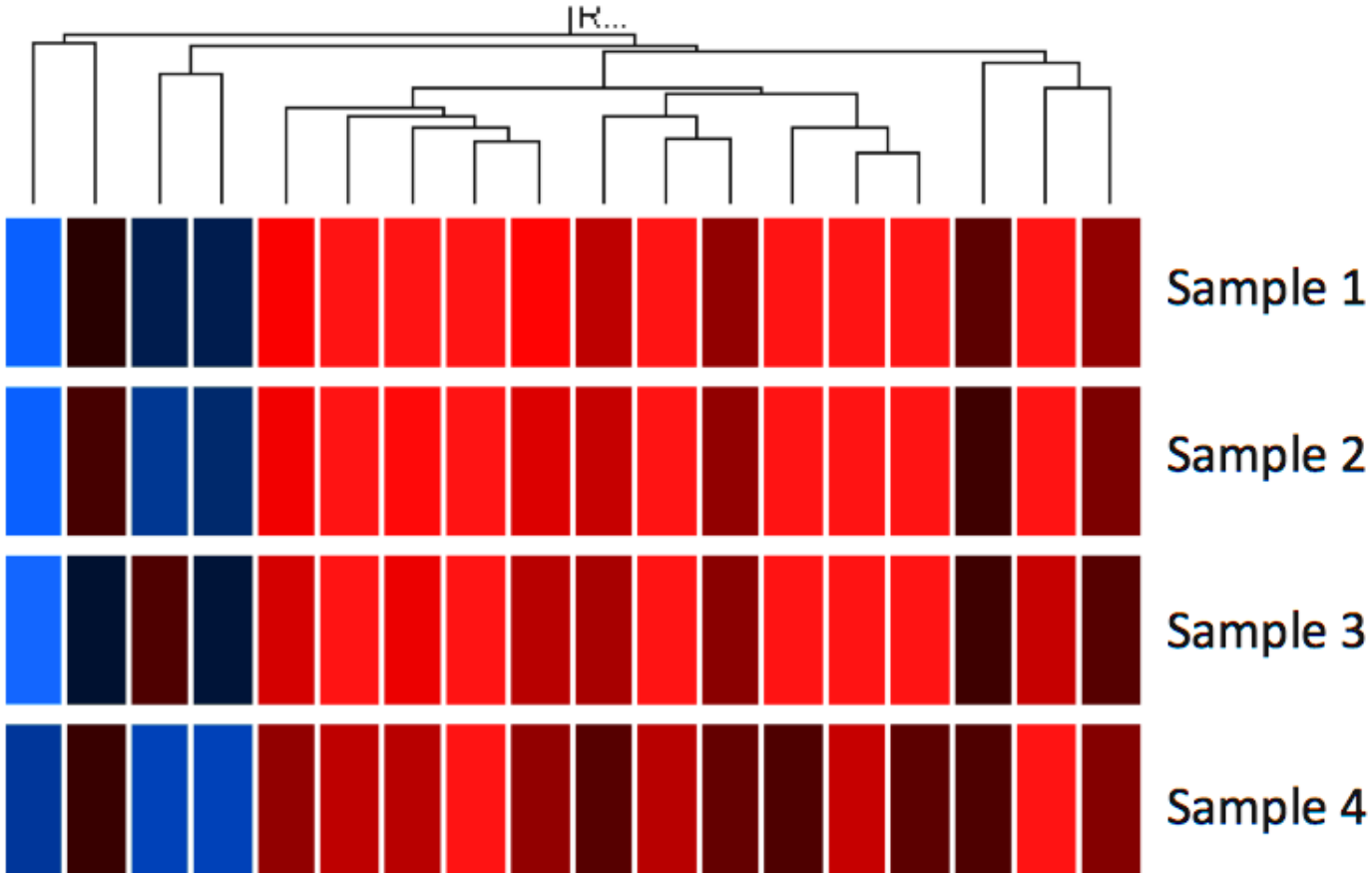




Transcriptomics pre-proccesing workflow

Preprocessing





Brain Transcriptome Databases

Table 1. Highlighted brain transcriptome databases[a]

