生物資訊系列 NGS Workshop 5:
Differential Analysis of Transcriptomics

講師：張耀明 博士

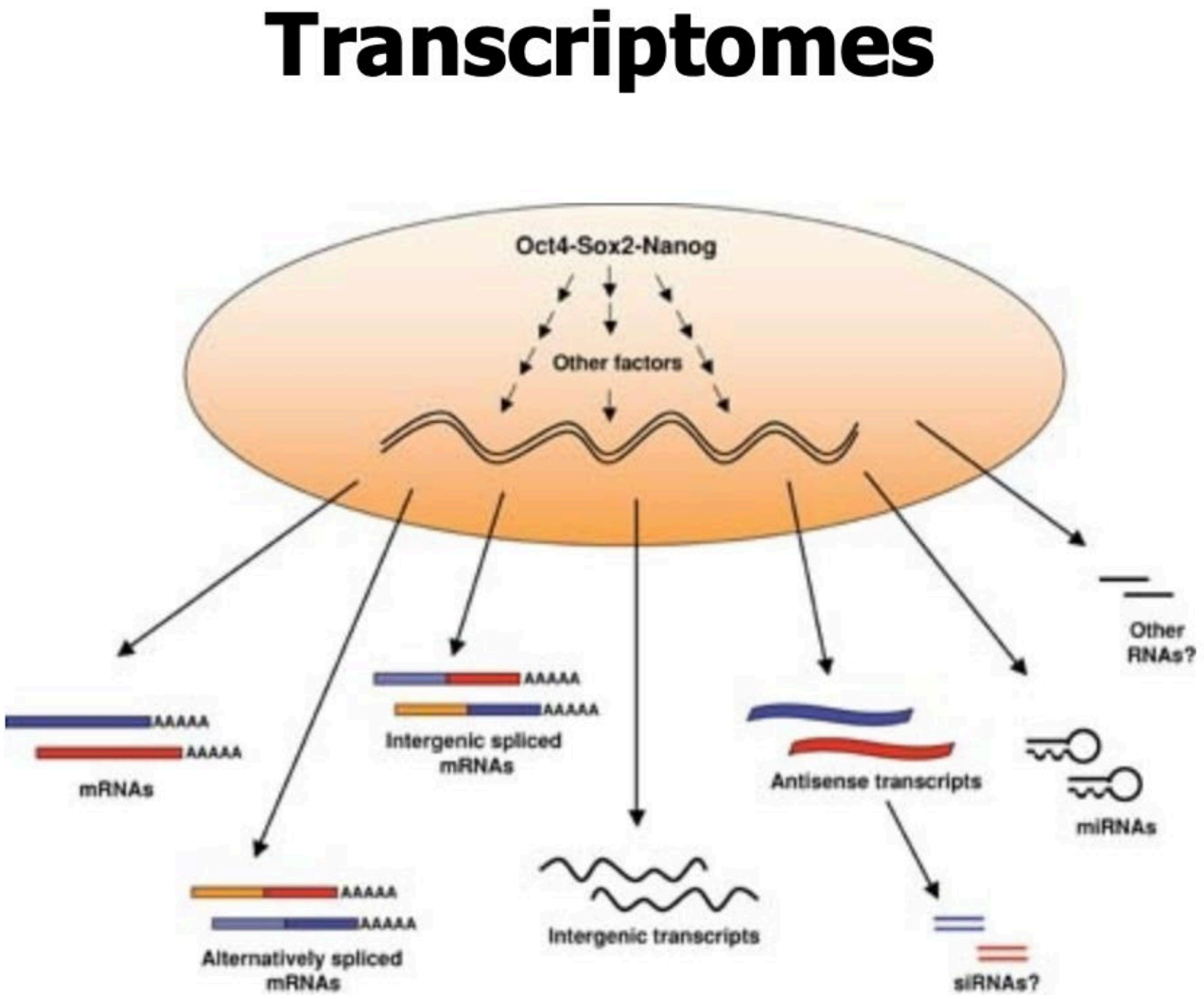
研究副技師
中研院 生醫所 計算醫學核心實驗室

Nov. 6, 2025

Core Questions

1. **What is Transcriptomics?**
2. Where is Transcriptomics in the Omics Universe?
3. Why is Transcriptomics so important?
4. Who can bridge Transcriptomics and biological research?
5. How to use Transcriptomics in biological research?

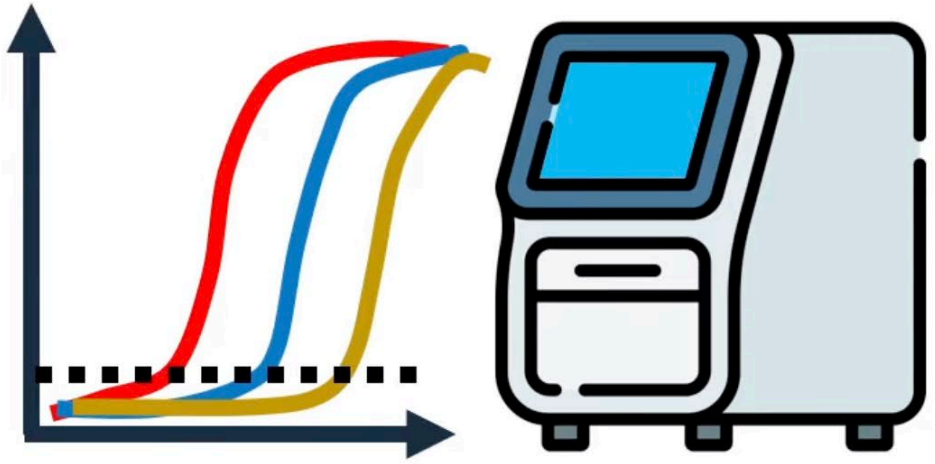
What is Transcriptomics?



Transcriptomic revolution

Timeline

qRT-PCR

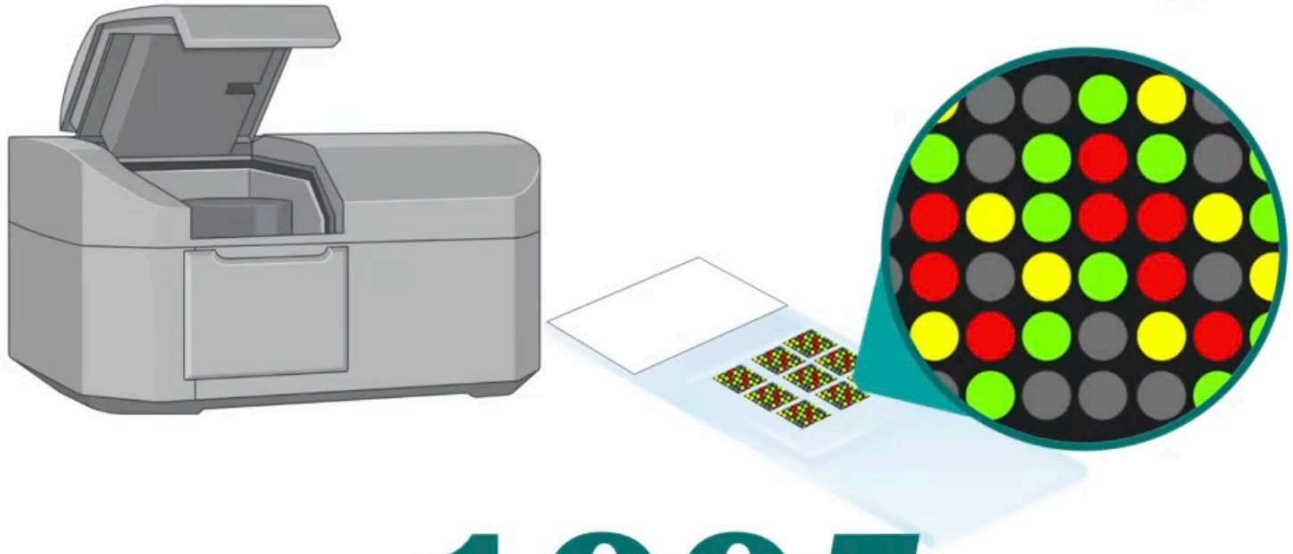


1994

Transcriptomic revolution

Timeline

Micro Array

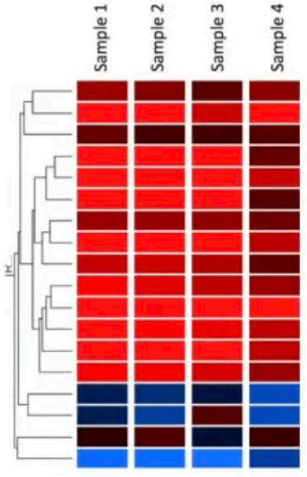


1995

Transcriptomic revolution

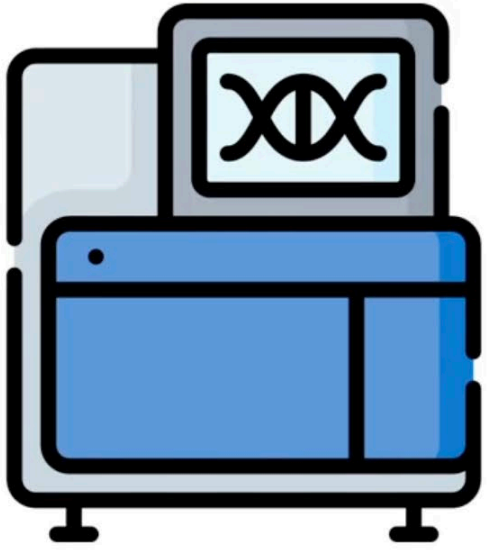
Timeline

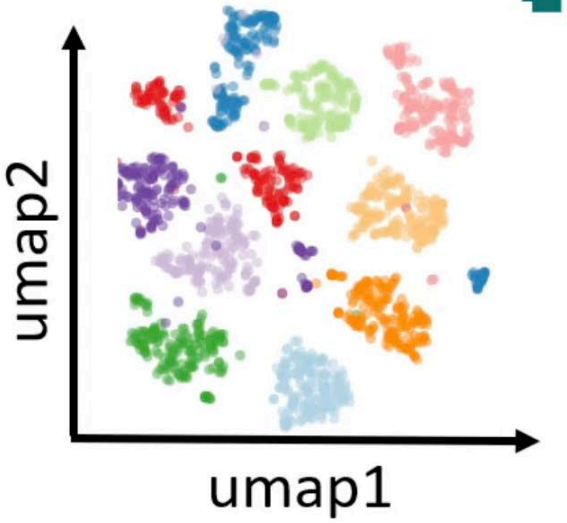
Bulk RNA Seq



2006

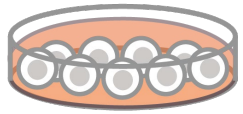
Sc RNA seq





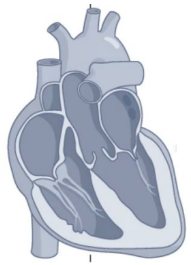
2009

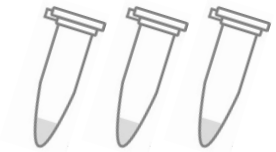
What is Transcriptomics?



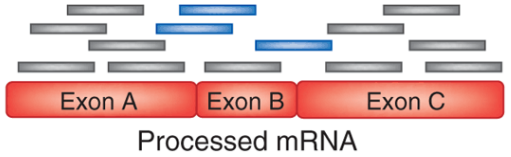
RNA

Sequencing





Sequencing reads alignment

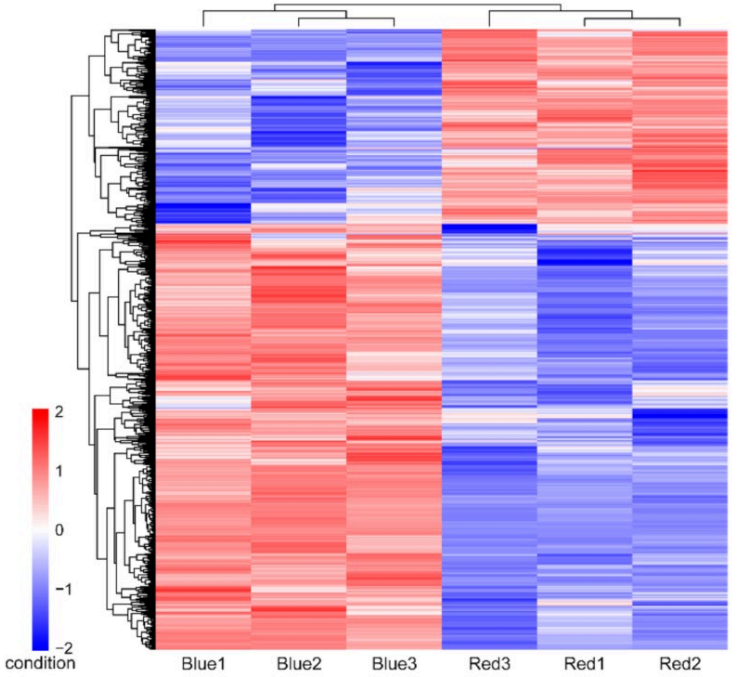


Data analysis

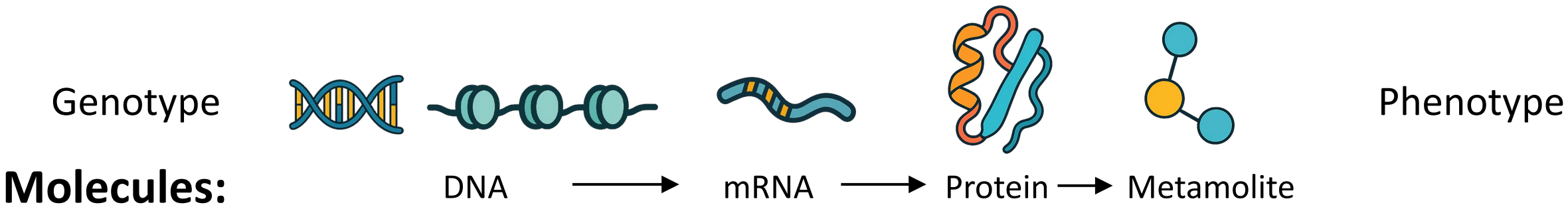
transcriptomes x samples

Acc ID	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
NM_007818	67540.89	70924.09	80243.76	3501.2	5697.47	2426.72
NM_001105160	811.93	801.36	740.71	128.67	104.42	101.33
NM_028089	190.41	211.06	236.19	9.05	23.33	8.44
NM_016696	66.77	57.56	101.09	750.9	659.84	491.89
NM_013459	3.3	11.29	1.89	735.82	816.46	118.22
NM_007809	45.34	36.12	51.02	245.27	372.13	335.67
NM_009999	103.04	370.21	200.29	17.09	13.33	8.44
NM_133960	7708.78	6976.38	6569.04	1731	1641.81	1853.55
NM_027881	31.32	10.16	24.56	268.39	186.62	135.11
NM_054053	31.32	24.83	19.84	323.68	428.78	116.11
NM_007377	47.81	89.17	70.86	370.93	378.79	279.72
NM_028064	703.95	689.62	662.29	214.11	168.85	144.61
NM_008182	222.56	339.73	226.75	30.16	63.32	26.39
NM_013661	12.36	11.29	8.5	97.51	77.76	71.78
NM_007815	20613.09	25218.13	31540.46	5209.07	7680.3	6312.2

Transcriptomic data analysis



Where is Transcriptomics in Omics Universe?



