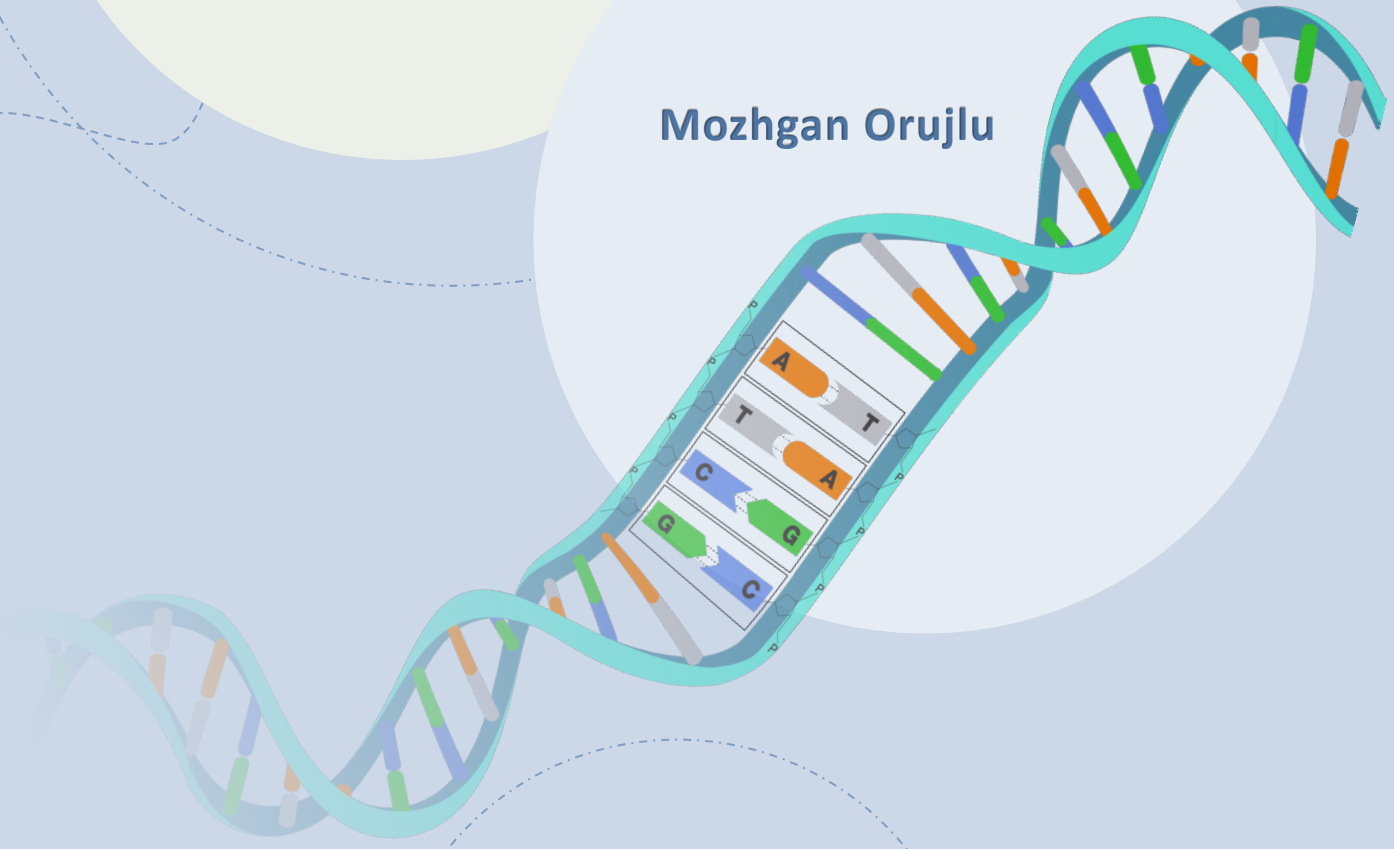
An Introduction to

Single-Cell RNA Sequencing Data

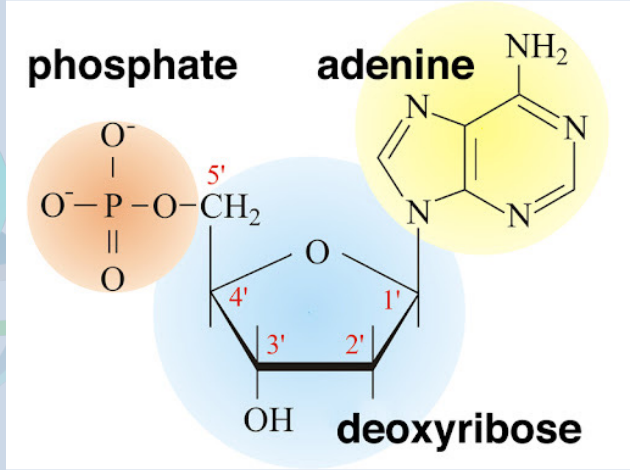
Gene Regulation

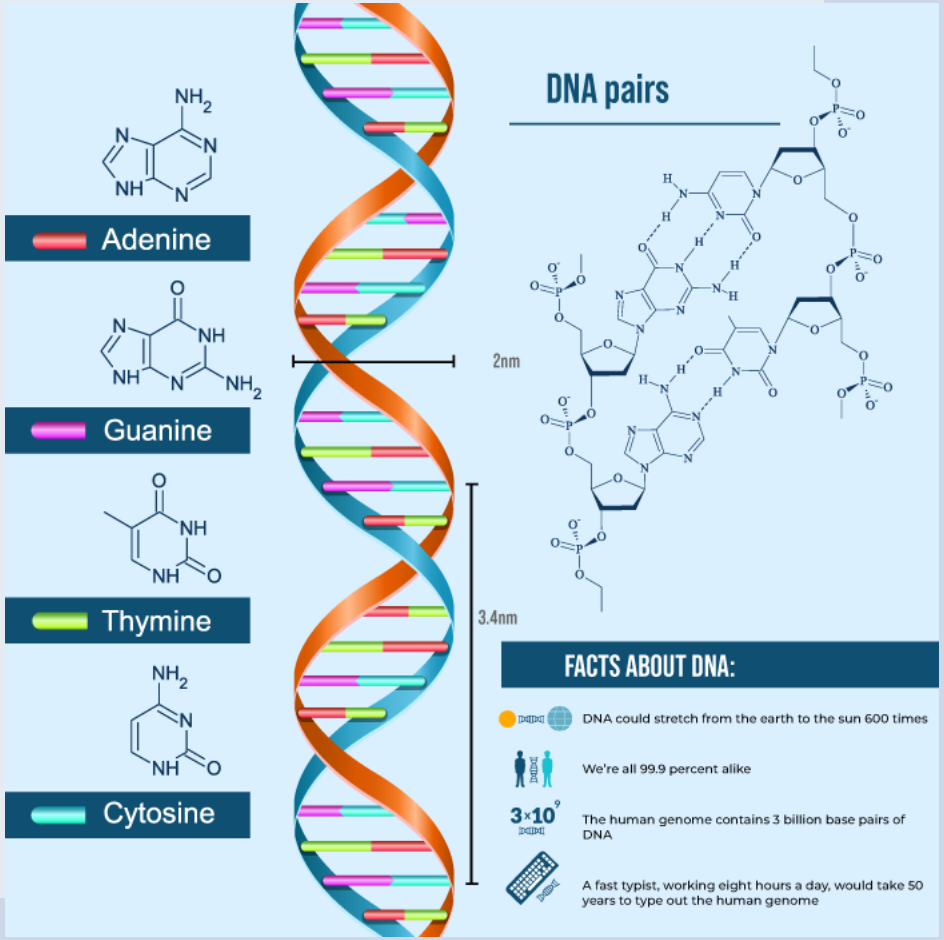
Mozhgan Orujlu



DNA

- DNA is made up of small repeating units called **nucleotides**.
- Each nucleotide consists of three components:
  - **Phosphate Group (PO_4^{3-})**
  - **Deoxyribose Sugar ($C_5H_{10}O_4$)**
  - **Nitrogenous Base**
    - **Adenine (A)** → $C_5H_5N_5$
    - **Thymine (T)** → $C_5H_6N_2O_2$
    - **Cytosine (C)** → $C_4H_5N_3O$
    - **Guanine (G)** → $C_5H_5N_5O$