Analysis	Web Interface	Reference	Species	Age	Sample	Method	Isoform	Accession
Spatiotemporal	http://hbatlas.org	Johnson et al., 2009	Human	Lifespan	Multi, macrodissection	Microarray	–	GSE13344
		Kang et al., 2011						GSE25219
	http://hbatlas.org/mouseNCXtranscriptome	Fertuzinhos et al., 2014	Mouse	Postnatal	Ctx layer, microdissection	RNA-seq	–	SRP031888
	http://www.blueprintnhpatlas.org	Bakken et al., 2016	Macaque	Lifespan	Multi, macrodissection, and LMD	Microarray	–	At database
Spatial	http://human.brain-map.org	Hawrylycz et al., 2012	Human	Adult	Multi, macrodissection, and LMD	Microarray	–	At database
	http://genserv.anat.ox.ac.uk/layers	Belgard et al., 2011	Mouse	Adult	Ctx layer, microdissection	RNA-seq	+	GSE27243
	http://rakiclab.med.yale.edu/transcriptome	Ayoub et al., 2011	Mouse	Embryonic	Ctx embryonic layer, LMD	RNA-seq	+	GSE30765
	http://www.brainspan.org/lcm	Miller et al., 2014	Human	Midfetal	Multi, LMD	Microarray	–	At database
	https://www.gtexportal.org	GTEx Consortium, 2015	Human	Adult	Many tissues and cell lines	RNA-seq	+	At database
Temporal	http://braincloud.jhmi.edu	Colantuoni et al., 2011	Human	Lifespan	Prefrontal Ctx, macrodissection	Microarray	–	GSE30272
Cell type- specific	http://brainrnaseq.org	Zhang et al., 2014	Mouse	Adult	Ctx, genetic labeling, immunopanning	RNA-seq	+	GSE52564
		Zhang et al., 2016	Human	Fetal/adult	Ctx, Hp, immunopanning	RNA-seq	–	GSE73721
	http://genetics.wustl.edu/jdlab/csea-tool-2	Doyle et al., 2008	Mouse	Adult	Multi, genetic labeling, ribosome affinity purification	Microarray	–	GSE13379
		Xu et al., 2014						
	http://decon.fas.harvard.edu	Molyneaux et al., 2015	Mouse	Embryonic	Ctx, transcription factor FACS	RNA-seq	+	GSE63482
	http://hipposeq.janelia.org	Cembrowski et al., 2016	Mouse	Adult	Hp, genetic labeling, manual selection	RNA-seq	–	GSE74985
	http://neuroseq.janelia.org	Sugino et al., 2017	Mouse	Adult	Multi, genetic labeling, manual selection	RNA-seq	+	GSE79238
Single-cell	http://linnarssonlab.org/cortex	Zeisel et al., 2015	Mouse	Adult	Ctx, Fluidigm	RNA-seq	–	GSE60361
	http://genebrowser.unige.ch/science2016	Telley et al., 2016	Mouse	Embryonic	Ctx, ventricle dye, FACS, Fluidigm	RNA-seq	–	NA
	https://portals.broadinstitute.org/single_cell	Shekhar et al., 2016	Mouse	Adult	Retina, genetic labeling, Drop-seq	RNA-seq	–	GSE81905
	https://portals.broadinstitute.org/single_cell	Habib et al., 2016	Mouse	Adult	Hp, single nuclei, FACS, sNuc-seq	RNA-seq	–	GSE84371
	https://bit.ly/cortexSingleCell	Nowakowski et al., 2017	Human	Fetal	Ctx, ganglionic eminence, Fluidigm	RNA-seq	–	PRJNA295469
	http://gbmseq.org	Darmanis et al., 2017	Human	Adult	Ctx tumor, immunopanning, FACS	RNA-seq	–	GSE84465
Integrative	https://www.encodeproject.org	ENCODE Project Consortium, 2012	Many	Many	Many tissues and cell lines	Multiomics	+	Many
	http://celltypes.brain-map.org	Tasic et al., 2016	Mouse	Adult	Ctx, genetic labeling, FACS	RNA-seq	–	GSE71585

[a]Ctx, Cortex; Hp, hippocampus; multi, multiple brain regions. Isoform column indicates availability of isoform information via web interface.

Keil et al. J Neurosci. 2018

Keeping up with the current research

BMC Part of Springer Nature

BMC Genomics

PLOS ONE

Bioinformatics

OXFORD ACADEMIC

Bioinformatics

nature neuroscience

BMC Part of Springer Nature

BMC Bioinformatics

JNeurosci
THE JOURNAL OF NEUROSCIENCE

An Official Journal of SfN Society for Neuroscience