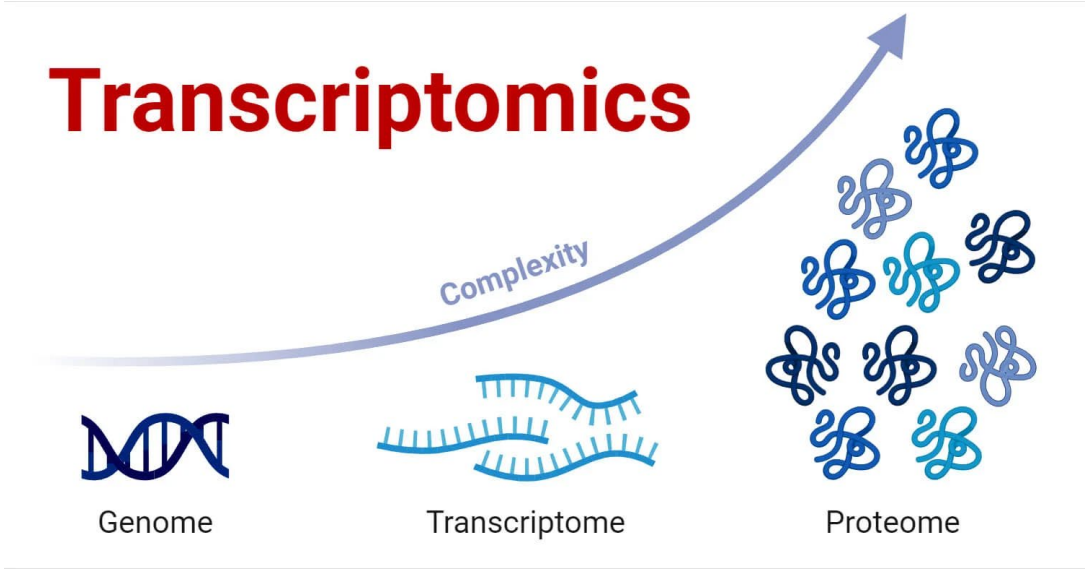
Omics:

Genomics | Epigenomics | **Transcriptomics** | Proteomics | Metabolomics

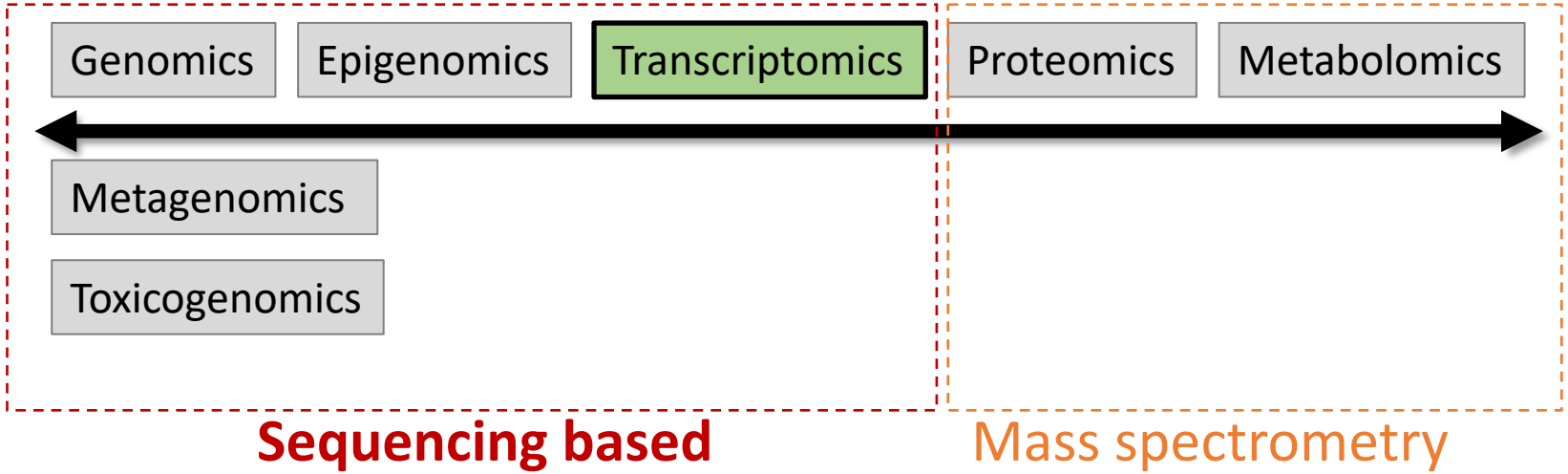
Metagenomics

Toxicogenomics

Why is Transcriptomics so important?

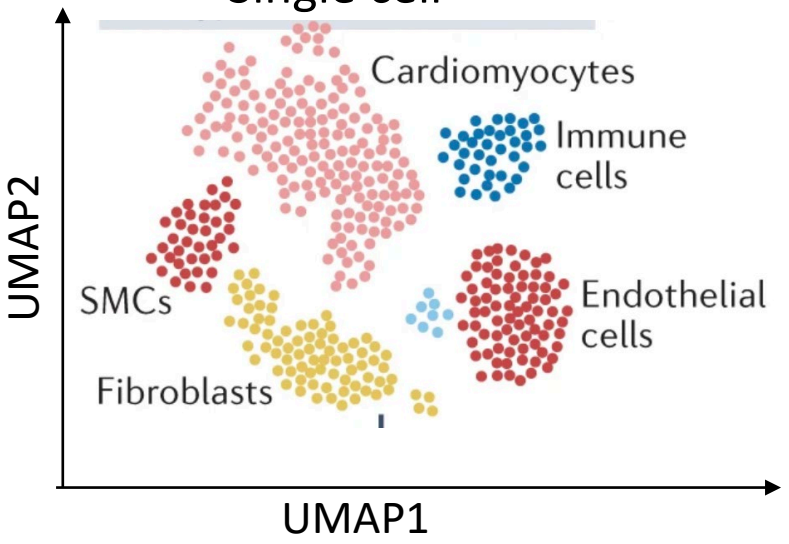


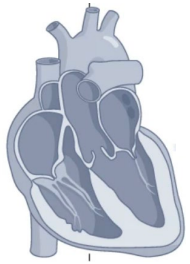
Omics:



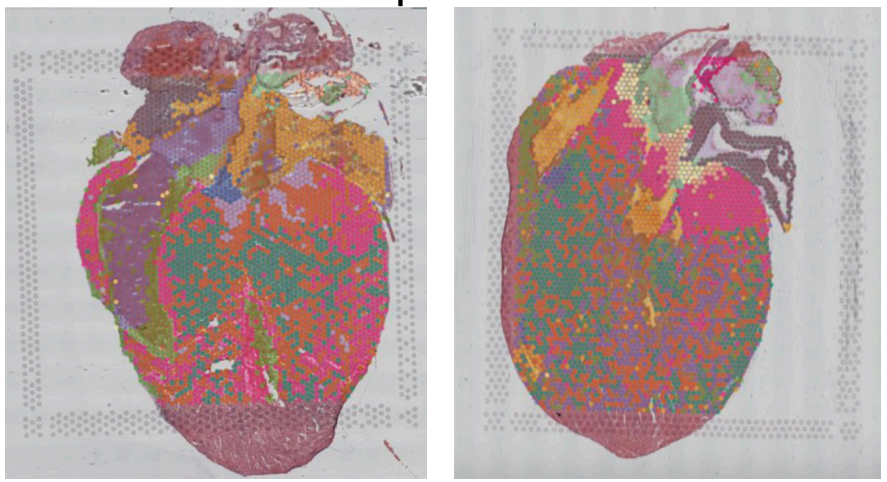
Why is Transcriptomics so important?

Single cell





Spatial



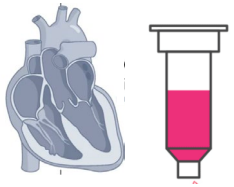
resolution

Omics:

Genomics | Epigenomics | **Transcriptomics** | Proteomics | Metabolomics

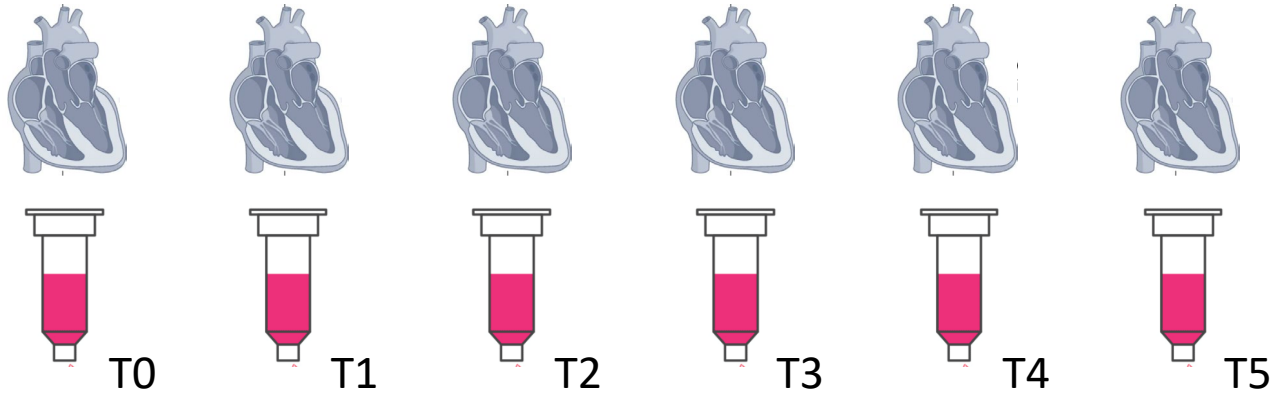
Metagenomics

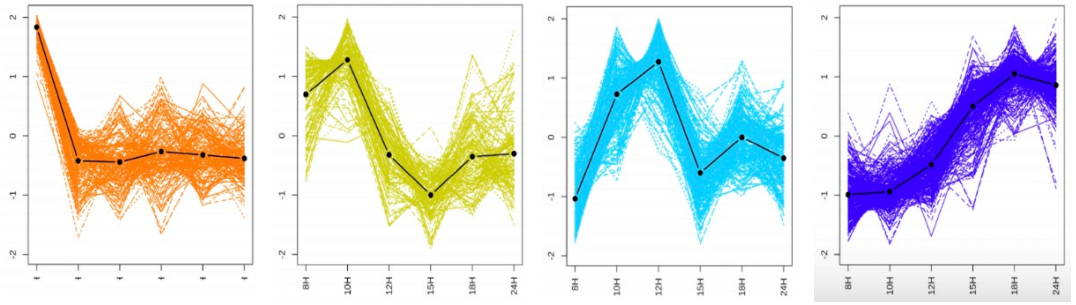
Toxicogenomics



bulk

Why is Transcriptomics so important?





Time course

Omics:

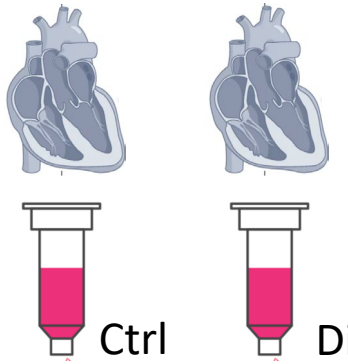
Genomics | Epigenomics | **Transcriptomics** | Proteomics | Metabolomics

Metagenomics

Toxicogenomics

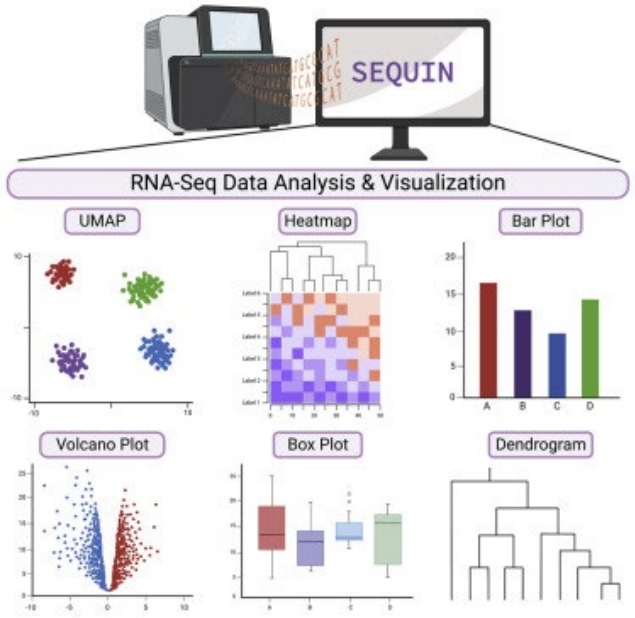
experiment design

Ctrl vs. Disease

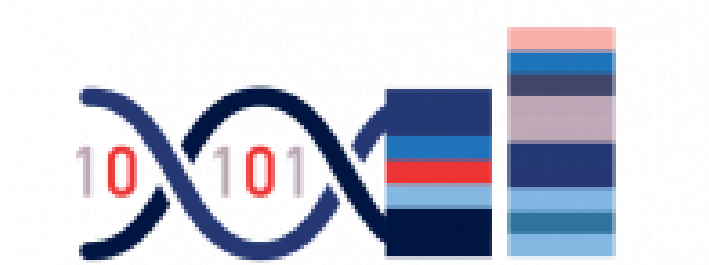


Who can bridge Transcriptomics & biological research?

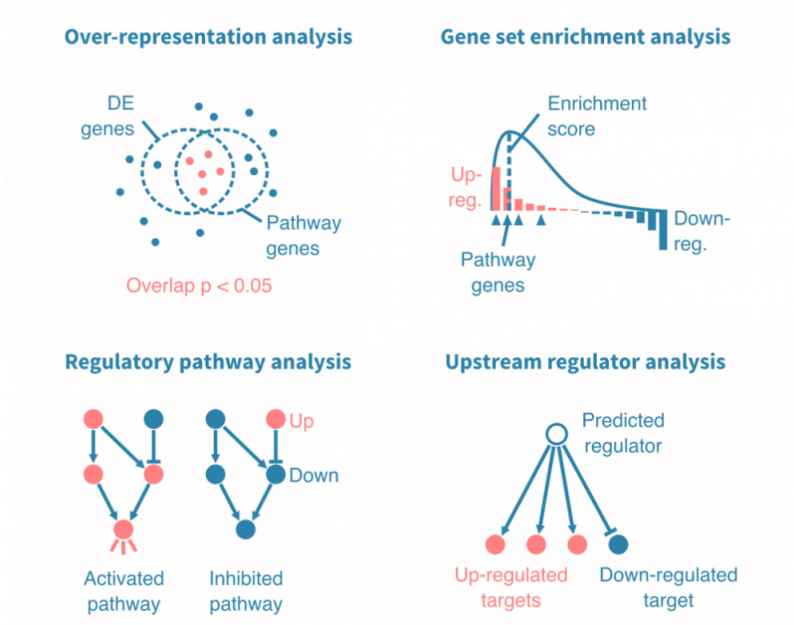
Transcriptomics ⟷ **Biological research**



Computational Analysis



Computational Biology Core



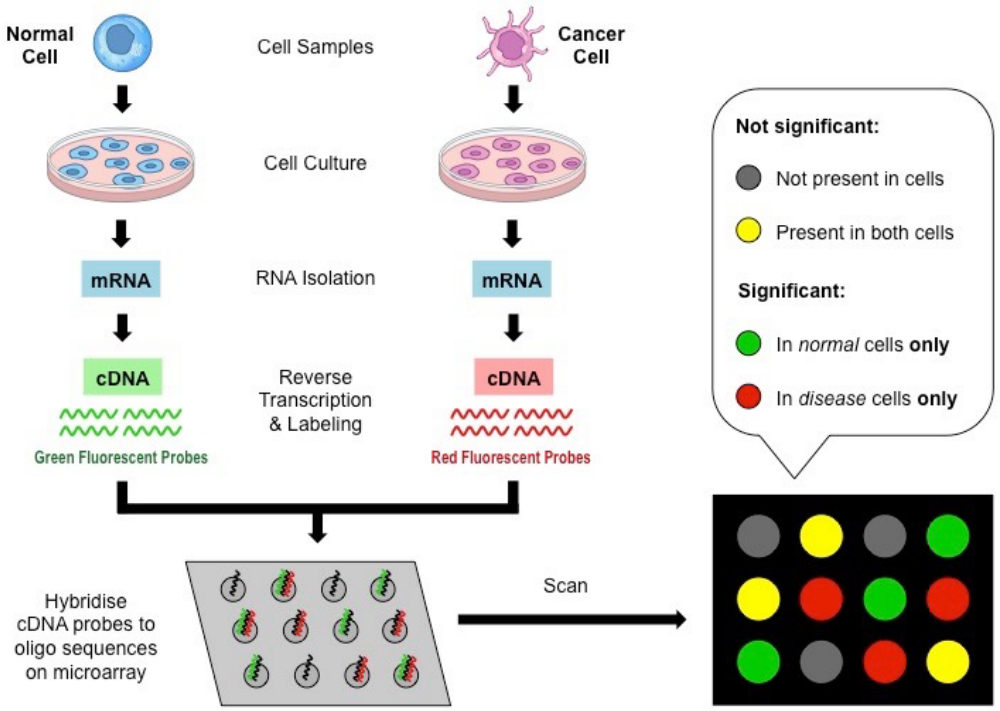
How to use Transcriptomics in biological research?

- Differential Gene Expression (DGE) Analysis
- Longitudinal/Time-Series RNA-seq
- Pathway Enrichment & Network Analysis
- Single-cell RNA-seq (scRNA-seq or snRNA-seq)
- Spatial Transcriptomics
- Cross-Species Analysis
- Multi-omics Integration Analysis
- Transcriptome-Wide Association Studies (TWAS)
- scGPT: single cell Transcriptomics with generative AI

Bulk RNA-seq data analysis

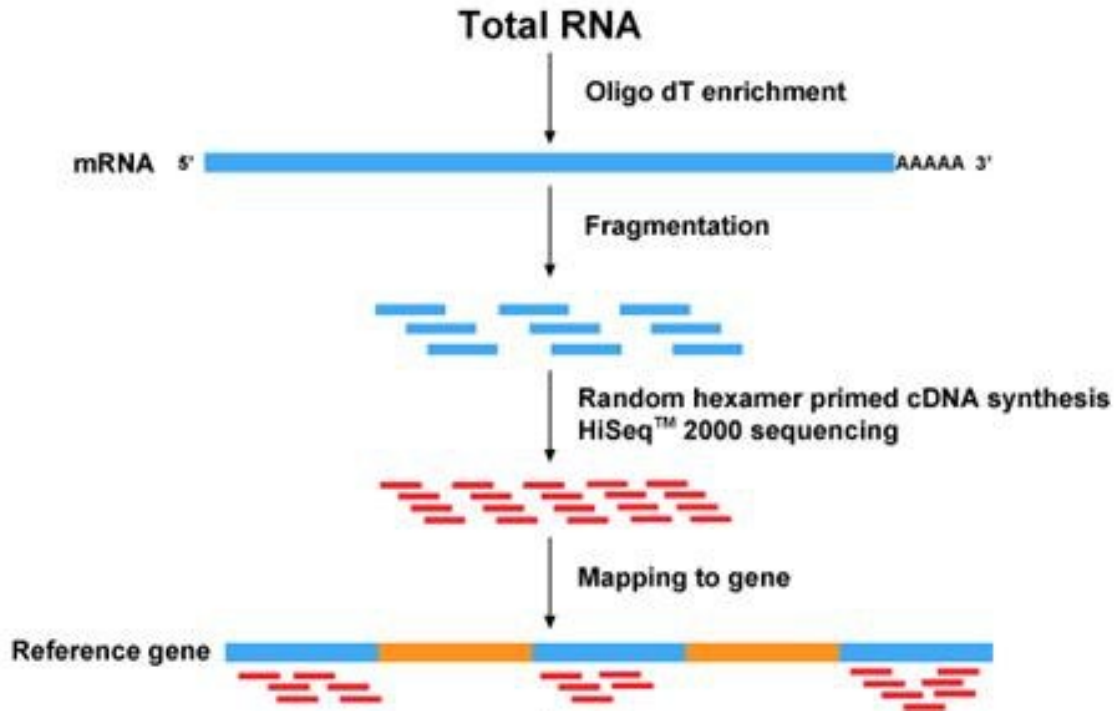
How do we get the transcriptome data?