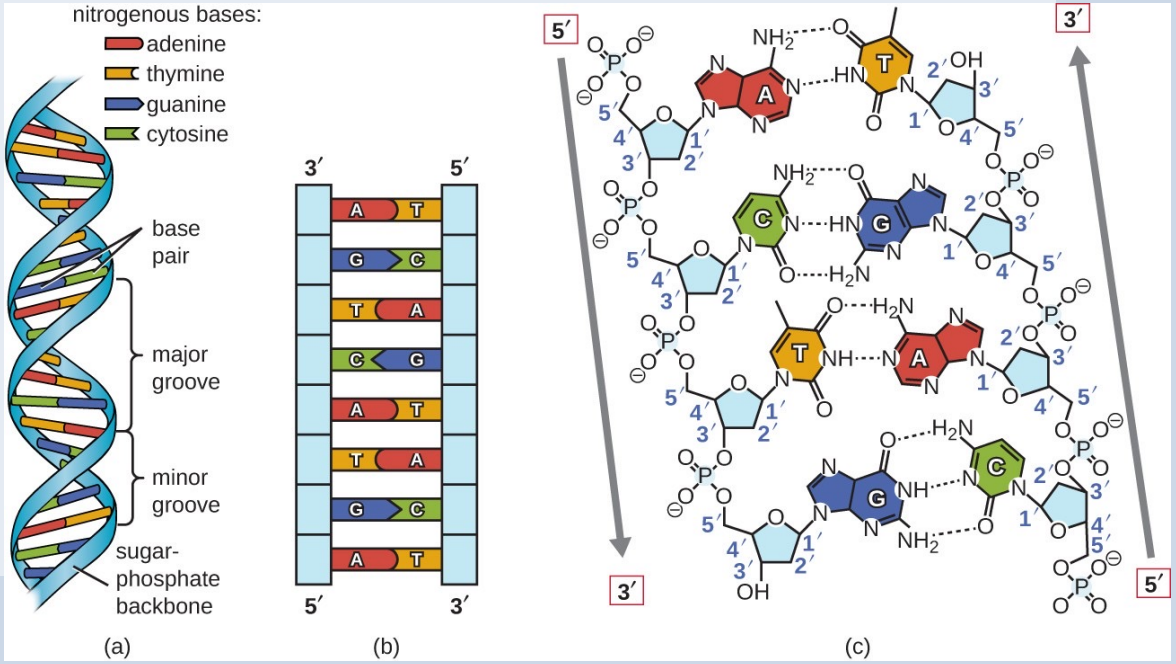




DNA

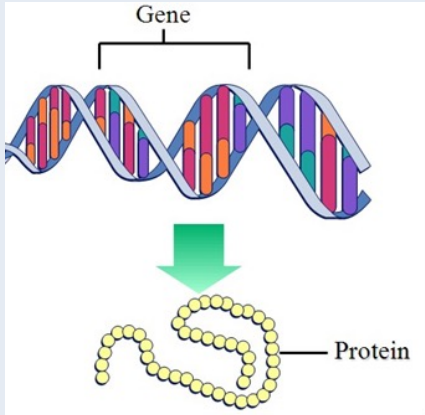
- The two strands of DNA are held together by **hydrogen bonds** between nitrogenous bases.
  - Adenine (A) pairs with Thymine (T) (A–T) with 2 hydrogen bonds
  - Cytosine (C) pairs with Guanine (G) (C–G) with 3 hydrogen bonds

- The two DNA strands are **oriented in opposite directions** (one 5' → 3', the other 3' → 5').

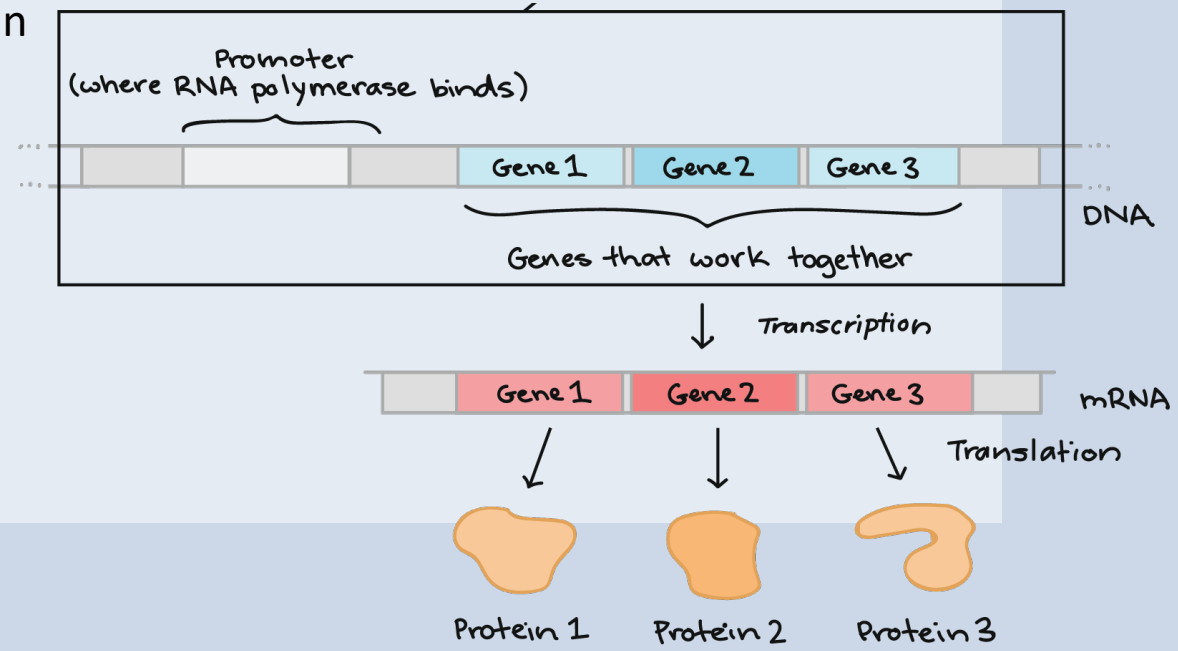


Gene

- a sequence of **nucleotides** in DNA that encodes instructions for building a **specific protein**.
  - Promoter region
  - Coding region
  - Terminator region

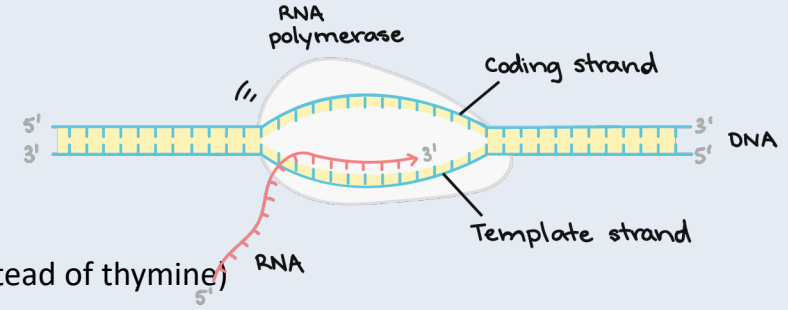


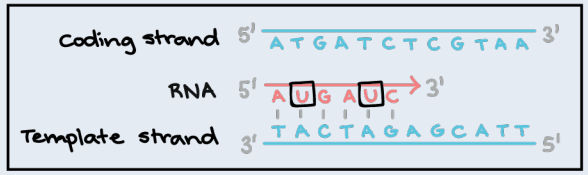
- **Two main steps:**
- **Transcription:** DNA to mRNA
- **Translation:** mRNA to Protein

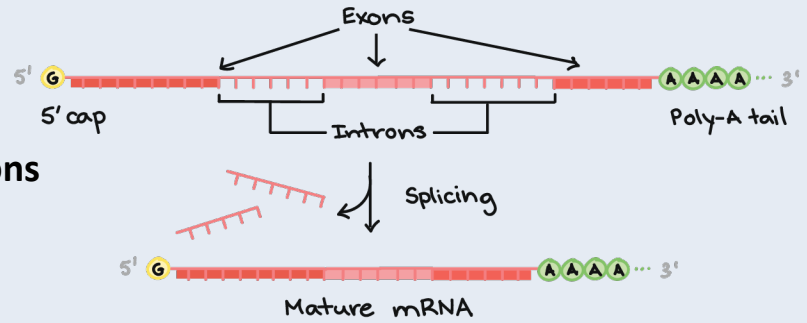


Transcription

- **Steps:**
- **Initiation:**
  - Binding and Unwinding
- **Elongation:**
  - A (Adenine) → U (Uracil) (RNA has uracil instead of thymine)
  - T (Thymine) → A (Adenine)
  - C (Cytosine) → G (Guanine)
  - G (Guanine) → C (Cytosine)
- **Termination**
- **Post-Transcriptional Modifications (in Eukaryotes):**
  - 5' Cap Addition
  - Poly-A Tail Addition
  - Splicing
    - Removing **introns** and joining **exons**