- Microarray



intensity from microarray (since 1981)

- RNA sequencing (RNA-seq)



read counts from RNA-seq (since 2011)

Two major parts for RNA-seq data analysis

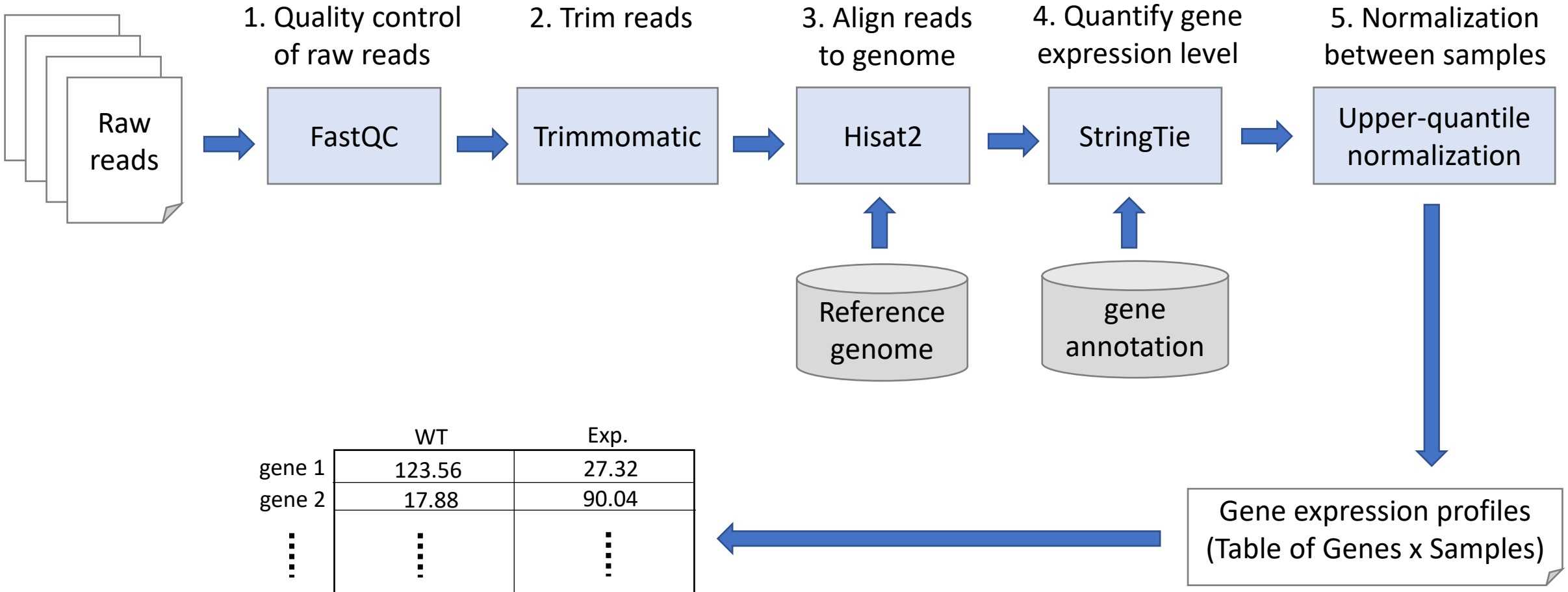
Part 1: RNA-sequencing data processing

- File size is very big (more than 1GB for a sample)
- Hard to read by human eyes
- Computing power and memory demanding

Part 2: Transcriptome (gene expression) analysis

- Various methods for different needs
- Many tools (R packages mostly) can run on personal computer

Pipeline of bulk RNA-seq data processing



Raw reads (short reads)

- Source
  - Sequencing results from your own experiments
  - Download from NCBI (GEO: gene expression omnibus or SRA: sequence read archive)

- Data format: FASTQ = sequence + quality score
  - Pair-end (mate-pair): SampleID_R1.fastq, SampleID_R2.fastq
  - Single-end: SampleID.fastq

- Sample description file

PI / Resource		Case ID	20210202-S2	Sequencer	HiSeq / D00368 (HS_Y)
Institution	IBMS	Application Type	(S-Ta) Stranded RNA Lib Prep, Ribo	Read Length	PR 201_dual10

Sample Name	Sample ID	Library ID	Barcode Index	Seq Date	Dataset ID	Lane ID	Yield (Mb)	# of Clusters
4DC-Hi-1	ST20-VI01	LTT21-VI01	ACGGTCCAAC+TCTTACATCA	3/17/21	sd485B	1	6,338	15,767,125
4DT-Hi-1	ST20-VI02	LTT21-VI02	GATCGTCGCG+CTGGATATGT	3/17/21	sd485B	1	6,275	15,610,036
8DT-Hi-1	ST20-VI03	LTT21-VI03	GTAACTTGGT+CGCCATACCT	3/17/21	sd485B	1	6,135	15,261,080
8DT-Low-1	ST20-VI04	LTT21-VI04	AGGCGTTCGC+TTCATCCAAC	3/17/21	sd485B	1	5,526	13,747,268
4DC-Hi-2	ST20-VI05	LTT21-VI05	AGTACTCATG+ATGTCGTGGT	3/17/21	sd485B	1	5,960	14,825,928
4DT-Hi-2	ST20-VI06	LTT21-VI06	CCGTGACCGA+CCGAACGTTG	3/17/21	sd485B	1	6,605	16,430,128
8DT-Hi-2	ST20-VI07	LTT21-VI07	GACGAGATTA+ACATTATCCT	3/17/21	sd485B	1	6,791	16,894,237
8DT-Low-2	ST20-VI08	LTT21-VI08	GACTGAGTAG+GTTGATAGTG	3/17/21	sd485B	1	6,554	16,303,329
4DC-Hi-3	ST20-VI09	LTT21-VI09	AGTCAGACGA+ACCAGCGACA	3/17/21	sd485B	1	6,619	16,466,240
4DT-Hi-3	ST20-VI10	LTT21-VI10	CCGTATGTTC+CATACACTGT	3/17/21	sd485B	1	6,357	15,813,810
8DT-Hi-3	ST20-VI11	LTT21-VI11	GAGTCATAGG+GTGTGGCGCT	3/17/21	sd485B	1	6,121	15,225,820
8DT-Low-3	ST20-VI12	LTT21-VI12	CTTGCCATTA+ATCACGAAGG	3/17/21	sd485B	1	6,014	14,960,676

Raw reads files (FASTQ format)

```
1  @SEQ_ID
2  GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
3  +
4  !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Line 2: Sequencing letters

Line 4: Quality score

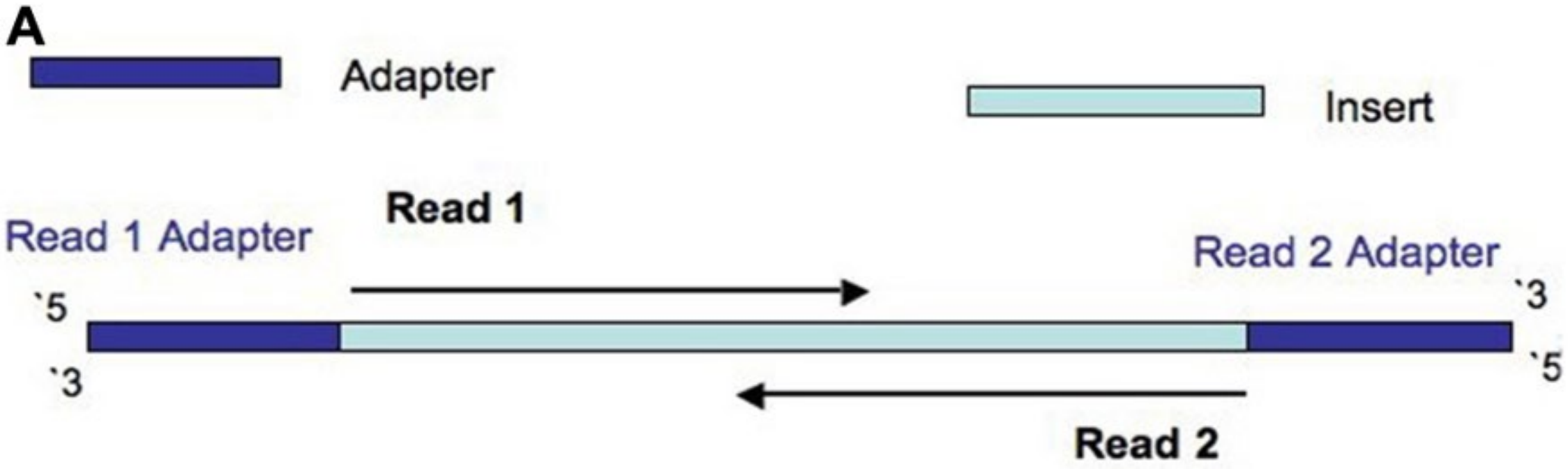
```
FFFFFFF:FFFFFFFFFFFFFFFFF:FFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:5104:1125 1:N:0:GAGATTCC+AGGCGAAG
GGCAAGTCAAAATTGATAACATGTTTCACATTTGAAATGTCCAGTCCTCTTGCTGCTACTGCTGTAGCCACTAAAATTGGGCTTTTTCCTGAGCGGAACTGGTGAAGGGCCTCTTCTCTATCCCAGATCGGAAGAGCACACGTCTGAACTC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:8250:1125 1:N:0:GAGATTCC+AGGCGAAG
AGGATCATTGCCTTTGTTTTCTGCTTTAAGACTTGGGAGGTTAGCAGGTGGAGGCATACGCCGTGAAATACCGACTTTTCCAAGACTCTGTAATCCATGTCGAGCTGCAACTGTGGTTTTCTGTGTTTCTAATGATTTCCCCTTGTAAGTA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:10908:1125 1:N:0:GAGATTCC+AGGCGAAG
GAAGTATTCAGCCACAGTCAGCCCAGTCAGAGCTACCCGGGCACGAGCACCAGGTGGTTCATTCATTTGACCATATACCAGCGCTACCTTAGAGGTGGCATCTTTTAAGTTGATAACACCAGATTCAATCATTTCATGGTATAAATCAAGA
+
FFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFF
@A00838:218:H72CYDSXY:4:1101:15501:1125 1:N:0:GAGATTCC+AGGCGAAG
GTCCGGTAGGAGACCTGAGCCTTCTGCAGCATTTCCAGGTGTGTCTCTGTTTCCTTCAGTTTCTCTTTTAACTCTTGCTTTTCCTCTTTGGTTCTTTCCAGCTCAGCTTTCAGAATCTGGAATTCAGCTTGGCGATAACCCATGTGTTTCT
+
FFFFFFFFFFFFFF:FFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFF,FFFFFF:FFFF:FFFFFFFFFFF:,FFF:FFFFFFFFF:FFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:16405:1125 1:N:0:GAGATTCC+AGGCGAAG
CACCAGGAAGAGGAAGACACCCACTGCGATGATGACCACTGGCAACAGAGAGCCAGGGGTAGCCCCCTGGATTATGGTCTGACTCAGGACAAGCTGTGCCCCGACACCCACGGCAATCAGTCCCACTGCACAGGCGCAAAAGGCCAGCAGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:19605:1125 1:N:0:GAGATTCC+AGGCGAAG
GCCTTTGACAATGTCATCAACAGACCAATTTACAGTGCCCTGGTTGTTGCGGTTTTCCTGCAGCGGAGAAGTAGCATCATCAGGAAATGAGCTTACATTTCTCCTCTTCAGCATCTGGTCATCCTTCTTAGCTTTCCTAGATCGGAAGAGC
+
FFFFFFFFFFFFFFFFFFFFFFFFFF,:F:FFFFFFFFF:FFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:19822:1125 1:N:0:GAGATTCC+AGGCGAAG
CCTTGAGCTCAGTCTCTGACTTCTCCATGATGGTCTGAAGGATAGTGGAAGTTACATCTCCATTGCAGGACCCTACTGCTATCATTCCACAGGCTAAAGCTGTGACACCTGCCACCTCCATGCTGGACTTTGAATCTCCCATCACAGGCAG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:20546:1125 1:N:0:GAGATTCC+AGGCGAAG
CTCCTGAGTAACAGCATCCTTCTCTATGCCTGCATCTTCTAACAGATGTACCTTTACTTCCTTAAAAACAGGCGTGCTTTGAGCTTTTTGCTTTAAAACTGTTCTTAGAGCATCTTTCTGTCCCATTGTATACTGAGAAATGAAGATGTCA
+
FFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFF,FF:FFFFFFFFFFFFFFFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:20943:1125 1:N:0:GAGATTCC+AGGCGAAG
GCTCATTGCCAATGGTGATGACCTGGCCATCGGGCAGCTCGTAGCTCTTCTCCAGAGAAGAGGAGGATGCGGCGGTGGCCATCTCCTGCTCGAAGTCCAGGGCGACGTAGCACAGCTTCTCCTTGATGTCGCGCACGATTTCCCGCTCGGC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00838:218:H72CYDSXY:4:1101:22516:1125 1:N:0:GAGATTCC+AGGCGAAG
CAACTGTAATCTTATTCTGGGGGGTCTGTTCTTCCTTTAGAAGATTATAAATCAGCTGATCCTTTAGAGTTGCCATATTGGACTTGGAACCAAAAGGAATCGGGAATGCACGTCGGGCGGTCGTCGGGGGCGCGTGGCAATGAGAGATCGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFF:FF
@A00838:218:H72CYDSXY:4:1101:23583:1125 1:N:0:GAGATTCC+AGGCGAAG
CCGTAATATTCAGCTCCCTGAGCTGAGCCTTGAGGTCCGAGTTCATCTCCAGCTCCAGAAGAGCCTGGGAGATGCCGGACTCGAACTCGTCCGGCTTCTCGCCATTGGGCTTCACGATCTTGGGGCTCGAACTGAACATGGCTTTCTCCTG
+
FFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF::FF:F,FFF:FFFFFFF,,FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF:FFFFFFFFFFFFFFFFF:FF
```

Quality score Definition

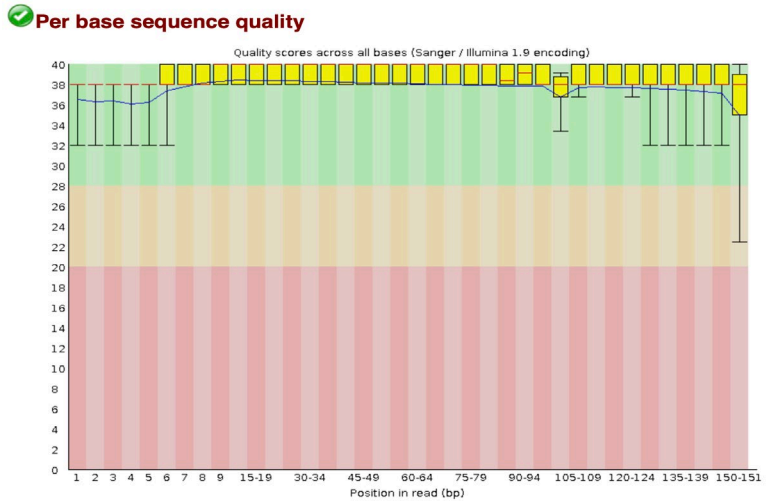
- Q = -10log_{10}(e), where e is estimated probability of the base call being wrong (error rate).

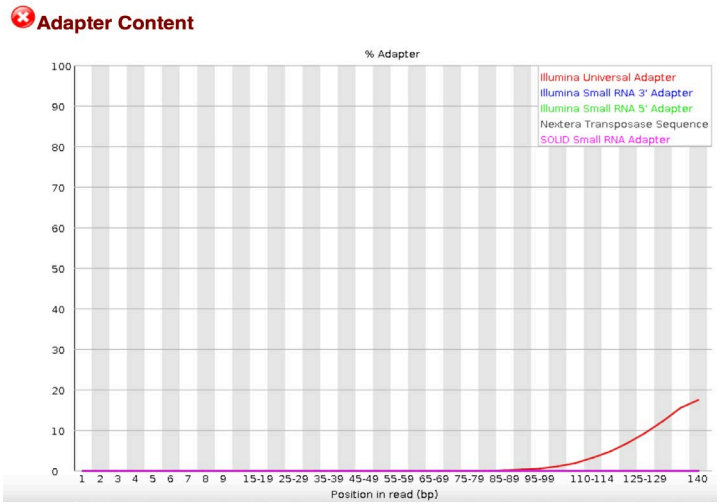
Relationship Between Sequencing Quality Score and Base Call Accuracy		
Quality Score	**Probability of Incorrect Base Call**	**Inferred Base Call Accuracy**
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

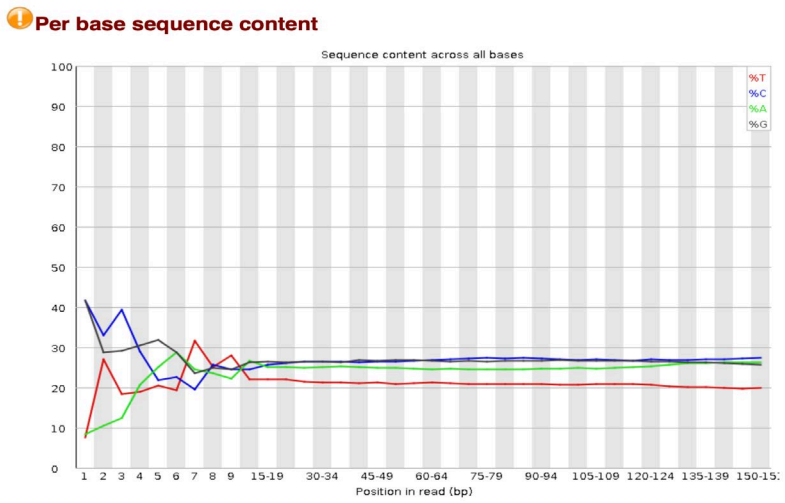
Adapter in read sequence

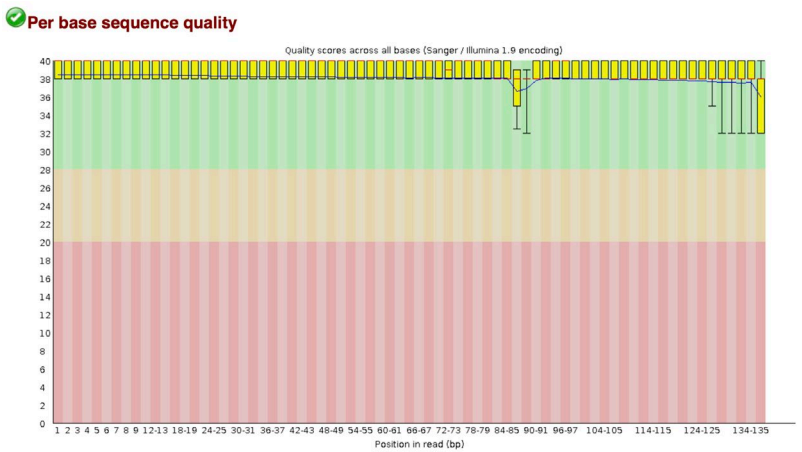


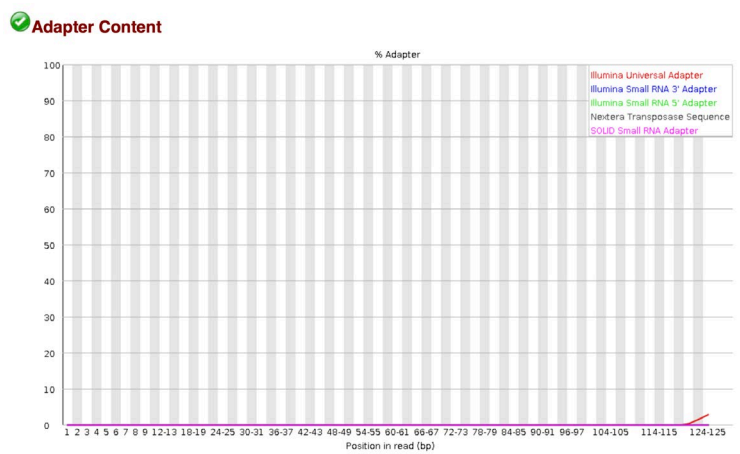
FastQC + Trimmomatic

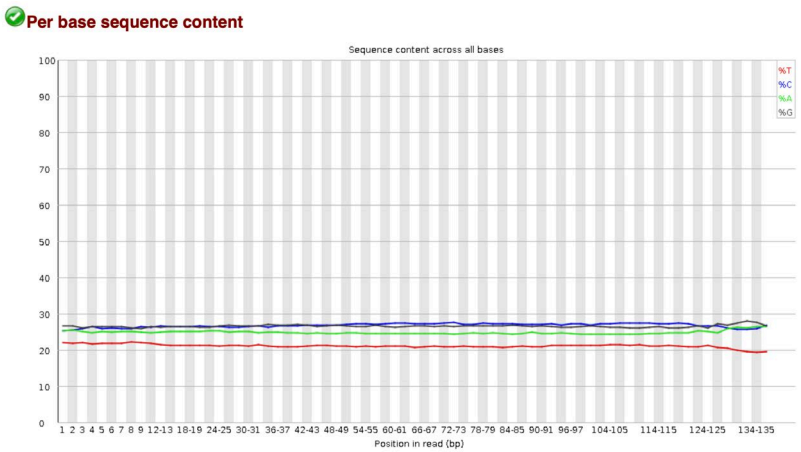












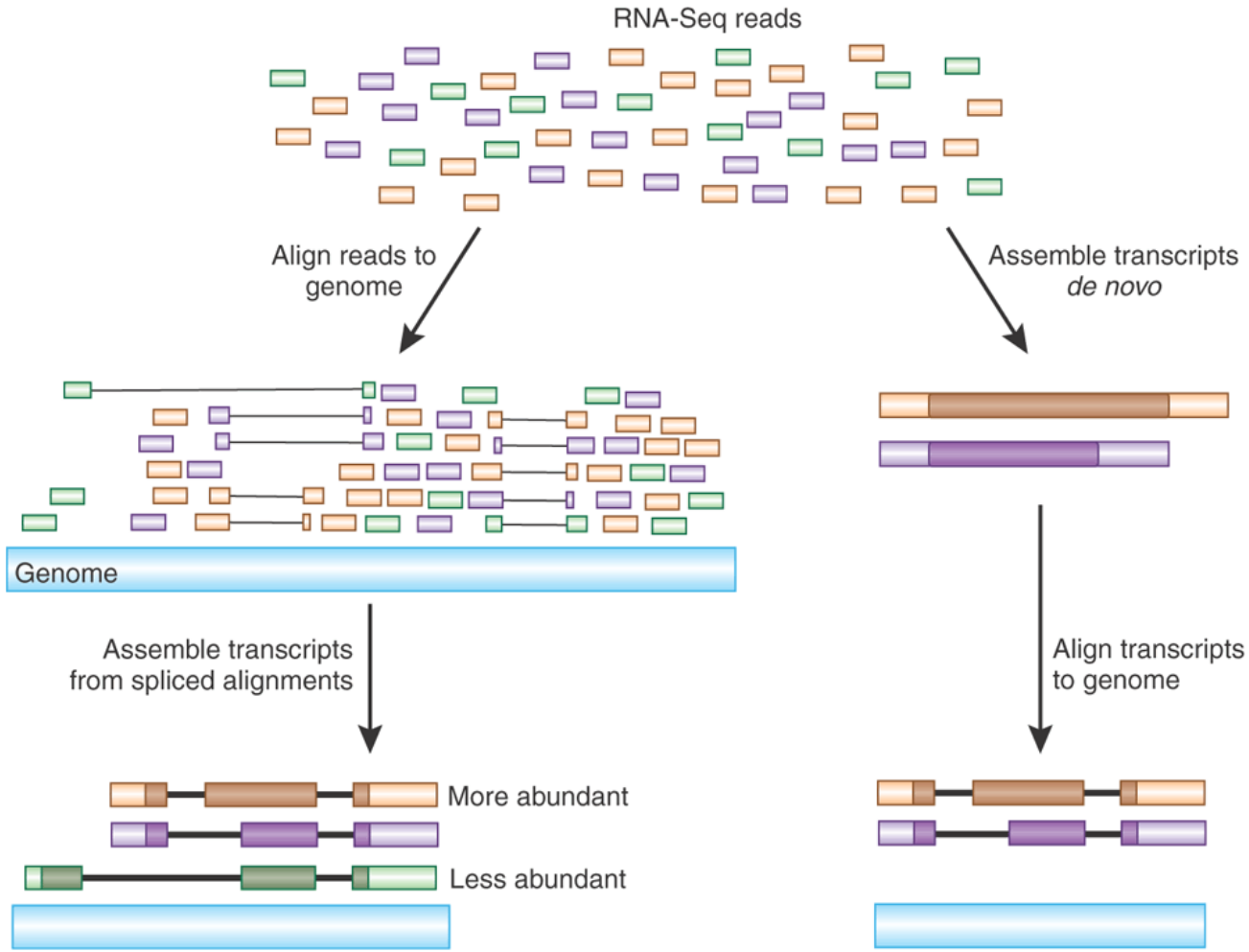
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aligning reads to genome with gene annotation

Trimmed RNA-seq reads

Reference genome sequence

Reference gene annotation



HISAT2 + StringTie pipeline

PROTOCOL

Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea[1,2], Daehwan Kim[1], Geo M Pertea[1], Jeffrey T Leek[3] & Steven L Salzberg[1–4]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [2]Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence should be addressed to S.L.S. (salzberg@jhu.edu).

Published online 11 August 2016; doi:10.1038/nprot.2016.095

Align the RNA-seq reads to the genome ● TIMING <20 min

1| Map the reads for each sample to the reference genome:

```
$ hisat2 -p 8 --dta -x chrX_data/indexes/chrX_tran -1
chrX_data/samples/ERR188044_chrX_1.fastq.gz -2
chrX_data/samples/ERR188044_chrX_2.fastq.gz -S ERR188044_chrX.sam
```

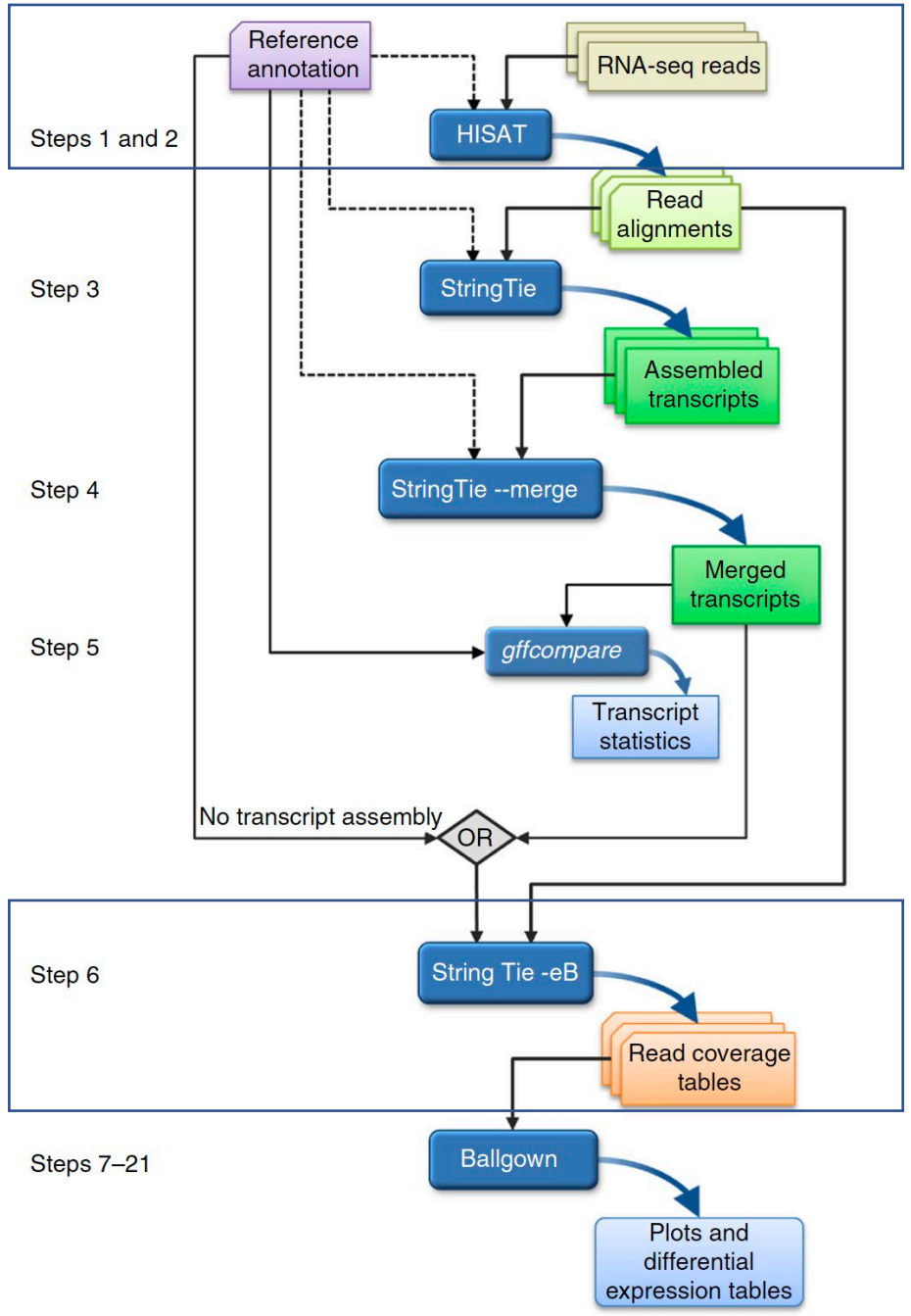
2| Sort and convert the SAM files to BAM:

```
$ samtools sort -@ 8 -o ERR188044_chrX.bam ERR188044_chrX.sam
```

Assemble and quantify expressed genes and transcripts ● TIMING ~15 min

6| Estimate transcript abundances and create table counts for Ballgown:

```
$ stringtie -e -B -p 8 -G stringtie_merged.gtf -o
ballgown/ERR188044/ERR188044_chrX.gtf ERR188044_chrX.bam
```

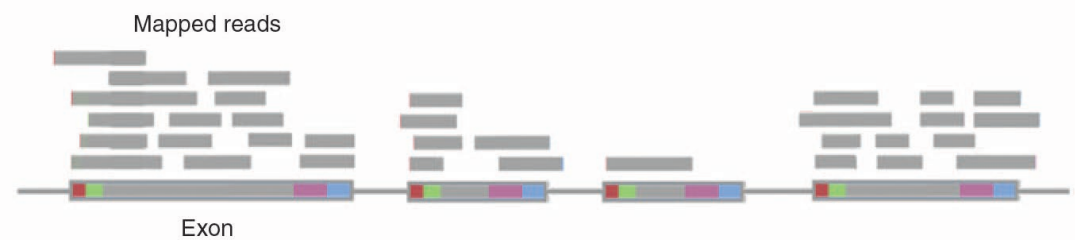


Alignment (mapped) rates

Report from HISAT2: mouse Pair-end reads