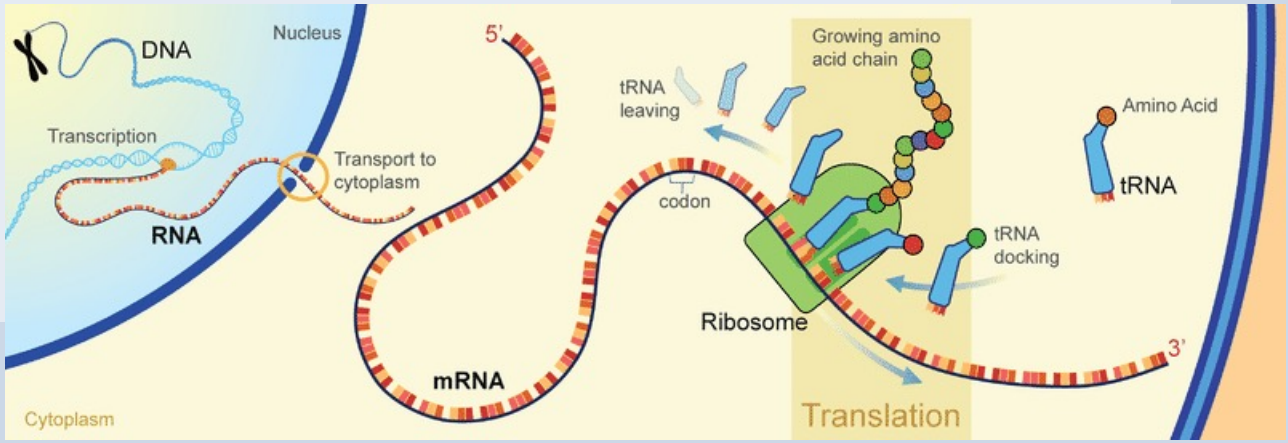


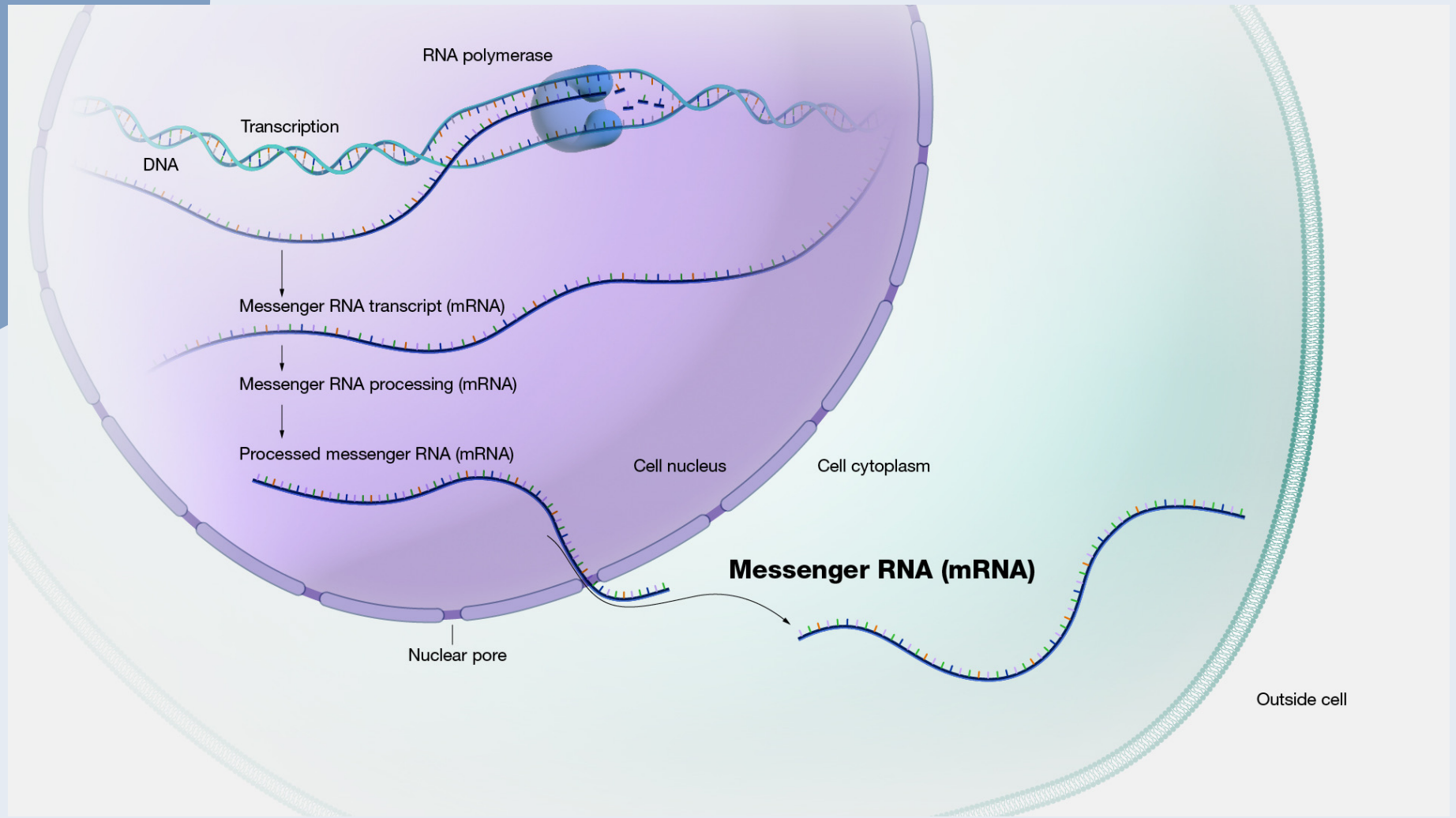




Translation

- **Initiation:**
  - Attaching ribosome
  - **Start Codon: AUG**
- **Elongation:**
  - **tRNA molecules** bring amino acids to the ribosome
  - Each tRNA has an **anticodon** that matches a specific **codon** on mRNA.
  - The ribosome forms **peptide bonds** between amino acids, linking them into a growing **polypeptide chain**.
- **Termination:**
  - **Stop Codon: UAA, UAG, UGA**
- **Post-Translational Modifications**

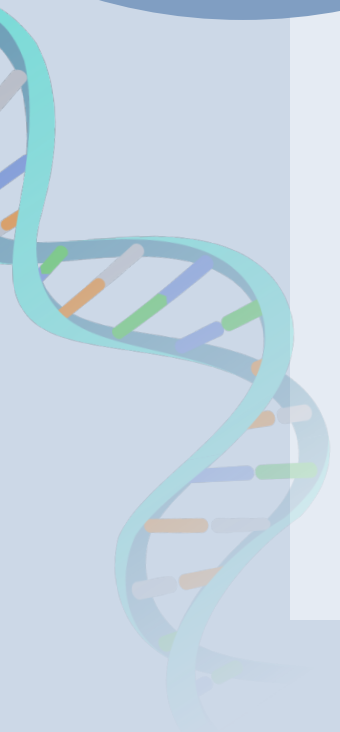



RNA polymerase
Transcription
DNA
Messenger RNA transcript (mRNA)
Messenger RNA processing (mRNA)
Processed messenger RNA (mRNA)
Cell nucleus
Cell cytoplasm
Messenger RNA (mRNA)
Nuclear pore
Outside cell

Gene Regulation

- Gene regulation is the process by which cells control the expression of genes, determining when, where, and to what extent specific genes are **turned on (activated) or off (repressed).**

- Gene regulation primarily involves **chemical modifications to DNA** rather than physically removing or attaching DNA segments.



Gene Modification

- **Gene modification?**
- adding, removing, or modifying specific genes or sequences of DNA.

Dynamic RNA **modifications** in **gene expression** regulation
IA Roundtree, ME Evans, T Pan, C He - Cell, 2017 - cell.com
… are also heavily **modified** and depend on the **modifications** for their … of these different chemical **modifications** is beginning to take … **modifications** represent a new layer of control of **genetic** …
☆ Save 99 Cite Cited by 2993 Related articles All 10 versions

RNA **modifications** modulate **gene expression** during development
M Frye, BT Harada, M Behm, C He - Science, 2018 - science.org
… of mRNA is an essential regulator of mammalian **gene expression** (4, 5). Other **modifications** … the roles of RNA **modifications** in modulating **gene expression** throughout cell differentiation …
☆ Save 99 Cite Cited by 1031 Related articles All 11 versions

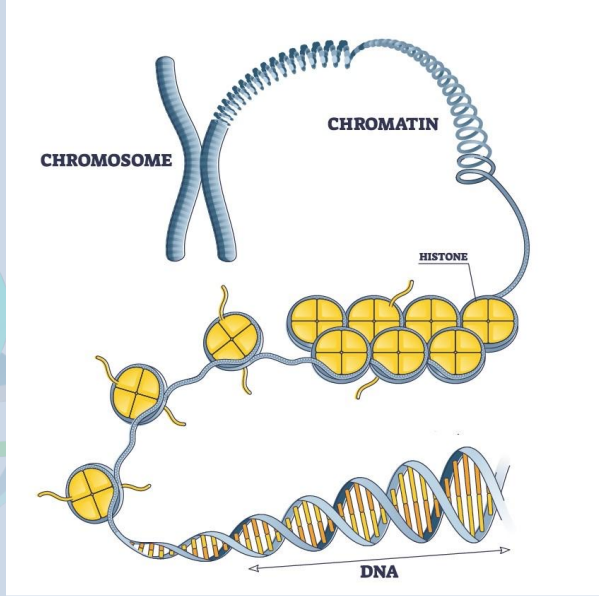
Nucleic acid **modifications** in regulation of **gene expression**
K Chen, BS Zhao, C He - Cell chemical biology, 2016 - cell.com
… Nucleic acids carry diverse **modifications** and employ these chemical … **modifications** that play important regulatory roles in biological systems, especially in regulation of **gene expression**: …
☆ Save 99 Cite Cited by 281 Related articles All 7 versions