```
70383031 reads; of these:
  70383031 (100.00%) were paired; of these:
    5951960 (8.46%) aligned concordantly 0 times
    61620434 (87.55%) aligned concordantly exactly 1 time
    2810637 (3.99%) aligned concordantly >1 times
    ----
    5951960 pairs aligned concordantly 0 times; of these:
      1124205 (18.89%) aligned discordantly 1 time
    ----
    4827755 pairs aligned 0 times concordantly or discordantly; of these:
      9655510 mates make up the pairs; of these:
        5995083 (62.09%) aligned 0 times
        3355722 (34.75%) aligned exactly 1 time
        304705 (3.16%) aligned >1 times
95.74% overall alignment rate
```

Genes x Samples table



		Control			Mutant		
		Rep 1	Rep2	Rep3	Rep 1	Rep2	Rep3
gene_id	Gene name	Control_1	Control_2	Control_3	Mutant_1	Mutant_2	Mutant_3
ENSG00000132680	KHDC4	2234	2252	2245	2143	2070	2518
ENSG00000186007	LEMD1	0	0	0	0	0	0
ENSG00000285839	AL445685.3	16	6	16	7	11	8
ENSG00000203663	OR2L2	25	11	9	4	3	3
ENSG00000173406	DAB1	1	0	0	0	0	1
ENSG00000260238	PMF1-BGLAF	103	46	44	85	5	38
ENSG00000198626	RYR2	2	5	0	3	0	4
ENSG00000177174	OR14C36	0	0	0	0	0	0
ENSG00000117640	MTFR1L	1450	1421	1255	526	1260	1301
ENSG00000143633	C1orf131	207	230	217	104	213	209
ENSG00000136628	EPRS	7964	6685	6690	1627	6879	6911
ENSG00000171819	ANGPTL7	82	51	80	16	75	58
ENSG00000143514	TP53BP2	2445	2557	2576	1313	2005	2251
ENSG00000162654	GBP4	0	5	14	11	0	11
ENSG00000198758	EPS8L3	4	0	5	3	2	3
ENSG00000188529	SRSF10	9102	9440	6754	2047	6718	8161
ENSG00000171786	NHLH1	0	4	7	0	0	7
ENSG00000173372	C1QA	0	0	0	0	0	0
ENSG00000131788	PIAS3	405	521	443	183	400	596
ENSG00000184022	OR2T10	0	0	0	1	0	4
ENSG00000169231	THBS3	935	847	898	741	840	1021
ENSG00000269113	TRABD2B	0	0	0	0	0	0
ENSG00000177151	OR2T35	0	0	0	0	0	0
ENSG00000284895	AC119676.1	0	1	1	1	1	1
ENSG00000143436	MRPL9	938	753	761	193	768	799
ENSG00000264343	NOTCH2NLA	118	86	114	286	122	164
ENSG00000142623	PADI1	0	0	0	0	0	0
ENSG00000120656	TAF12	314	292	368	129	276	351
ENSG00000162669	HFM1	0	3	1	8	0	7
ENSG00000186141	POLR3C	940	913	974	649	907	964
ENSG00000169418	NPR1	6	11	0	0	0	0
ENSG00000253304	TMEM200B	37	36	11	18	29	18
ENSG00000010803	SCMH1	1568	1489	1514	1248	1541	1556
ENSG00000137968	SLC44A5	0	5	5	4	0	0
ENSG00000158859	ADAMTS4	36	75	56	93	45	83
ENSG00000154358	OBSCN	1882	1125	2334	4522	2142	1619

Filter out non-nuclear-protein-coding genes

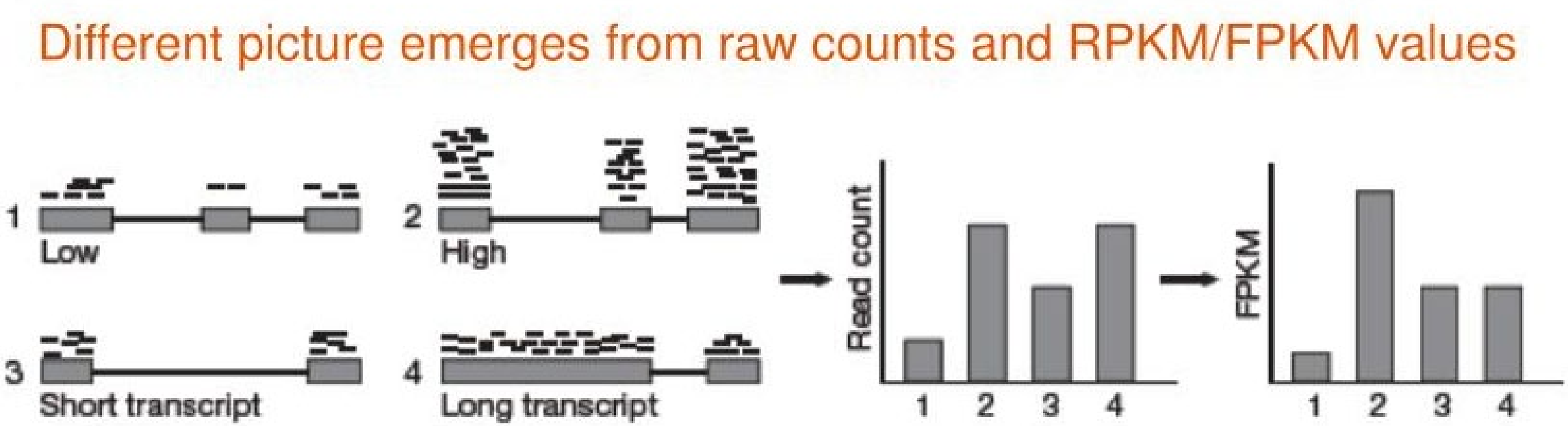
1	gene id	gene name	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
60587	ENSG00000279011	AC091905.2	0	0	0	0	0	0
60588	ENSG00000237801	AMD1P2	1	0	0	1	0	0
60589	ENSG00000228304	OR4K6P	0	0	0	0	0	0
60590	ENSG00000207237	RNU6-110P	0	2	2	0	1	2
60591	ENSG00000218213	FTH1P26	0	0	0	3	0	0
60592	ENSG00000223202	RN7SKP297	1	3	0	0	1	0
60593	ENSG00000089639	GMIP	1501	1647	1425	3033	1405	1788
60594	ENSG00000259614	AC087477.6	0	0	0	0	0	0
60595	ENSG00000241347	RN7SL466P	0	0	0	0	0	0
60596	ENSG00000233436	BTBD18	258	219	175	85	158	166
60597	ENSG00000156017	CARNMT1	1763	1771	1542	658	1575	1666
60598	ENSG00000176510	OR10AC1	0	2	0	3	0	2
60599	ENSG00000269524	AC245052.7	0	0	0	0	0	0
60600	ENSG00000265644	AC087399.1	0	0	0	0	0	0
60601	ENSG00000276417	AC092111.2	0	0	7	2	0	0
60602	ENSG00000272346	AC103810.8	1	1	0	2	0	0
60603	ENSG00000235780	USP17L27	1	1	2	3	2	1
60604	ENSG00000181325	OR9G3P	0	0	0	0	0	0