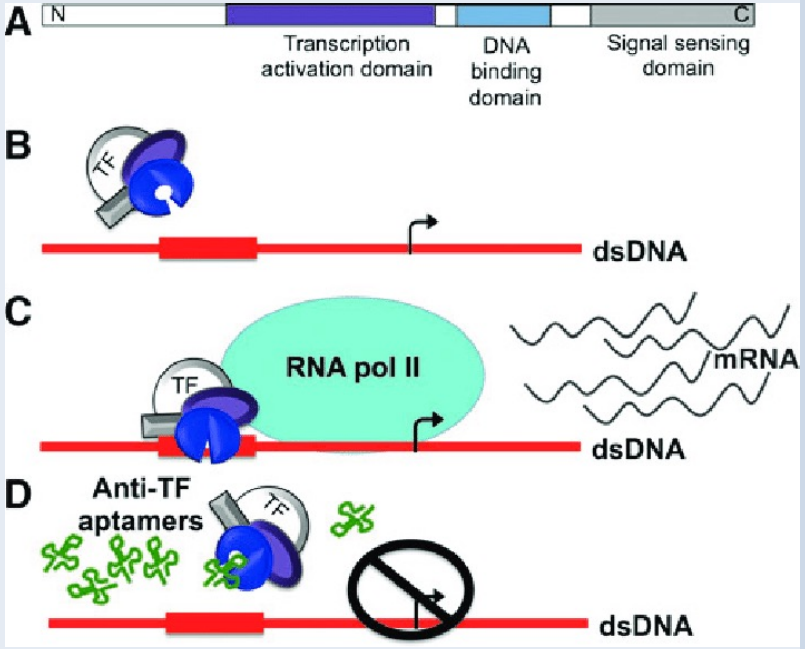
Histone **modification** levels are predictive for **gene expression**
R Karlić, HR Chung, J Lasserre, K Vlahoviček… - Proceedings of the …, 2010 - pnas.org
… of histone **modifications** are necessary to accurately predict **gene expression**. We show that different sets of histone **modifications** are necessary to predict **gene expression** driven by …
☆ Save 99 Cite Cited by 946 Related articles All 16 versions

- **Removing DNA Sequences : CRISPR-Cas9**
- **Attaching DNA Sequences : PCR**

Gene Modification

- Expression of genes
  - transcription factors [TFs]
  - Antagonists
  - consolidator
    - transcriptional co-factors
    - chromatin remodelers

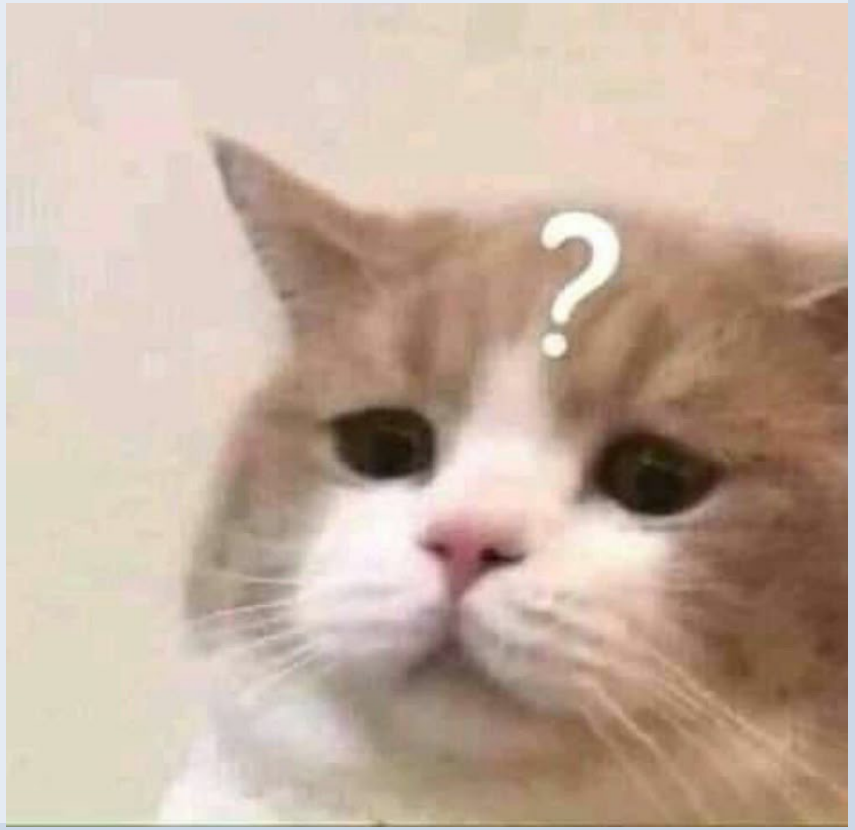




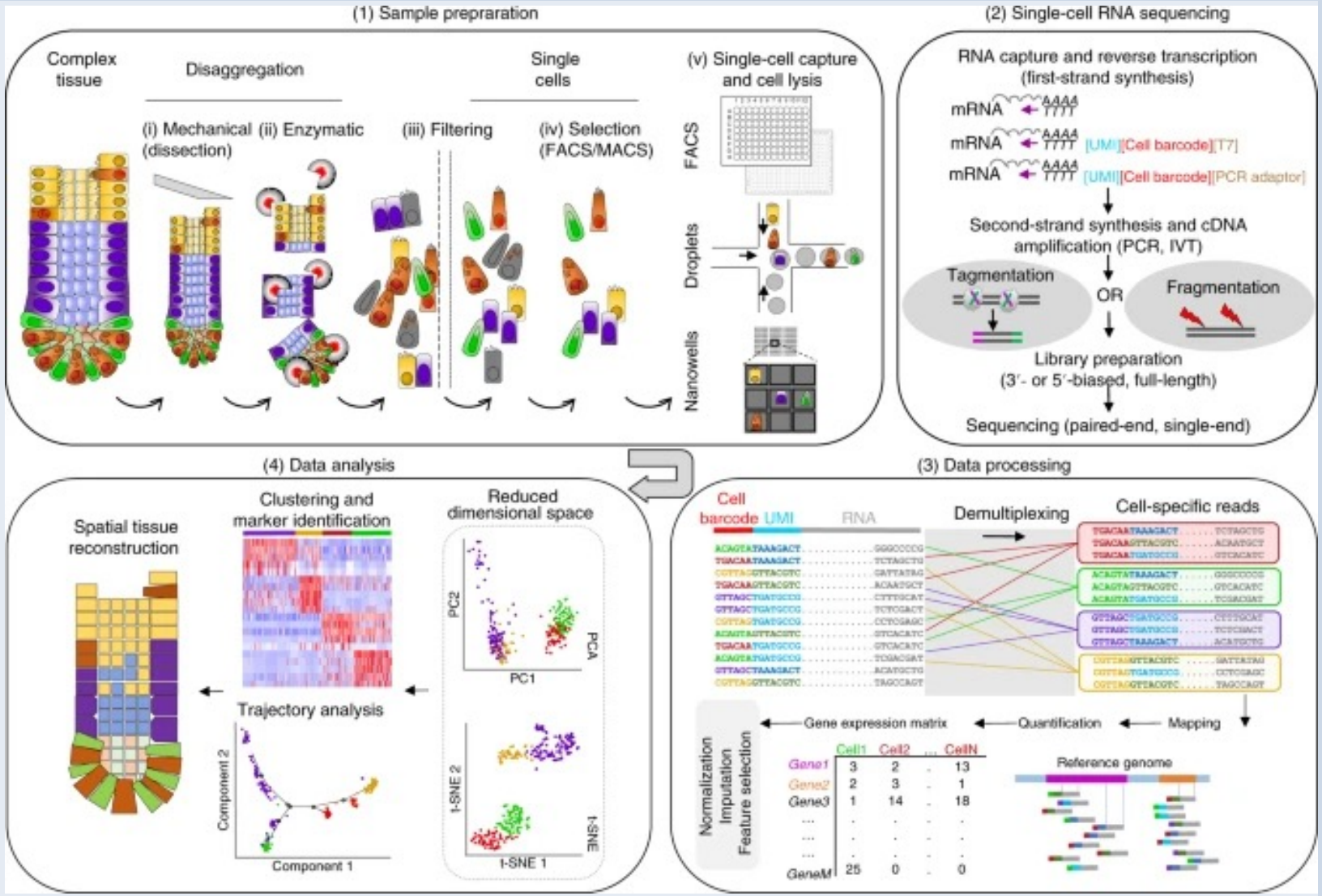
?