60605	ENSG00000217026	RPL10P1	0	0	0	0	0	1
60606	ENSG00000113971	NPHP3	711	839	847	603	756	727
60607	ENSG00000211452	DIO1	24	21	7	12	28	18
60608	ENSG00000109132	PHOX2B	0	0	0	1	0	0
60609	ENSG00000231814	LINC00210	0	0	0	3	0	0
60610	ENSG00000201410	RF00342	0	0	0	0	0	0
60611	ENSG00000117151	CTBS	1209	1342	1306	427	1051	1357
60612	ENSG00000282995	FRG1EP	34	56	29	29	25	29
60613	ENSG00000230772	VN1R108P	10	18	13	10	12	3
60614	ENSG00000259912	AC023813.1	0	0	0	0	0	0
60615	ENSG00000226582	UBE2V1P5	0	2	0	0	0	0
60616	ENSG00000156042	CFAP70	14	31	15	9	11	11
60617	ENSG00000261543	AC010931.3	0	0	0	0	0	0
60618	ENSG00000227233	CICP17	1	1	3	1	0	2

1	gene_id	Gene name	biotype	ome/scaff	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
19838	ENSG00000286135	AL009764.2	protein_coding	10	0	0	0	0	0	0
19839	ENSG00000286137	AC211486.8	protein_coding	7	19	13	23	10	25	18
19840	ENSG00000286140	DERPC	protein_coding	16	496	295	923	158	767	692
19841	ENSG00000286143	AC053503.7	protein_coding	2	4	0	0	15	0	0
19842	ENSG00000286165	AC012488.2	protein_coding	2	9	0	0	0	0	0
19843	ENSG00000286175	AC023490.4	protein_coding	22	0	0	0	0	0	0
19844	ENSG00000286185	AC242842.3	protein_coding	1	0	266	164	515	350	1104
19845	ENSG00000286190	AC055839.2	protein_coding	17	0	0	0	0	0	0
19846	ENSG00000286192	AC069288.1	protein_coding	7	139	164	392	238	138	202
19847	ENSG00000286219	AC242843.1	protein_coding	1	395	391	514	348	485	392
19848	ENSG00000286221	AC009070.1	protein_coding	16	0	0	0	3	9	10
19849	ENSG00000286224	AP000471.1	protein_coding	21	6	7	6	4	7	3
19850	ENSG00000286228	SPDYE17	protein_coding	7	64	32	63	19	71	50
19851	ENSG00000286231	AL445423.2	protein_coding	1	0	0	2	0	0	0
19852	ENSG00000286235	AL035461.3	protein_coding	20	0	0	0	0	0	0
19853	ENSG00000286237	ARMCX5-GF	protein_coding	X	14	32	29	0	7	14
19854	ENSG00000286239	AC093884.1	protein_coding	2	9	0	0	15	17	0
19855	ENSG00000286264	AP001453.5	protein_coding	11	58	9	30	66	29	18
19856	ENSG00000286268	AF196969.1	protein_coding	X	13	42	41	28	50	26
19857	ENSG00000286522	HIST1H3B	protein_coding	6	5784	4656	5886	2038	5343	5463
19858	ENSG00000286905	AC108488.2	protein_coding	2	118	121	107	82	119	124
19859	ENSG00000286920	AL662820.1	protein_coding	6	0	0	1	0	1	1
19860	ENSG00000287080	HIST1H3C	protein_coding	6	1516	1207	1631	1273	1304	1414
19861	ENSG00000287363	AL096814.2	protein_coding	6	0	0	0	0	0	0
19862	ENSG00000287542	AC098582.1	protein_coding	4	0	54	0	0	54	247
19863	ENSG00000287585	AC231656.1	protein_coding	X	0	0	0	0	0	0
19864	ENSG00000287694	AC106741.1	protein_coding	16	0	0	0	0	0	0
19865	ENSG00000287725	AP003071.5	protein_coding	11	71	42	33	7	59	41
19866	ENSG00000287856	AL445524.2	protein_coding	1	2	0	0	0	0	0
19867	ENSG00000287908	AC025165.6	protein_coding	12	0	1	24	27	8	2
19868	ENSG00000288000	AL031681.2	protein_coding	20	68	29	0	0	0	0
19869	ENSG00000288053	AC231657.3	protein_coding	X	0	32	4	2	0	0

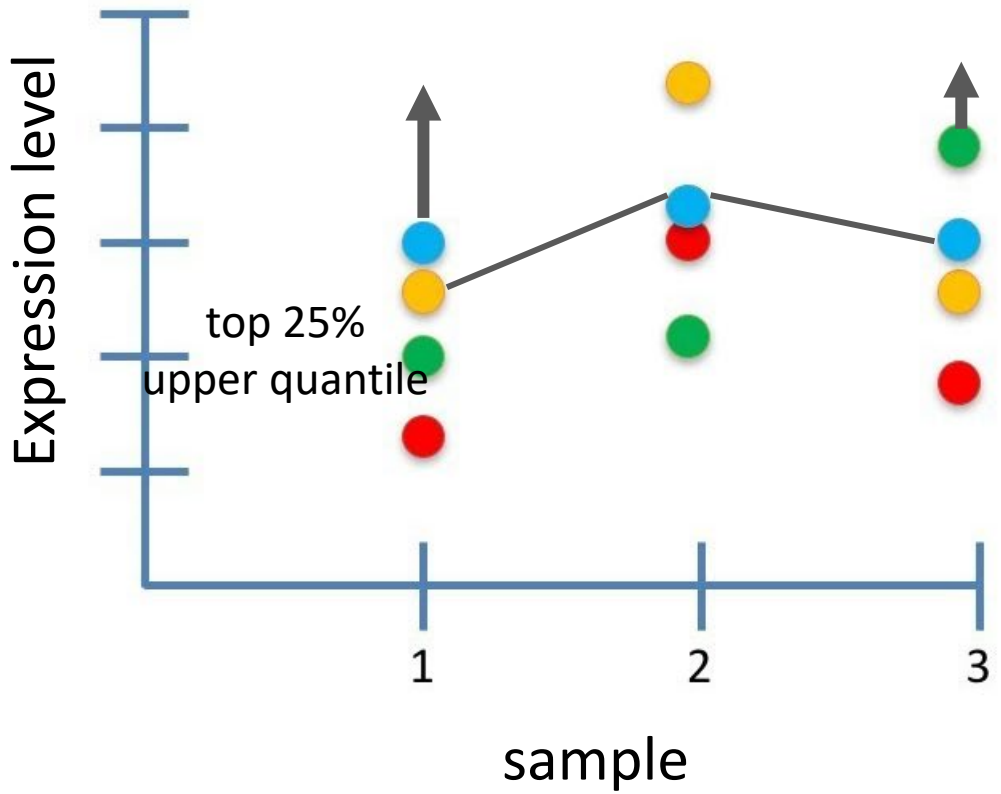
60,617 genes from gene annotation → 19,868 nuclear protein-coding genes

Gene expression quantification

- Read count
- Fragments per kilobase transcript, per million mapped reads (FPKM)

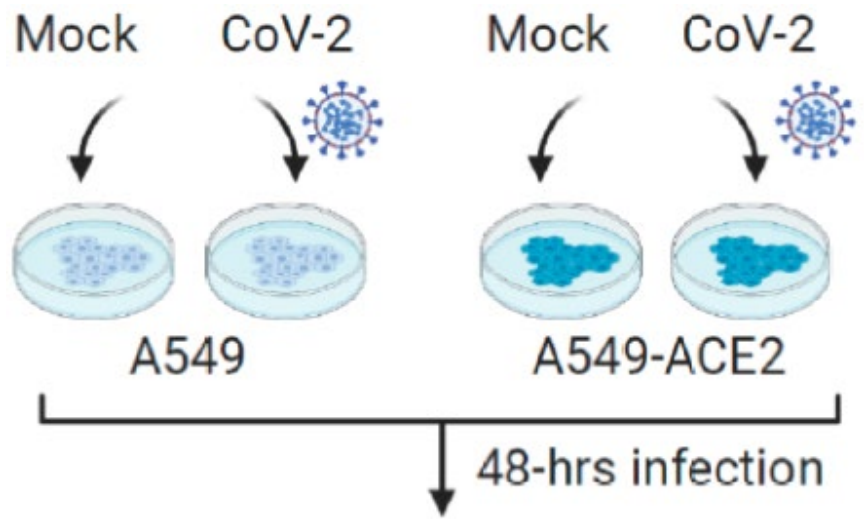


Inter sample normalization (upper quantile normalization)



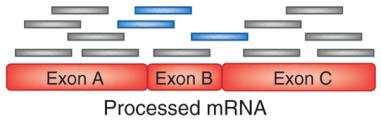
Basic transcriptome analysis

- Check the overall similarity among samples
  - Principal component analysis (PCA) plot