- How can we determine the expression of genes in a cell?
- The answer is:

Single-cell RNA Sequencing Data



sc-RNA seq


(1) Sample prepraration
Complex tissue
Disaggregation
Single cells
(i) Mechanical (dissection)
(ii) Enzymatic
(iii) Filtering
(iv) Selection (FACS/MACS)
(v) Single-cell capture and cell lysis
FACS
Droplets
Nanowells
(2) Single-cell RNA sequencing
RNA capture and reverse transcription (first-strand synthesis)
mRNA
[UMI][Cell barcode][T7]
[UMI][Cell barcode][PCR adaptor]
Second-strand synthesis and cDNA amplification (PCR, IVT)
Tagmentation
OR
Fragmentation
Library preparation (3′- or 5′-biased, full-length)
Sequencing (paired-end, single-end)
(3) Data processing
Cell barcode
UMI
RNA
Demultiplexing
Cell-specific reads
Mapping
Quantification
Gene expression matrix
Cell1
Cell2
CellN
Gene1
Gene2
Gene3
GeneM
Reference genome
Normalization
Imputation
Feature selection
(4) Data analysis
Spatial tissue reconstruction
Clustering and marker identification
Reduced dimensional space
PC1
PC2
PCA
Trajectory analysis
Component 1
Component 2
t-SNE 1
t-SNE 2
t-SNE

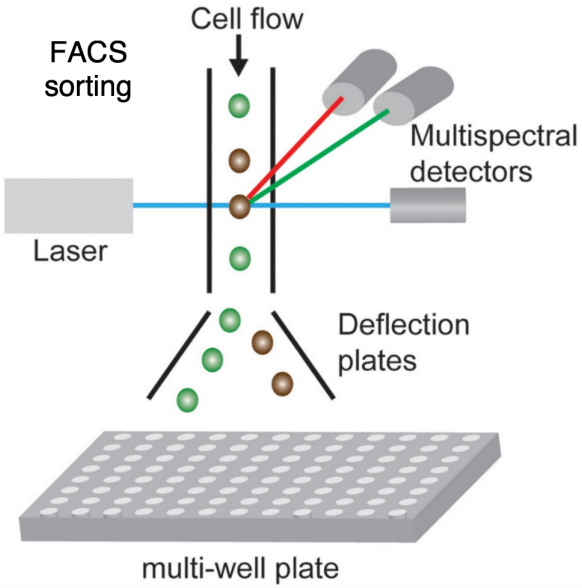
Sample preparation

- Steps:
  - Tissue dissection and cell dissociating to obtain a suspension of cells.
  - Optionally cells may be selected (e.g. based on membrane markers, fluorescent transgenes or staining dyes).
  - Capture single cells into individual reaction containers (e.g. wells or oil droplets).
  - Extracting the RNA from each cell.

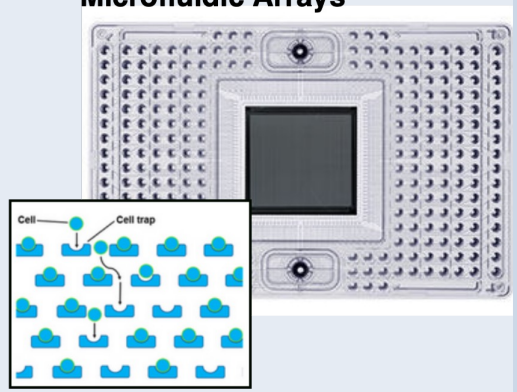
Cell Capture

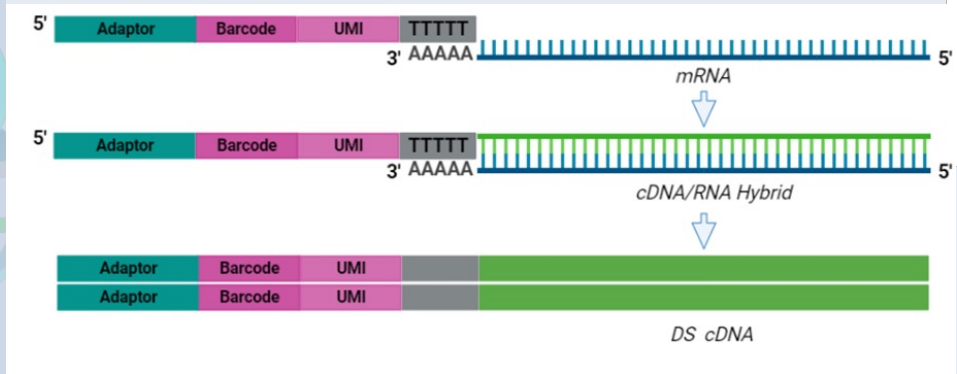
- Microtitre plate based,
- Microfluidic array based
- Microfluidic droplet based

Microtitre Plates

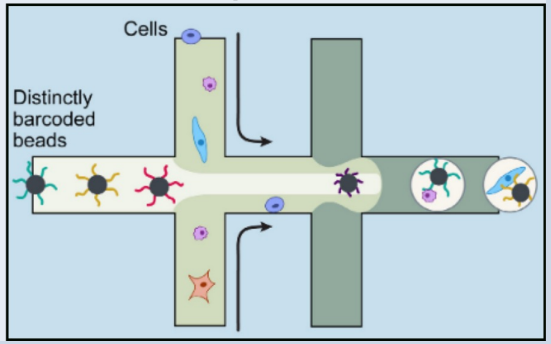


Microfluidic Arrays



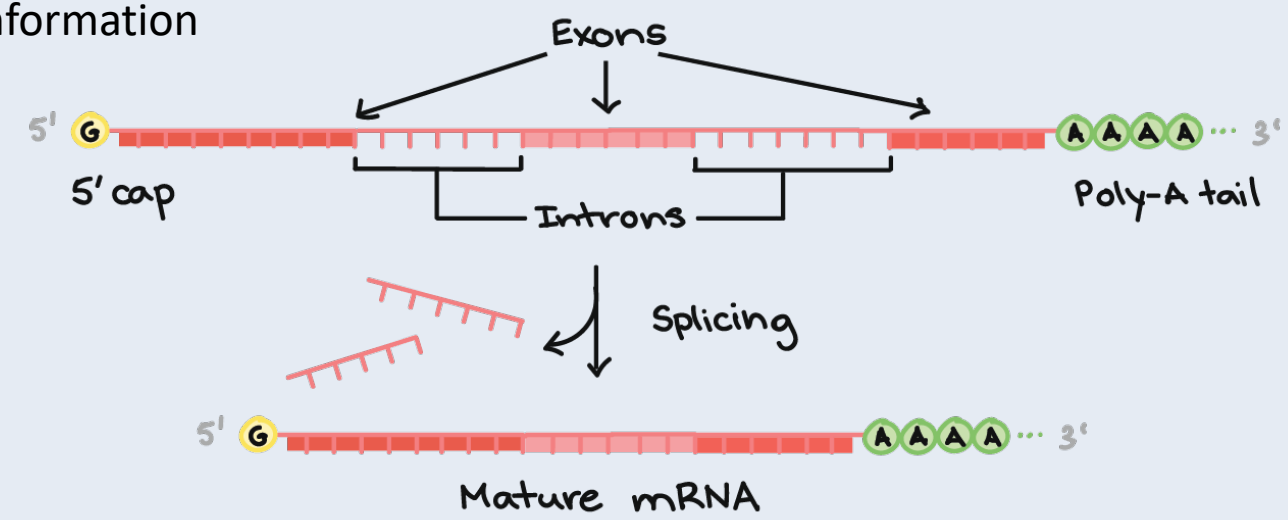


Microfluidic Droplets



Sequencing

- full-length
  - Captures the entire transcript,
  - from the 5' end to the 3' end.
  - more informative but expensive and lower throughput.
- tag-based
  - Only captures either the 5' end or the 3' end
  - Cheaper and more scalable
  - Loses information



Raw data

- **Raw data from sequencing**
- FASTQ format:
  - The first line begins with '@' followed by a sequence identifier.
  - The second line contains the actual nucleotide sequence.
  - The third line starts with a '+' sign and optionally contains the same identifier.
  - The fourth line encodes the quality scores corresponding to the nucleotide sequence.

```
@SEQ_ID_1
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SEQ_ID_2
AGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTA
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SEQ_ID_3
TTTAAAGGGCCCTTTAAAGGGCCCTTTAAAGGG
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SEQ_ID_4
CGTACGTACGTAGCTAGCGTGACGTAGCTAGCT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Raw data processing

```
@SEQ_ID_1
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

```
read_1  99  chr1  100  60  4M  5S  =  200  300  ACGTACGT  IIIIIIII
```

Count Matrix

- **From Reads to a Count Matrix**

Matrix is cells x genes

Needs to be filtered:

- gene3 - all zeros
- gene5 - mostly zeros
- cell3 - failed/rare cell
- cell5 - failed/overamplified cell

	cell1	cell2	cell3	cell4	cell5	...	cellM
gene1	93	25	0	52	3335		82
gene2	5	2	0	3	1252		12
gene3	0	0	0	0	0		0
gene4	98	21	1	1	5318		75
gene5	0	0	0	0	50		0
...							
geneN	22	52	0	31	4313		63

Count Matrix

- **From Reads to a Count Matrix**

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

Quality Control

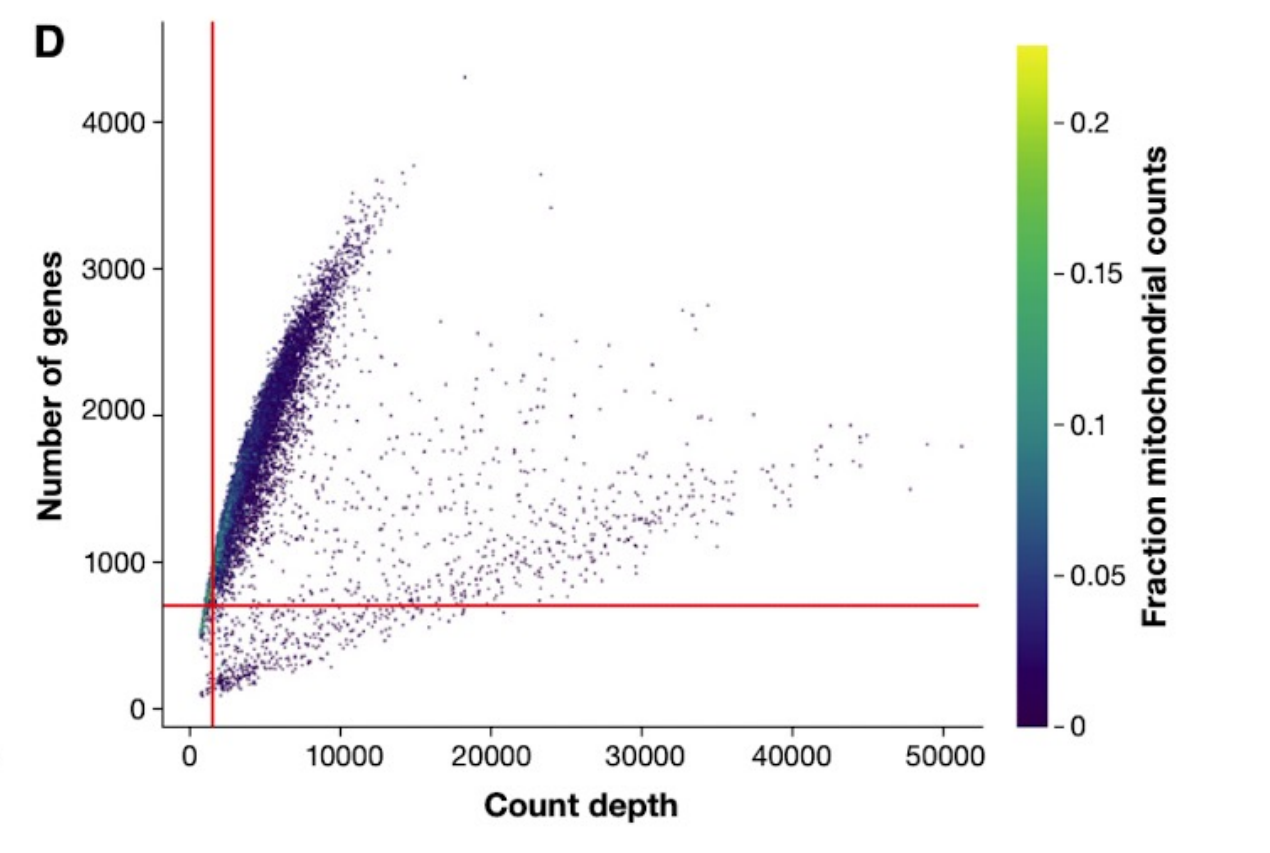
- Three main key factors : QC covariates
- Count depth
  - Low count depth ➡️ damaged or broken cells
  - Very high count depth ➡️ doublets
  - Threshold ➡️ 500, 20,000
- The number of genes per barcode
  - Low ➡️ dead cells or empty droplets
  - Very high ➡️ doublets
- The fraction of counts from mitochondrial genes per barcode
  - Very high ➡️ dying or stressed cells, broken cells
  - A normal mitochondrial fraction for a healthy cell is typically less than 5–10%