- Identify the differentially expressed genes (DEGs)
  - MA-plot for showing the up- or down-regulated DEGs
  - Volcano plot for narrow down the number of candidate genes
- Overrepresentation analysis (ORA)
  - Enriched GO terms
  - Enriched KEGG pathways
- Identify genes with co-expression patterns
  - Heatmap with clustering



RNA library

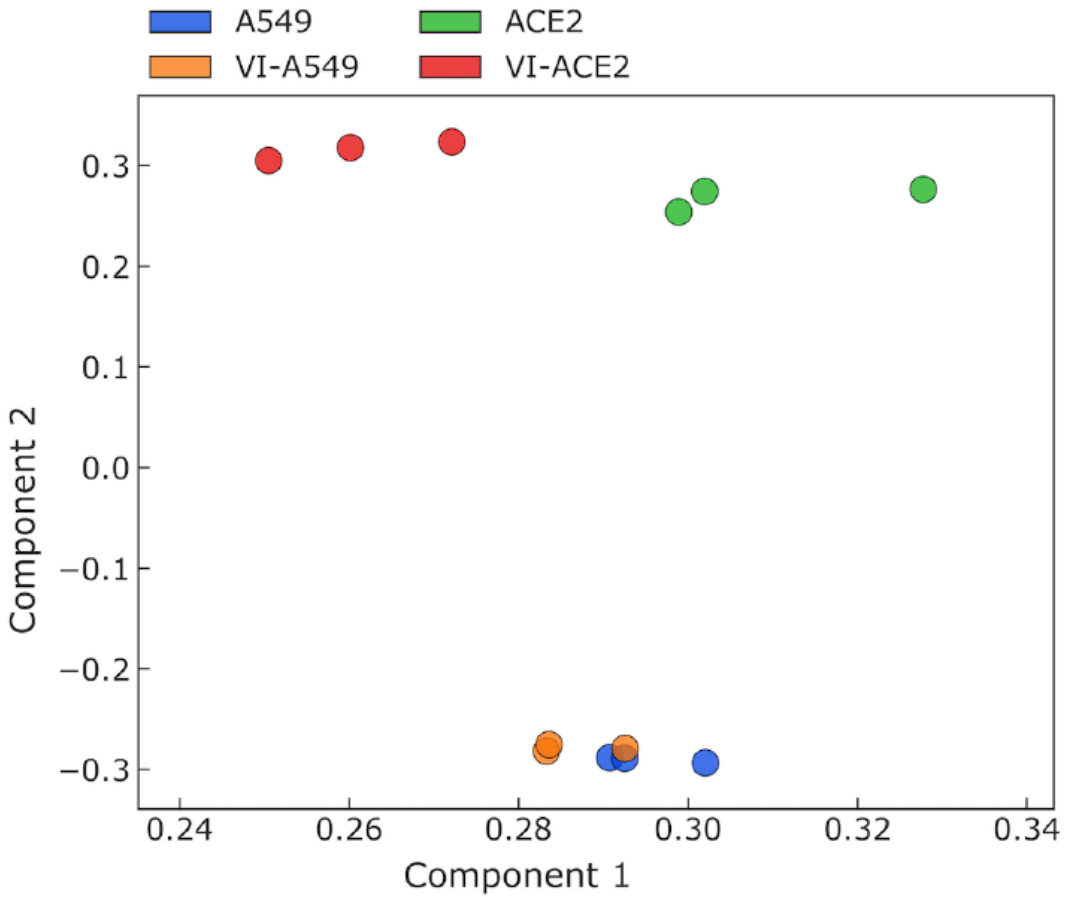
RNA-sequencing



expression

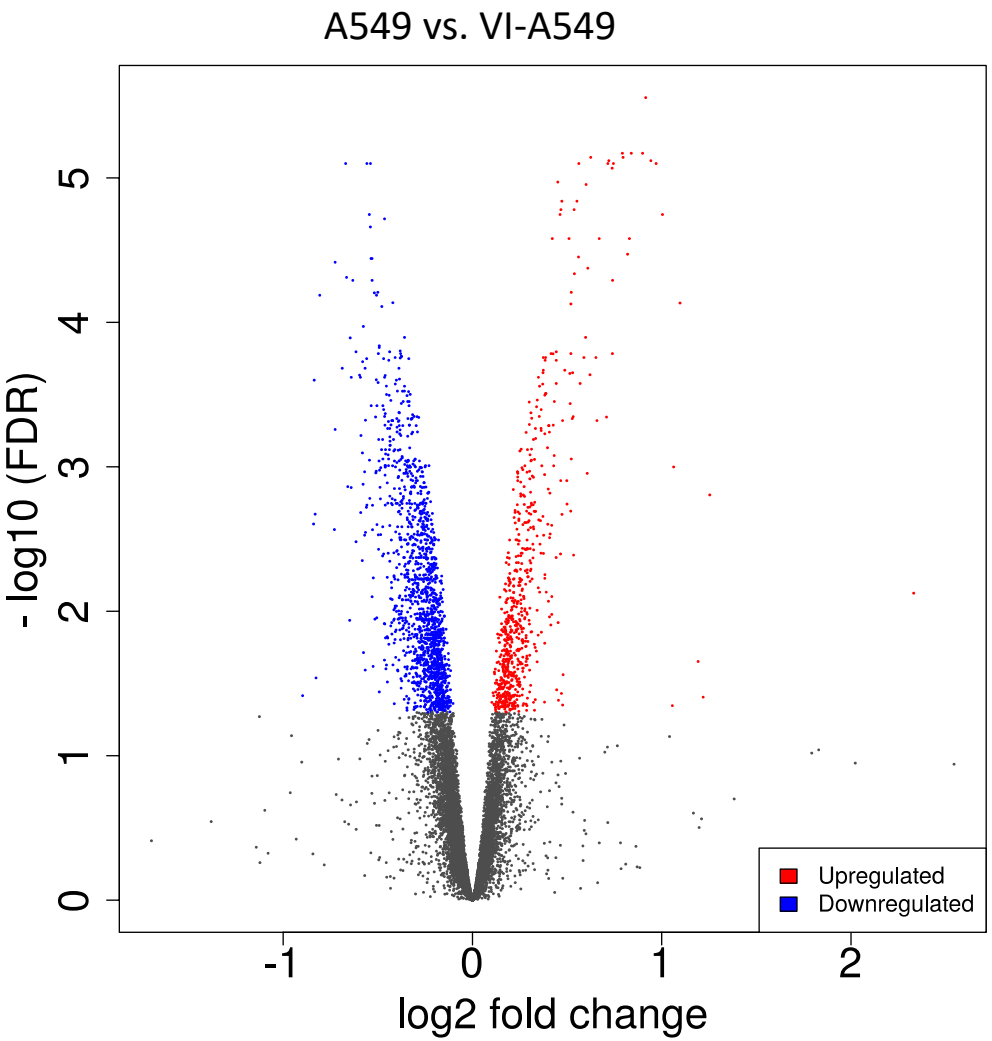
	A549	VI-A549	A549-ACE2	VI-A549-ACE2
gene 1	123.56	27.32	.....	.....
gene 2	17.88	90.04		
⋮	⋮	⋮	⋮	⋮
gene x	0.71	161.09		

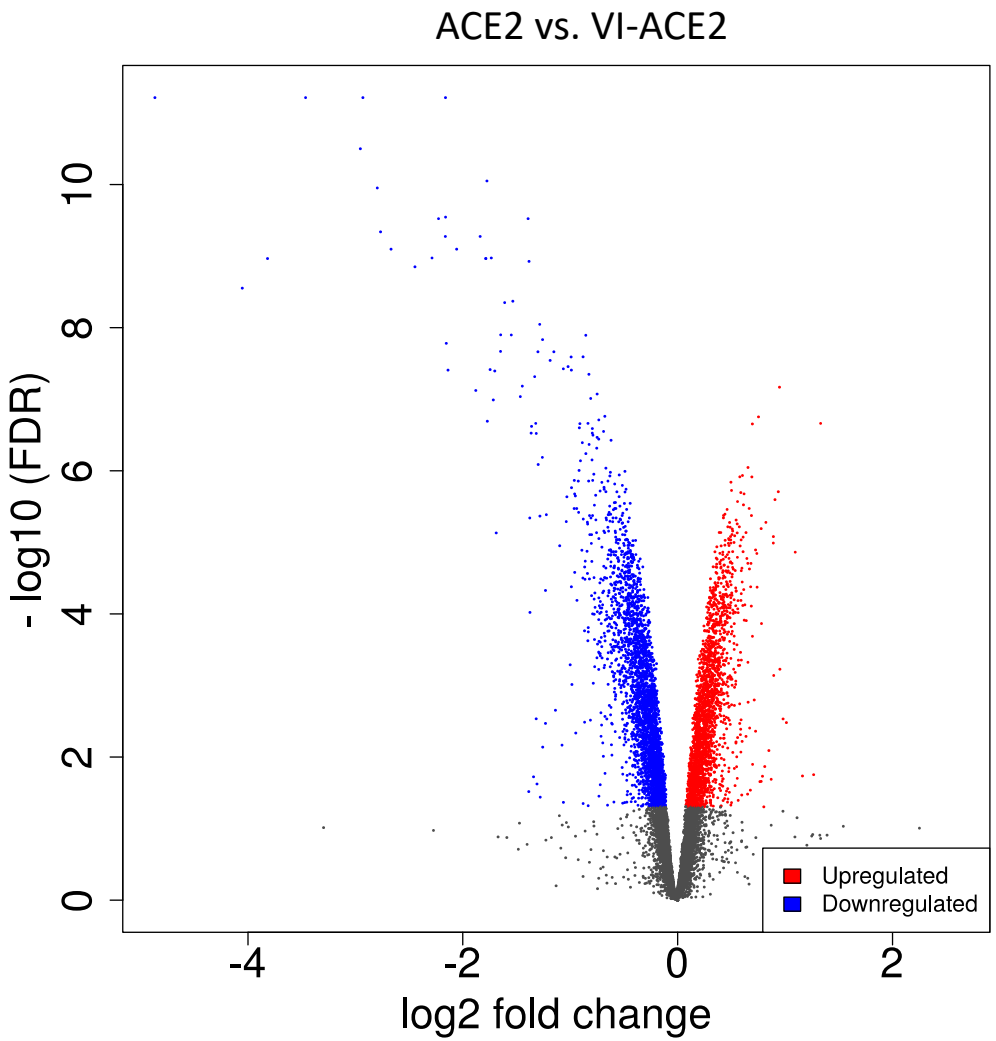
Principle component analysis (PCA)

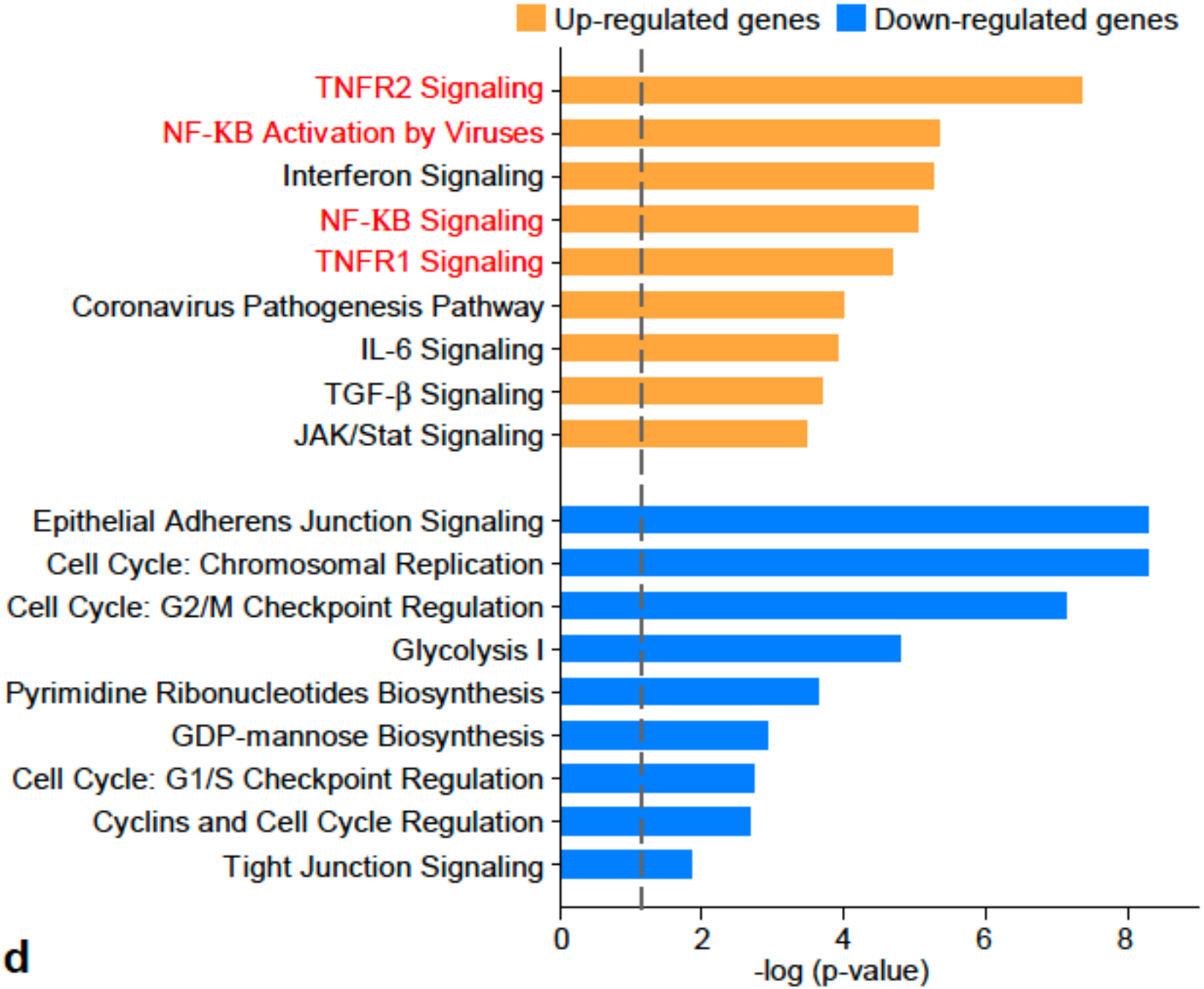


Data from Dr. Chia-Wei Li's lab, IBMS, AS

DEGs identification (volcano-plot)

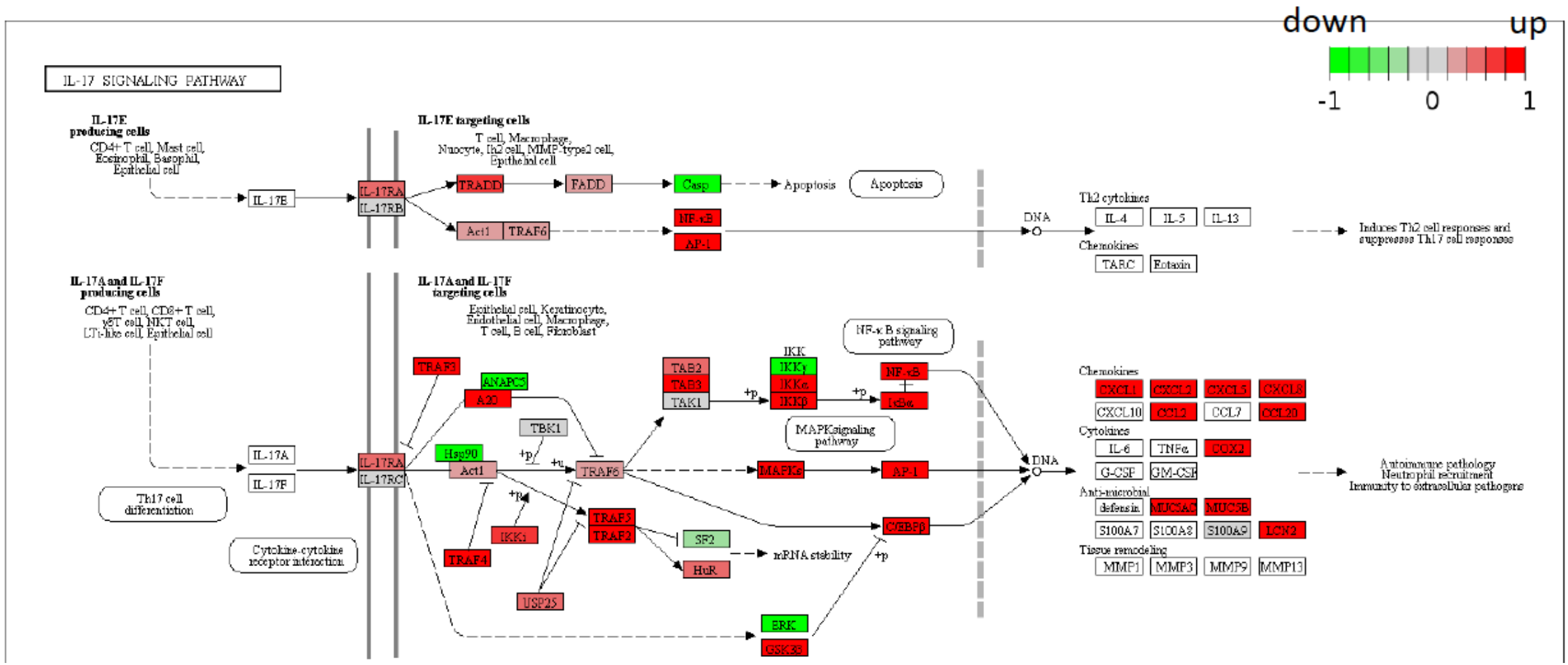




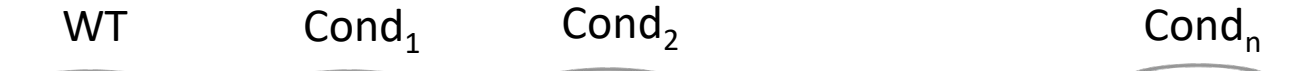

Up-regulated genes
Down-regulated genes
TNFR2 Signaling
NF-KB Activation by Viruses
Interferon Signaling
NF-KB Signaling
TNFR1 Signaling
Coronavirus Pathogenesis Pathway
IL-6 Signaling
TGF-β Signaling
JAK/Stat Signaling
Epithelial Adherens Junction Signaling
Cell Cycle: Chromosomal Replication
Cell Cycle: G2/M Checkpoint Regulation
Glycolysis I
Pyrimidine Ribonucleotides Biosynthesis
GDP-mannose Biosynthesis
Cell Cycle: G1/S Checkpoint Regulation
Cyclins and Cell Cycle Regulation
Tight Junction Signaling
0
2
4
6
8
-log (p-value)

d

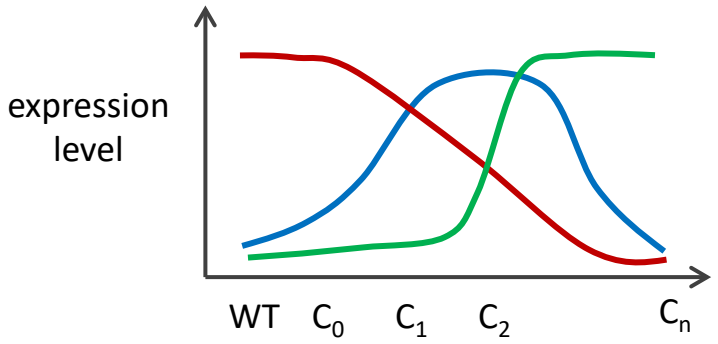
DEGs enriched in a pathway

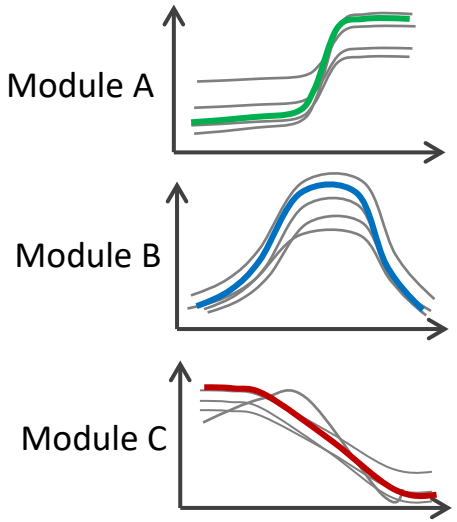


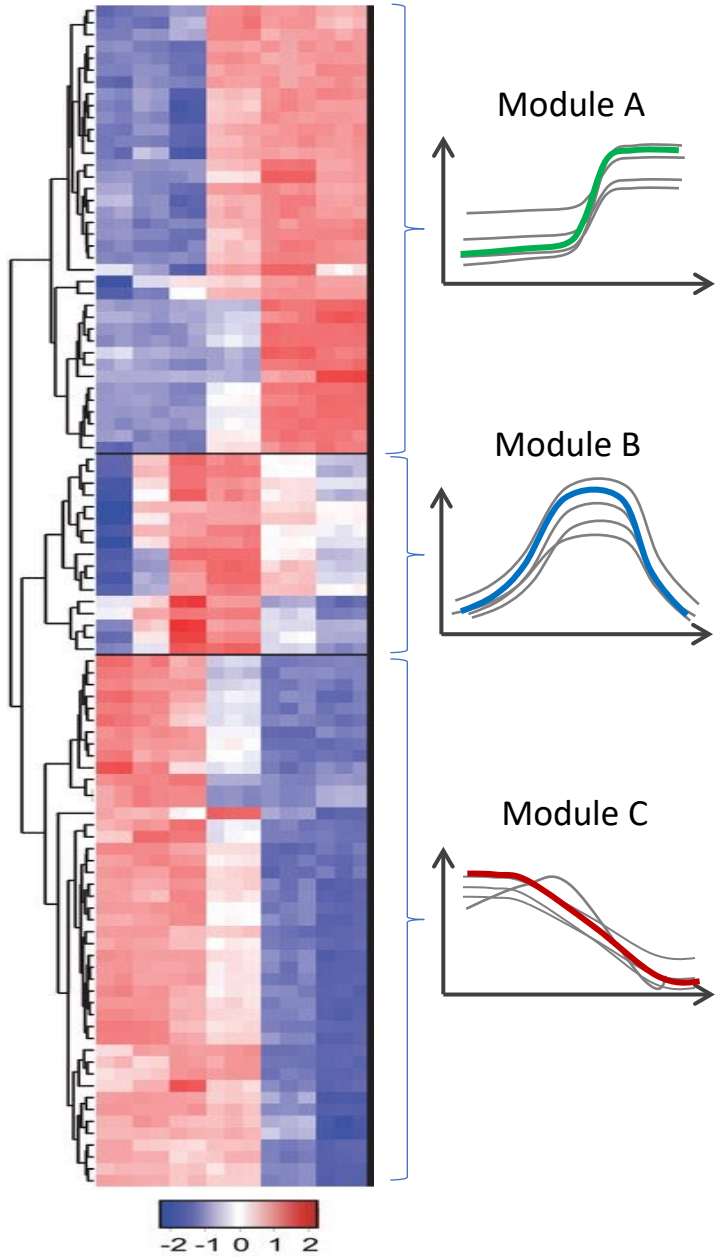
Clustering genes in modules

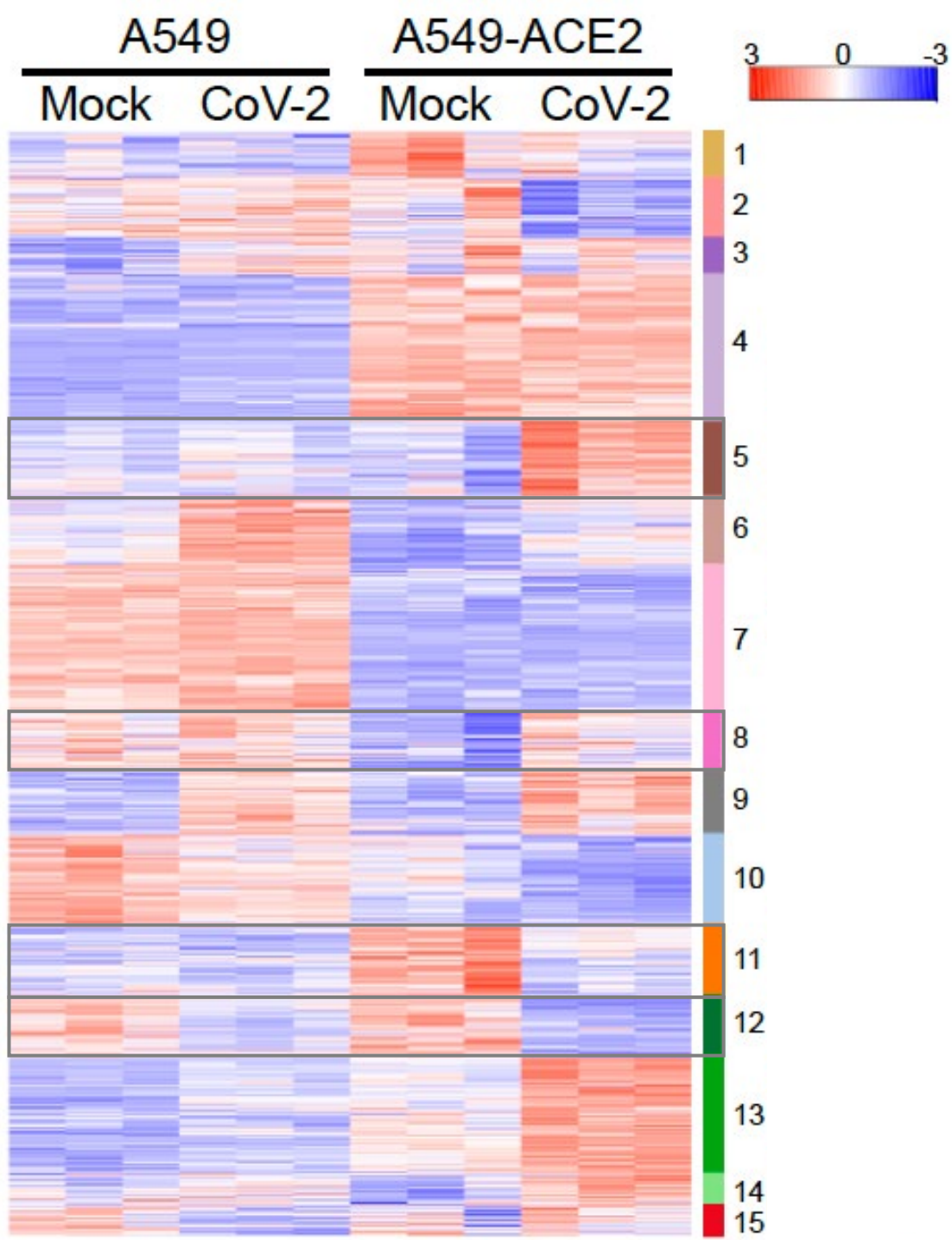


	WT	C_1	C_2		C_n
gene 1	123.56	27.32	20.06		5.99
gene 2	17.88	90.04	96.29		331.36
⋮	⋮	⋮	⋮		⋮
gene x	0.71	161.09	509.41		1.85

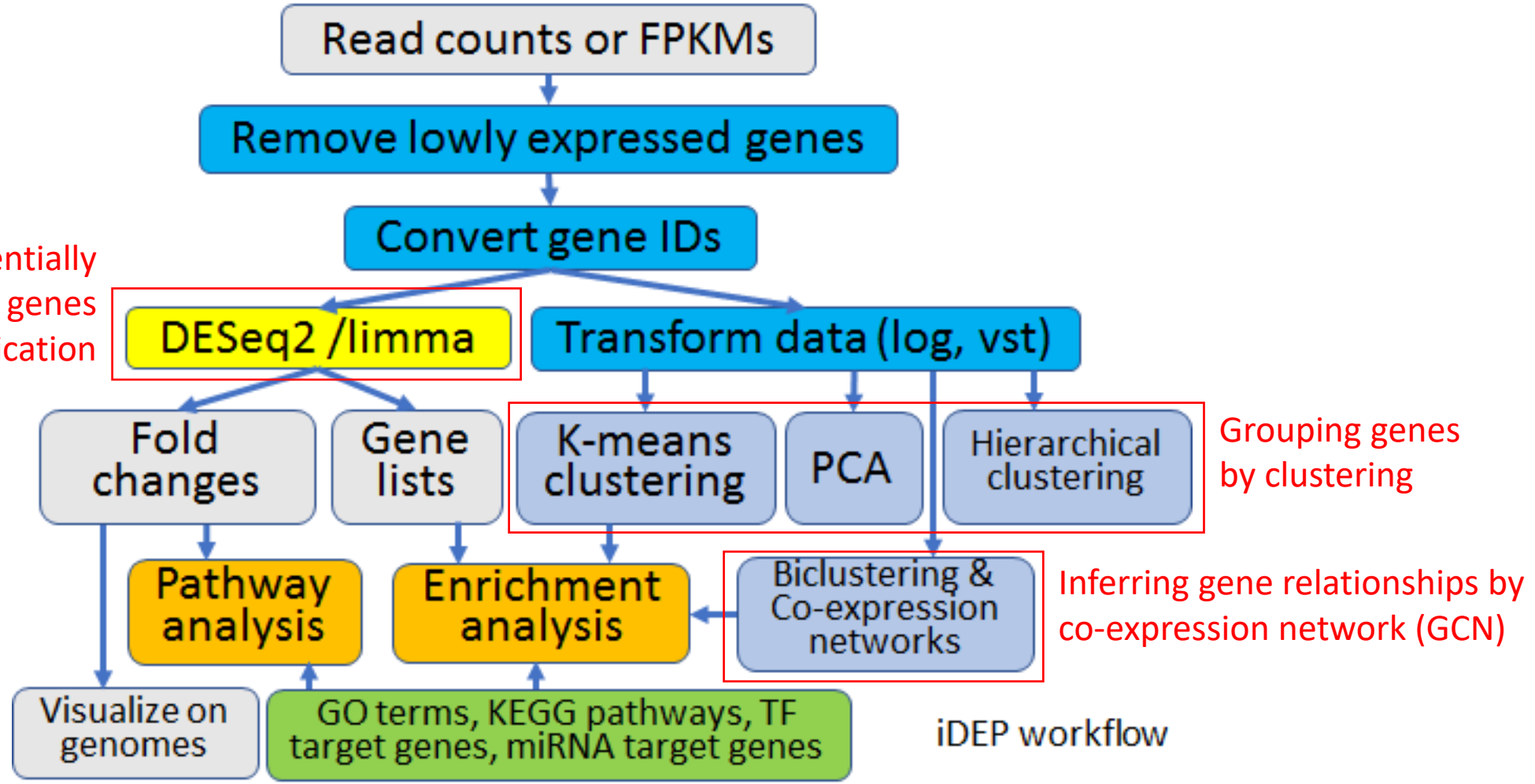







A549
A549-ACE2
Mock
CoV-2
Mock
CoV-2
3
0
-3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Web-based bulk RNA-seq data analysis tool



Ge et al., BMC Bioinformmatics, 2018

Reset

1. Species *Human*

Select

New

2. Data type

Read counts data

+ Settings

Public Data | Cite iDEP | Gene IDs

Using Human genome annotations and pathways.

Warning: Low conversion rate! 26009 out of 35237 genes (73.8%) converted to Ensembl/STRING IDs.

Study_design	hesc_cyto_rep1	hesc_cyto_rep2	hesc_monosome_rep1	hesc_monosome_rep2	hesc_nuc_rep1	hesc_nuc_rep2
cell	hesc	hesc	hesc	hesc	hesc	hesc
RNA	cyto	cyto	monosome	monosome	nuc	nuc

	hesc_cyto_rep1	hesc_cyto_rep2	hesc_monosome_rep1	hesc_monosome_rep2	hesc_nuc_rep1	hesc_nuc_
A1BG	348	515	467	593	314	
A1CF	30	78	9	14	40	
A2M	436	603	111	120	327	
A2ML1	584	839	93	77	386	
A2MP1	48	70	16	3	38	
A3GALT2	0	0	0	0	0	
A4GALT	300	328	272	269	76	
A4GNT	0	2	0	0	0	
AAAS	1982	2658	536	693	651	
AACS	3279	3818	857	1147	1458	

1. Filter genes with low counts

Min. CPM: 0.5

n libraries: 1

2. Transform counts data

EdgeR: log2(CPM+c)

Pseudo-count c: 4

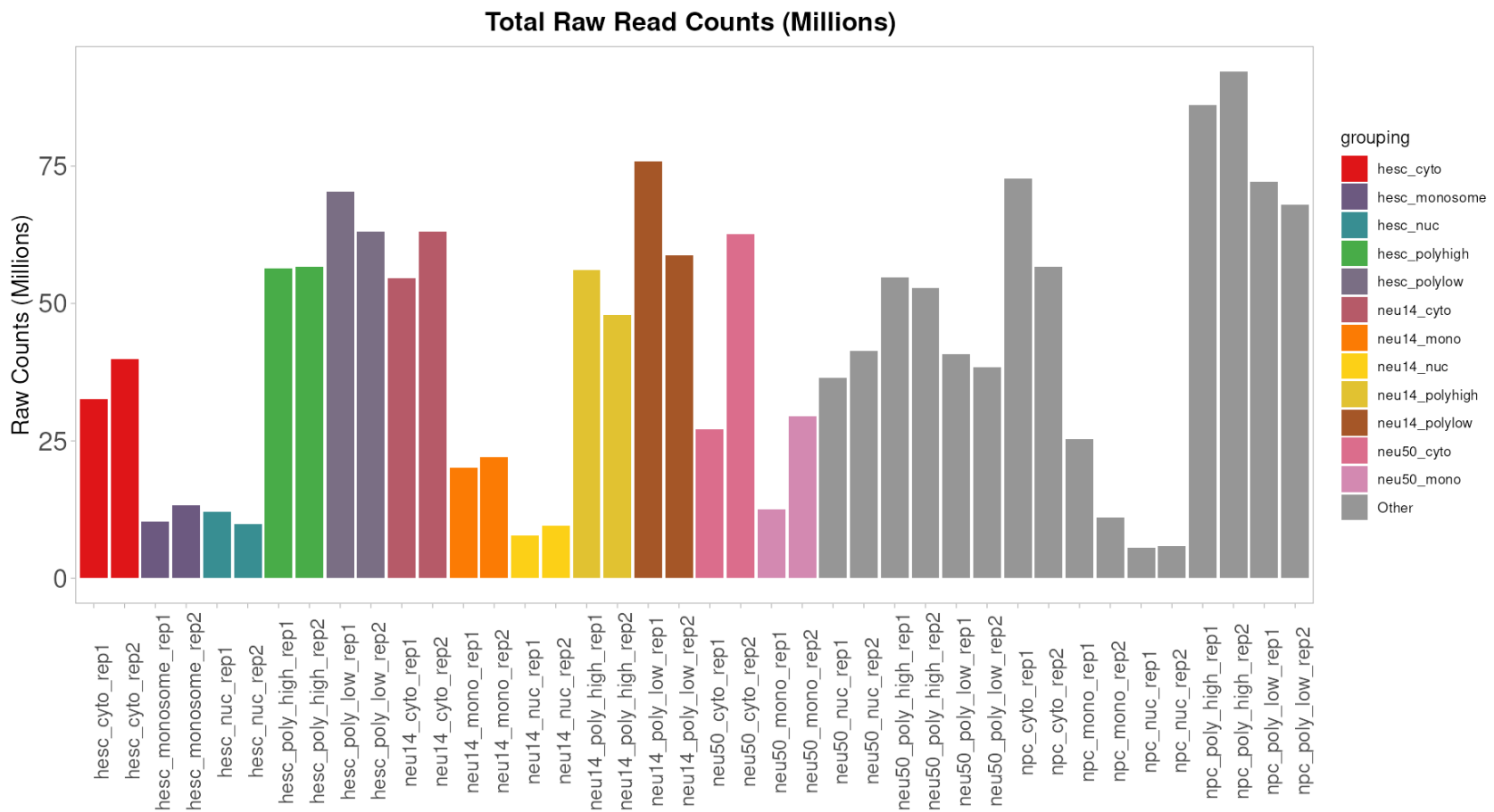
Missing values: Use gene median

Processed data | Converted counts

.RData | Report

After mapping, there are 35238 unique genes in 40 samples. Of the 24605 genes passed filter, 19987 were converted to Ensembl/STRING gene IDs in our database. The remaining 4618 genes were kept in the data using original IDs.

Reads | Distribution | Scatterplot | QC | Markers | Genes | Data