Each row is a gene

Each column is a sample

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

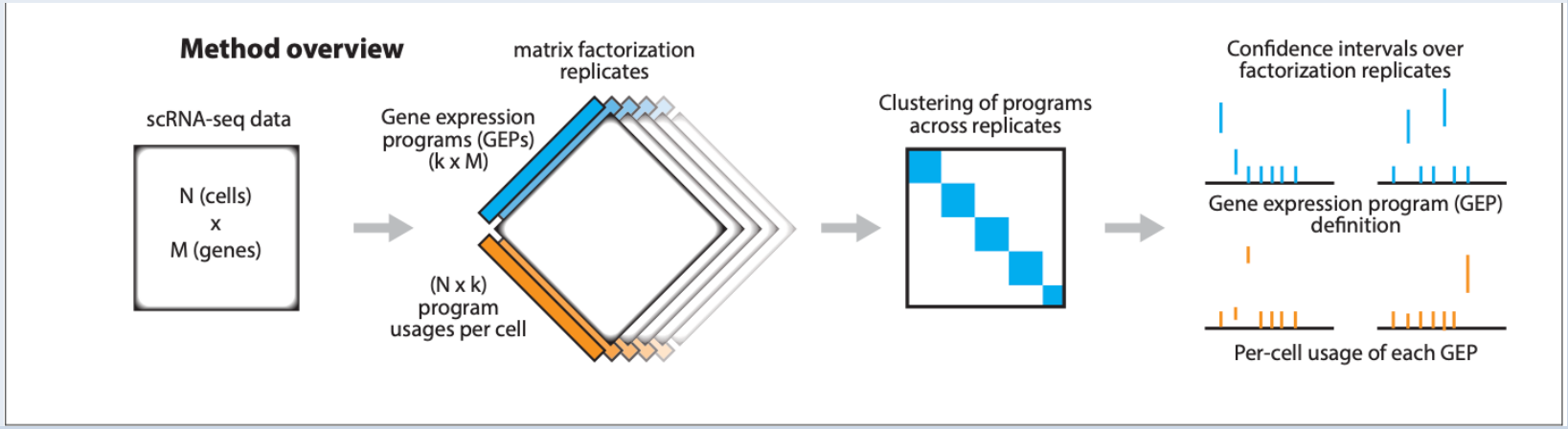
Quality Control

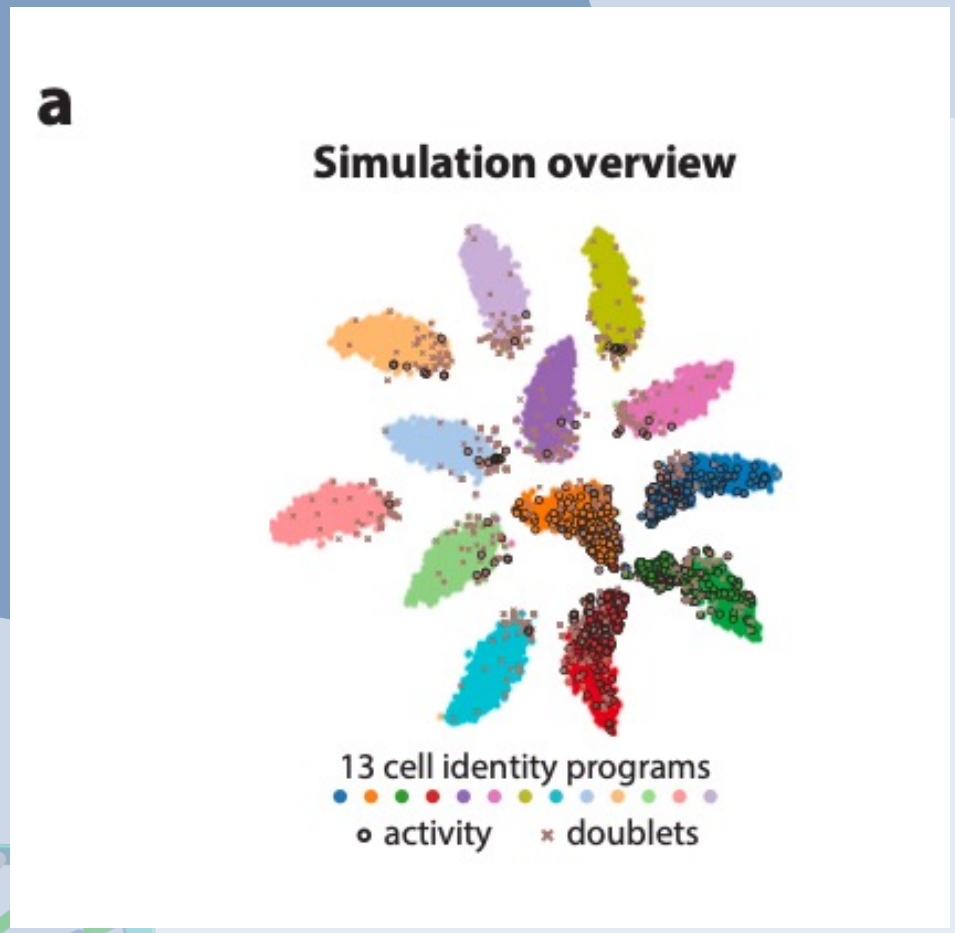


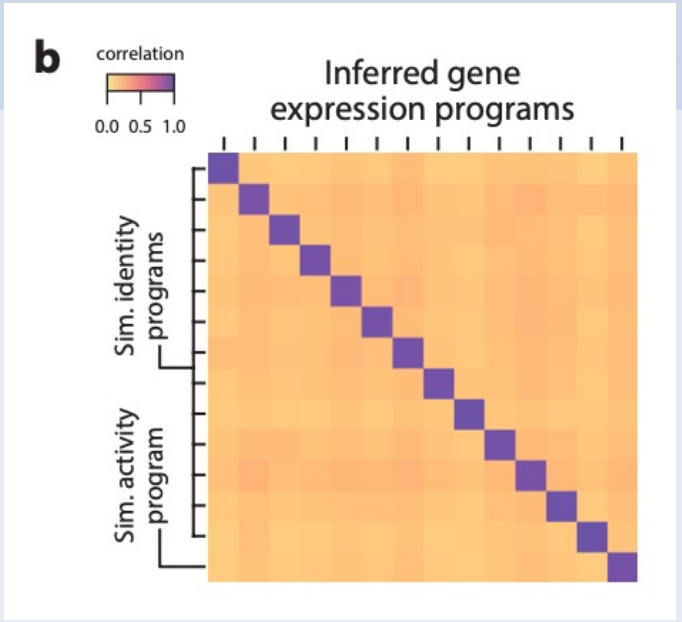
Papers

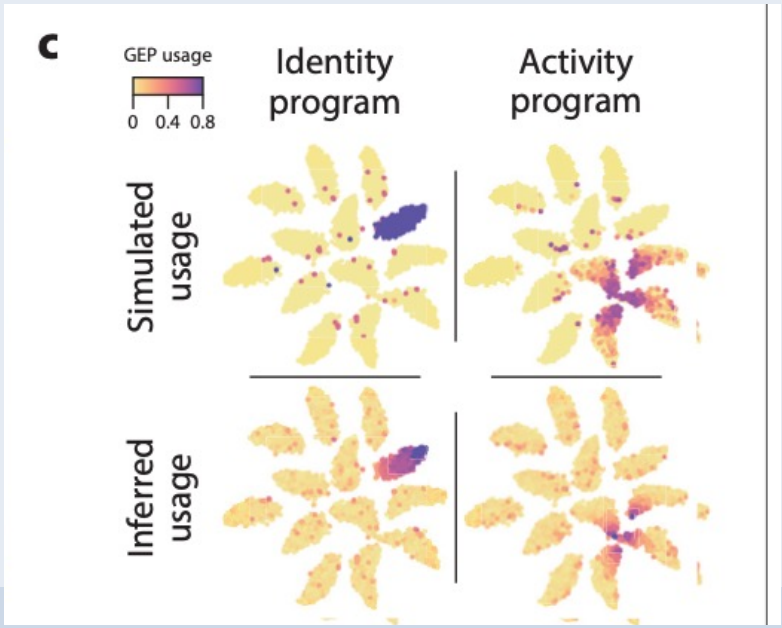
- Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq
  - Dylan Kotliar1,2,3†*, Adrian Veres1,3,4†, M Aurel Nagy3,5, Shervin Tabrizi2, Eran Hodis3,6, Douglas A Melton4,7, Pardis C Sabeti1,2,7

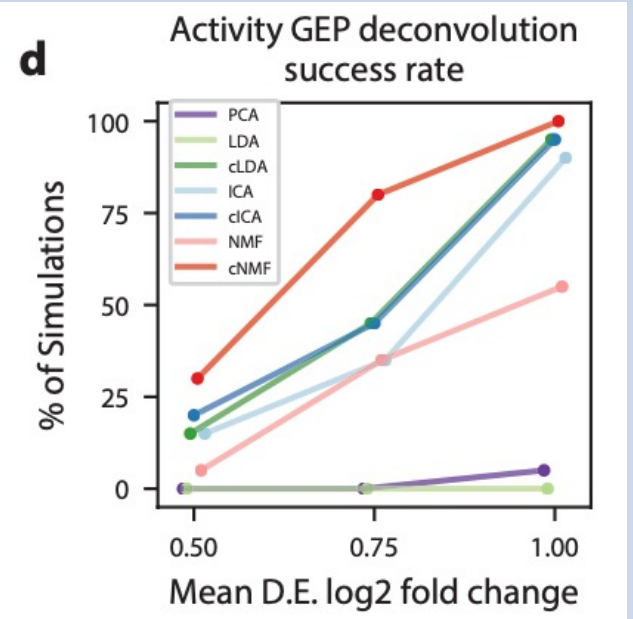
- cNMF (consensus non-negative matrix factorization)
- Activity Program
- Identity program



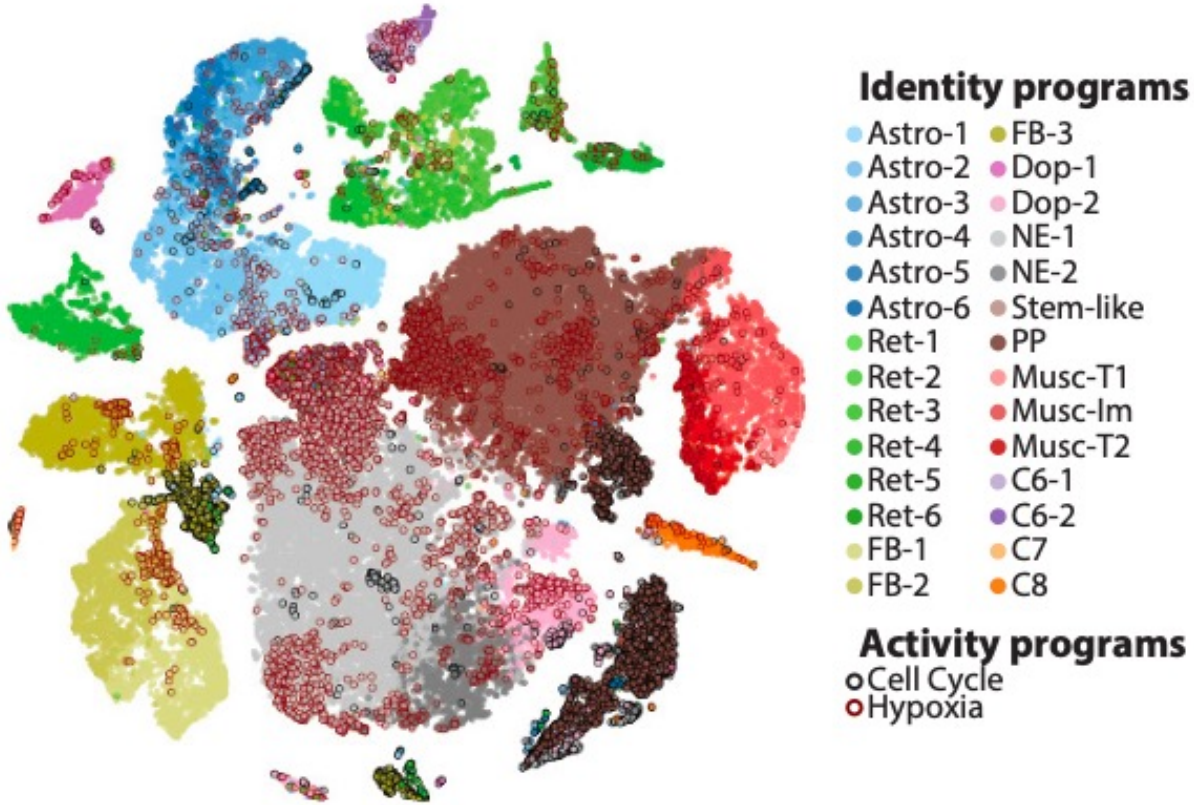

a
Simulation overview
13 cell identity programs
activity
doublets


b
correlation
0.0 0.5 1.0
Inferred gene expression programs
Sim. identity programs
Sim. activity program


c
GEP usage
0 0.4 0.8
Identity program
Activity program
Simulated usage
Inferred usage

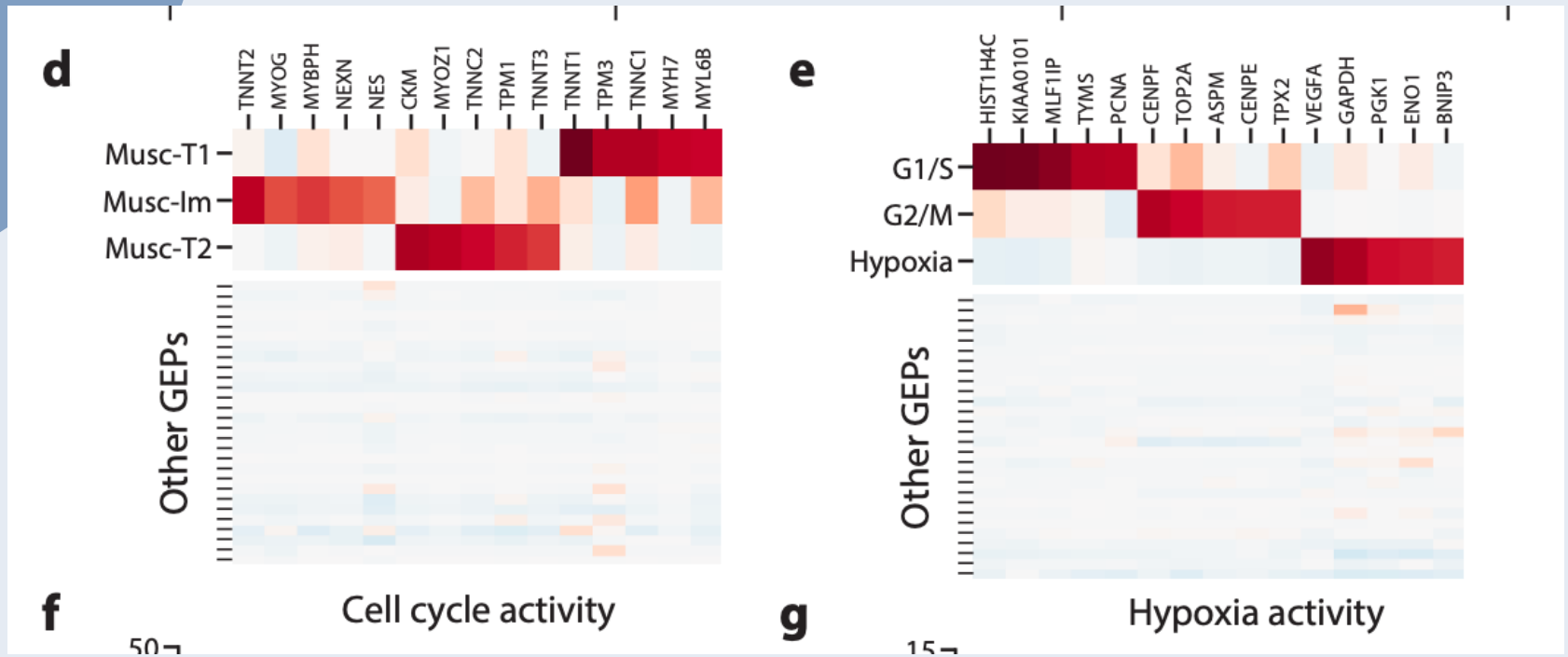

d
Activity GEP deconvolution success rate
PCA
LDA
cLDA
ICA
clCA
NMF
cNMF
% of Simulations
100
75
50
25
0
0.50
0.75
1.00
Mean D.E. log2 fold change

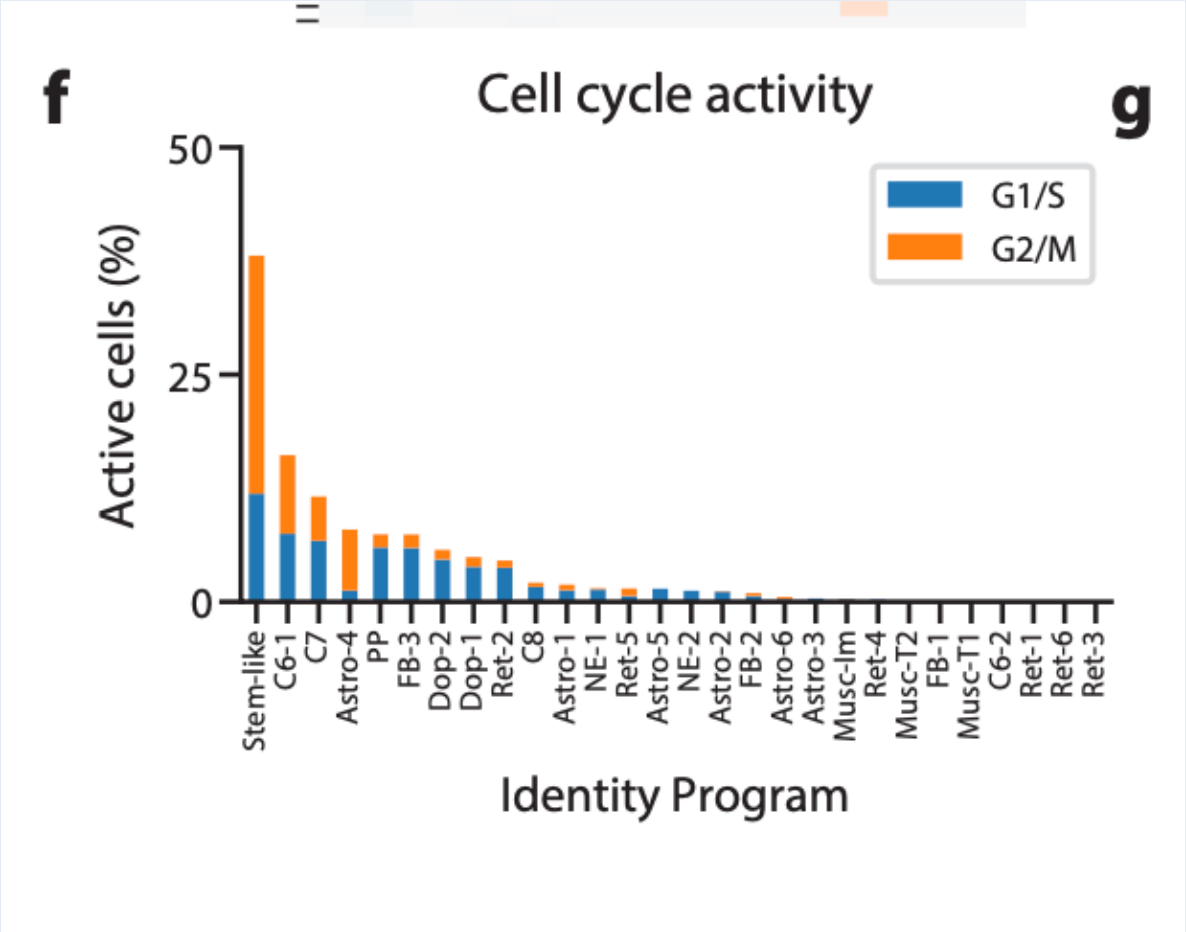
Organoid cell-types and activities



C

	G1/S Activity Program		G2/M Activity Program		Hypoxia Activity Program	
rank	GO term	p-val	GO term	p-val	GO term	p-val
1	Cell Cycle	$2x10^{-84}$	Mitotic Cell Cycle	$3x10^{-77}$	Establish. of Protein Loc. To Endoplasmic Reticulum	$3x10^{-37}$
2	Mitotic Cell Cycle	$8x10^{-81}$	Cell Cycle Process	$2x10^{-68}$	Protein Loc. To Endoplasmic Reticulum	$2x10^{-35}$
3	Cell Cycle Process	$3x10^{-77}$	Cell Cycle	$6x10^{-64}$	Translation Initiation	$6x10^{-32}$
4	Chromosome Organization	$3x10^{-70}$	Mitotic Nuclear Division	$6x10^{-61}$	Nuclear Transcribed mRNA Catabolic Process NMD	$4x10^{-31}$
5	DNA Metabolic Process	$4x10^{-69}$	Organelle Fission	$4x10^{-53}$	rRNA Metabolic Process	$8x10^{-28}$
6	DNA Repair	$5x10^{-55}$	Sister Chromatid Division	$1x10^{-47}$	Ribosome Biogenesis	$2x10^{-27}$


d
TNNT2
MYOG
MYBPH
NEXN
NES
CKM
MYOZ1
TNNC2
TPM1
TNNT3
TNNT1
TPM3
TNNC1
MYH7
MYL6B
Musc-T1
Musc-Im
Musc-T2
Other GEPs
e
HIST1H4C
KIAA0101
MLF1IP
TYMS
PCNA
CENPF
TOP2A
ASPM
CENPE
TPX2
VEGFA
GAPDH
PGK1
ENO1
BNIP3
G1/S
G2/M
Hypoxia
Other GEPs
f
Cell cycle activity
g
Hypoxia activity


f
Cell cycle activity
g
G1/S
G2/M
Active cells (%)
50
25
0
Stem-like
C6-1
C7
Astro-4
PP
FB-3
Dop-2
Dop-1
Ret-2
C8
Astro-1
NE-1
Ret-5
Astro-5
NE-2
Astro-2
FB-2
Astro-6
Astro-3
Musc-Im
Ret-4
Musc-T2
FB-1
Musc-T1
C6-2
Ret-1
Ret-6
Ret-3
Identity Program

Next session