Choose which principal component is shown on the x-axis.

X-axis: 1

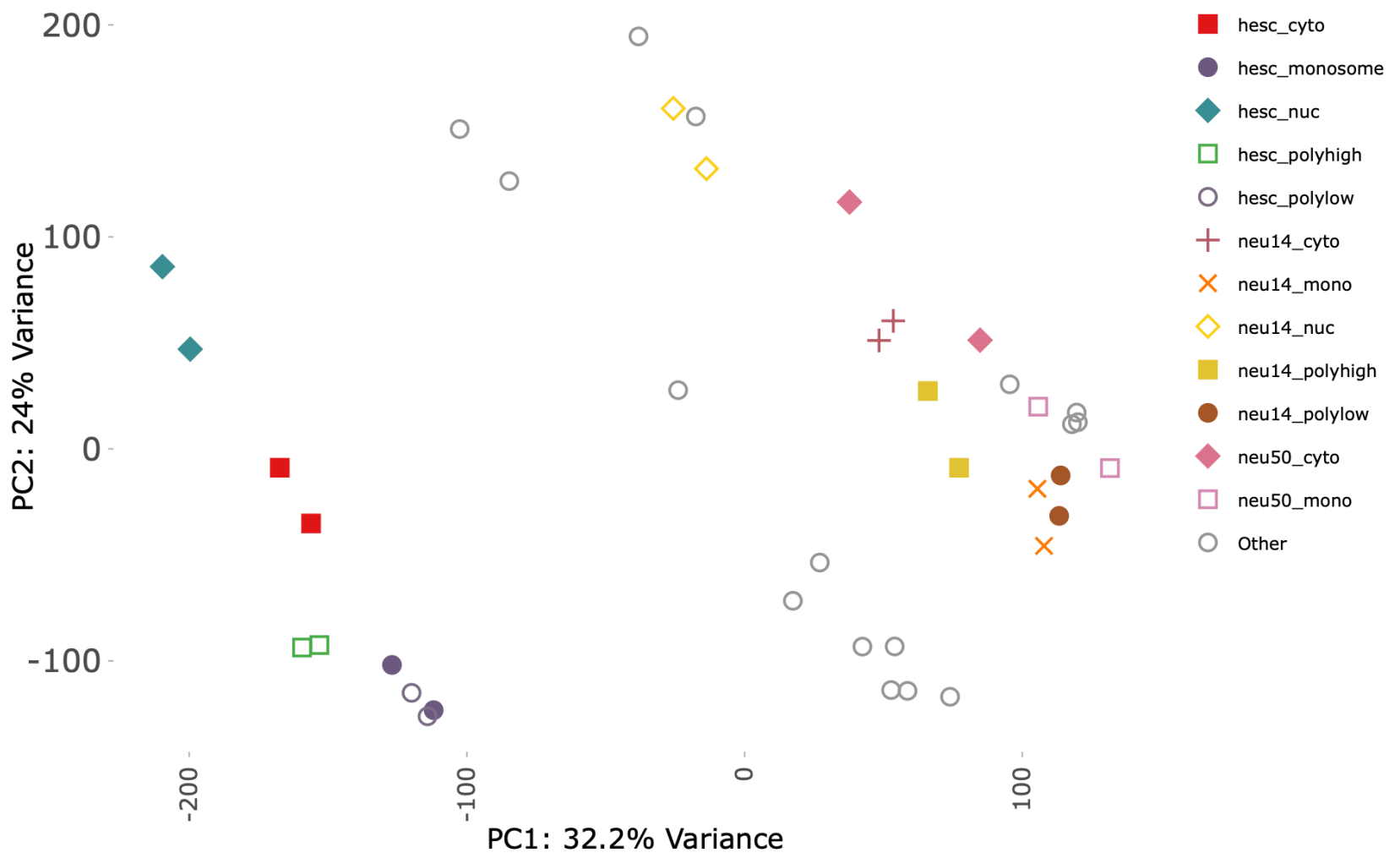
Y-axis: 2

Shape: Names

Color: Names

PCA data | Report

PCA | 3D | PCAtools | MDS | t-SNE



PC1 is correlated with cell (p=2.00e-10).PC2 is correlated with RNA (p=1.01e-06).PC3 is correlated with cell (p=3.65e-18).PC4 is correlated with RNA (p=5.34e-16).PC5 is correlated with cell (p=3.10e-17).

Identify DEGs

Submit

Method:

DESeq2

FDR cutoff: 0.05

Min fold-change: 1

- [ ] Threshold-based Wald Test
- [x] Independent filtering of lower counts

Results Gene Lists

Report

Experiment Design | Results | Venn Diagram | R Code

1. Select main factors (optional)

- [x] cell
- [ ] RNA

Select a batch/block factor (optional)

- [ ] RNA

Reference level for cell

npc

Model: expression ~ cell

2. Select at least one comparison:

- [ ] cell: hesc vs. neu14
- [ ] cell: hesc vs. neu50
- [x] cell: hesc vs. npc
- [ ] cell: neu14 vs. hesc
- [ ] cell: neu14 vs. neu50
- [x] cell: neu14 vs. npc
- [ ] cell: neu50 vs. hesc
- [ ] cell: neu50 vs. neu14
- [x] cell: neu50 vs. npc
- [ ] cell: npc vs. hesc
- [ ] cell: npc vs. neu14
- [ ] cell: npc vs. neu50

Identify DEGs

Submit

Method:

DESeq2

FDR cutoff: 0.05

Min fold-change: 1

☐ Threshold-based Wald Test

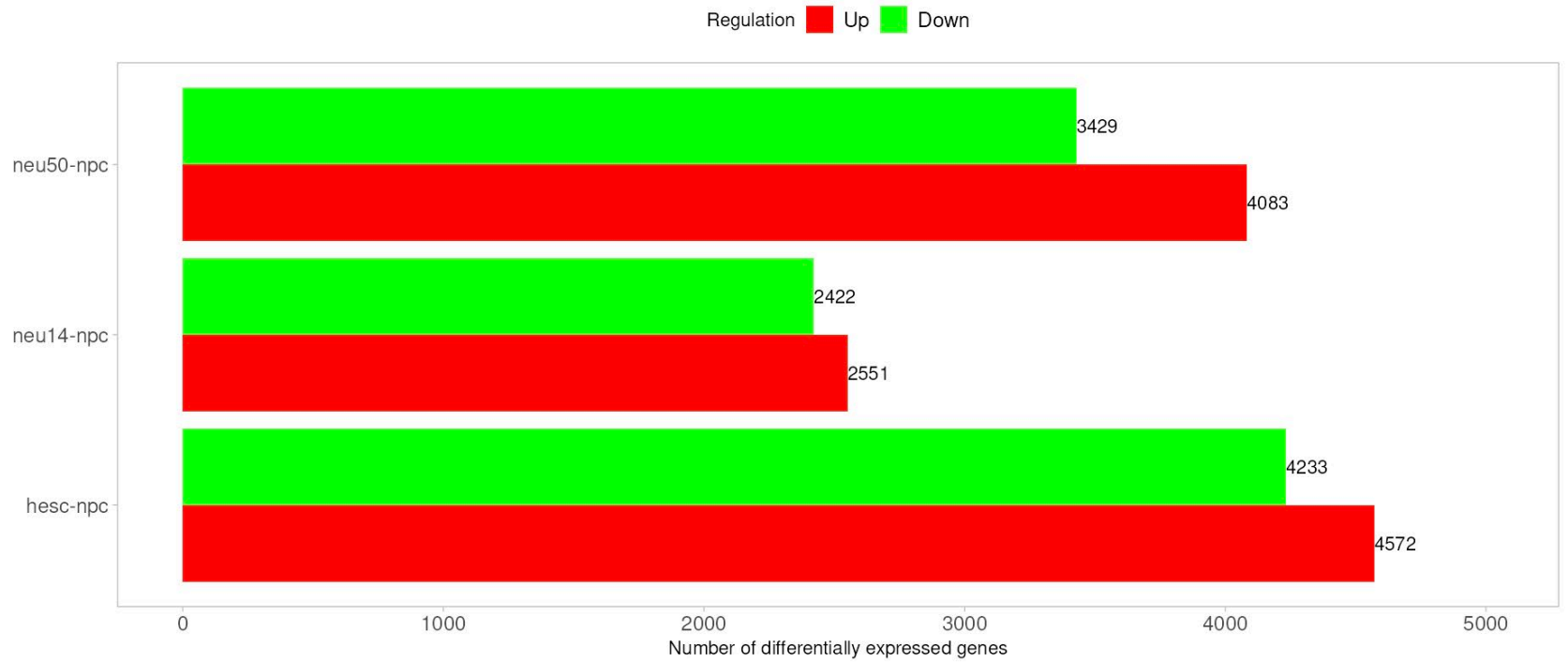
☑ Independent filtering of lower counts

Green-Red

Results Gene Lists

Report

Experiment Design | Results | Venn Diagram | R Code



Numbers of differentially expressed genes for all comparisons. "B-A" means B vs. A. Interaction terms start with "I:"

Comparisons	Up	Down
neu50-npc	4083	3429
neu14-npc	2551	2422
hesc-npc	4572	4233

Identify DEGs

Submit

Method:

DESeq2

FDR cutoff: 0.05

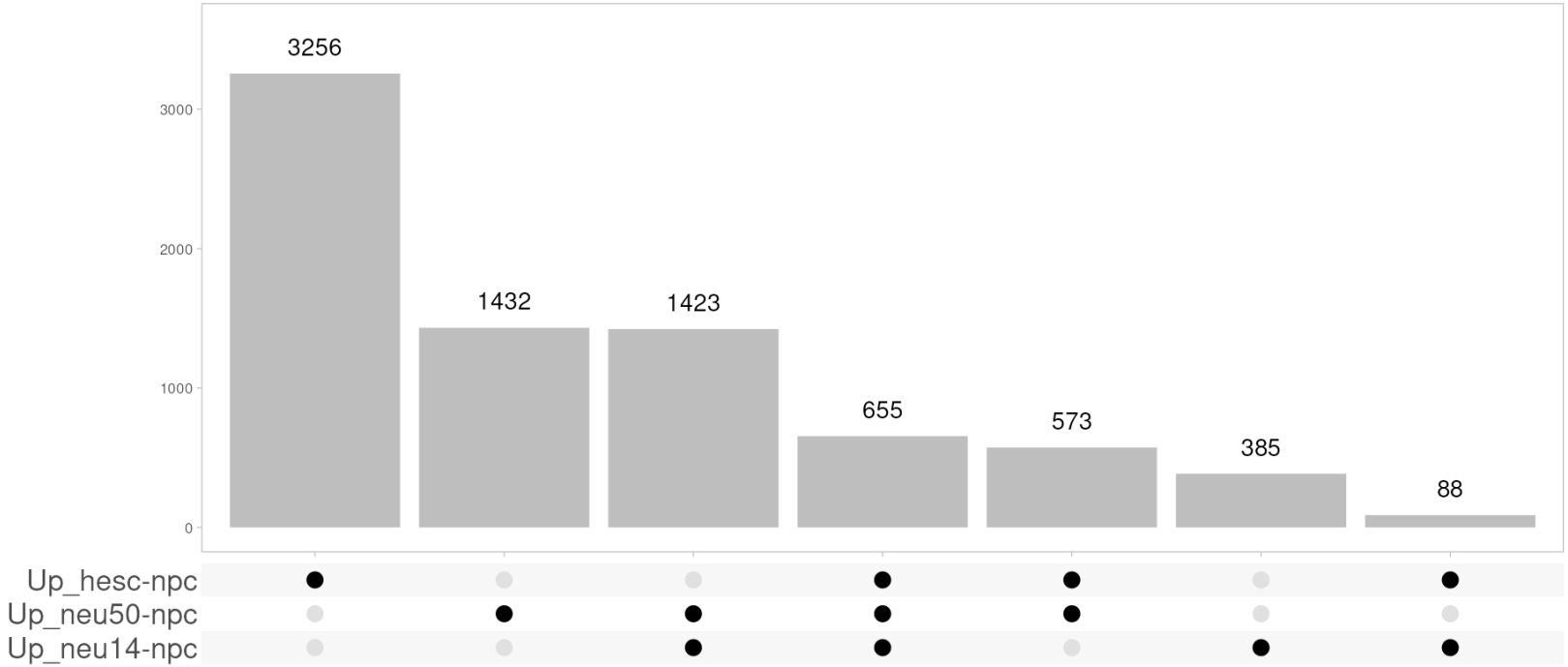
Min fold-change: 1

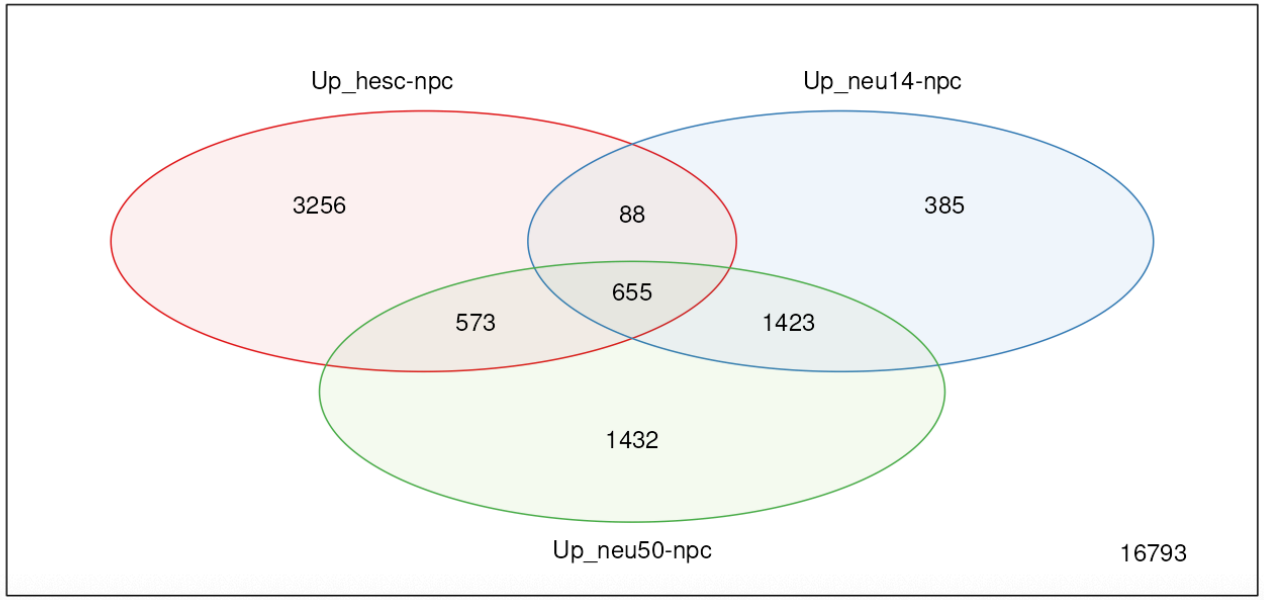
Threshold-based Wald Test

Independent filtering of lower counts

Results Gene Lists

Report





Investigate DEGs

Select a comparison:

neu50-npc

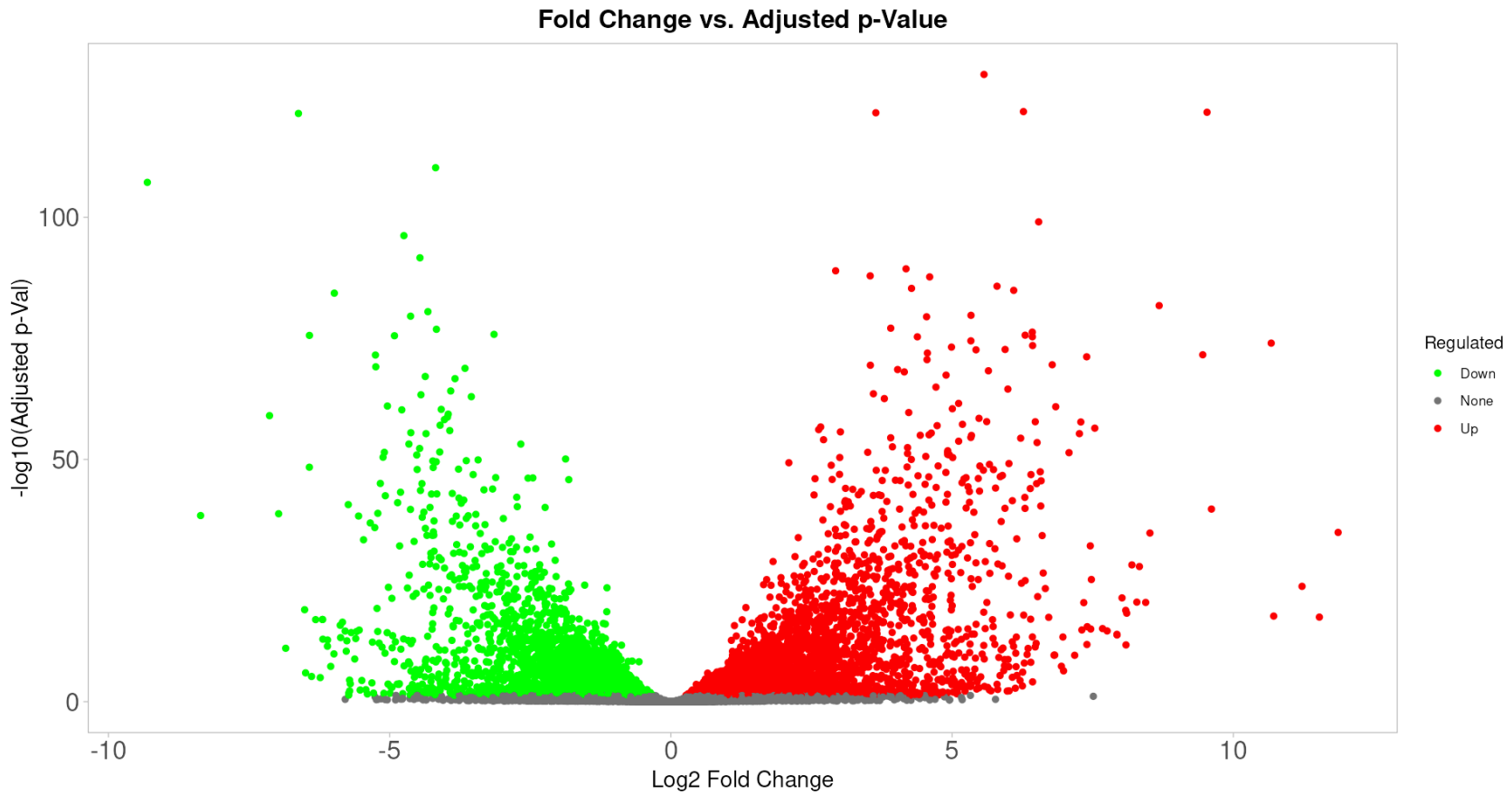
Up genes: 4,083; Down genes: 3,429

Green-Red

Label Genes

Report

Genes Heatmap Volcano Plot MA Plot Scatter Plot Enrichment



Investigate DEGs

Select a comparison:

neu50-npc

Up genes: 4,083; Down genes: 3,429

Report

Genes | Heatmap | Volcano Plot | MA Plot | Scatter Plot | Enrichment

KEGG | All Groups | More options

Pathways | Tree | Network | Plot | Genes | Details

Grp.	Adj.Pval	Fold	Pathway (Click for more info)
Upregulated	5.96E-13	1.9	Neuroactive ligand-receptor interaction
	1.02E-11	2	Calcium signaling pathway
	4.75E-9	2.6	Morphine addiction
	1.32E-8	2.3	Glutamatergic synapse
	1.32E-8	3.4	Nicotine addiction
	2.57E-8	2.5	GABAergic synapse
	3.70E-8	2.4	Hypertrophic cardiomyopathy
	5.94E-8	1.9	CAMP signaling pathway
	1.00E-7	2.1	Adrenergic signaling in cardiomyocytes
	1.45E-7	2.3	Dilated cardiomyopathy
Downregulated	3.47E-16	2.6	Cell cycle
	8.03E-7	3.3	DNA replication
	4.02E-6	3	Base excision repair
	4.58E-5	2.6	Fanconi anemia pathway
	4.86E-5	1.9	Spliceosome
	2.81E-3	2.5	Homologous recombination
	2.81E-3	2.2	Notch signaling pathway
	2.85E-3	1.5	Human papillomavirus infection
	3.34E-3	1.9	ATP-dependent chromatin remodeling
	3.49E-3	1.6	MicroRNAs in cancer

Most variable genes to include (< 3001)

10000

Soft Threshold

5

Min. Module Size

20

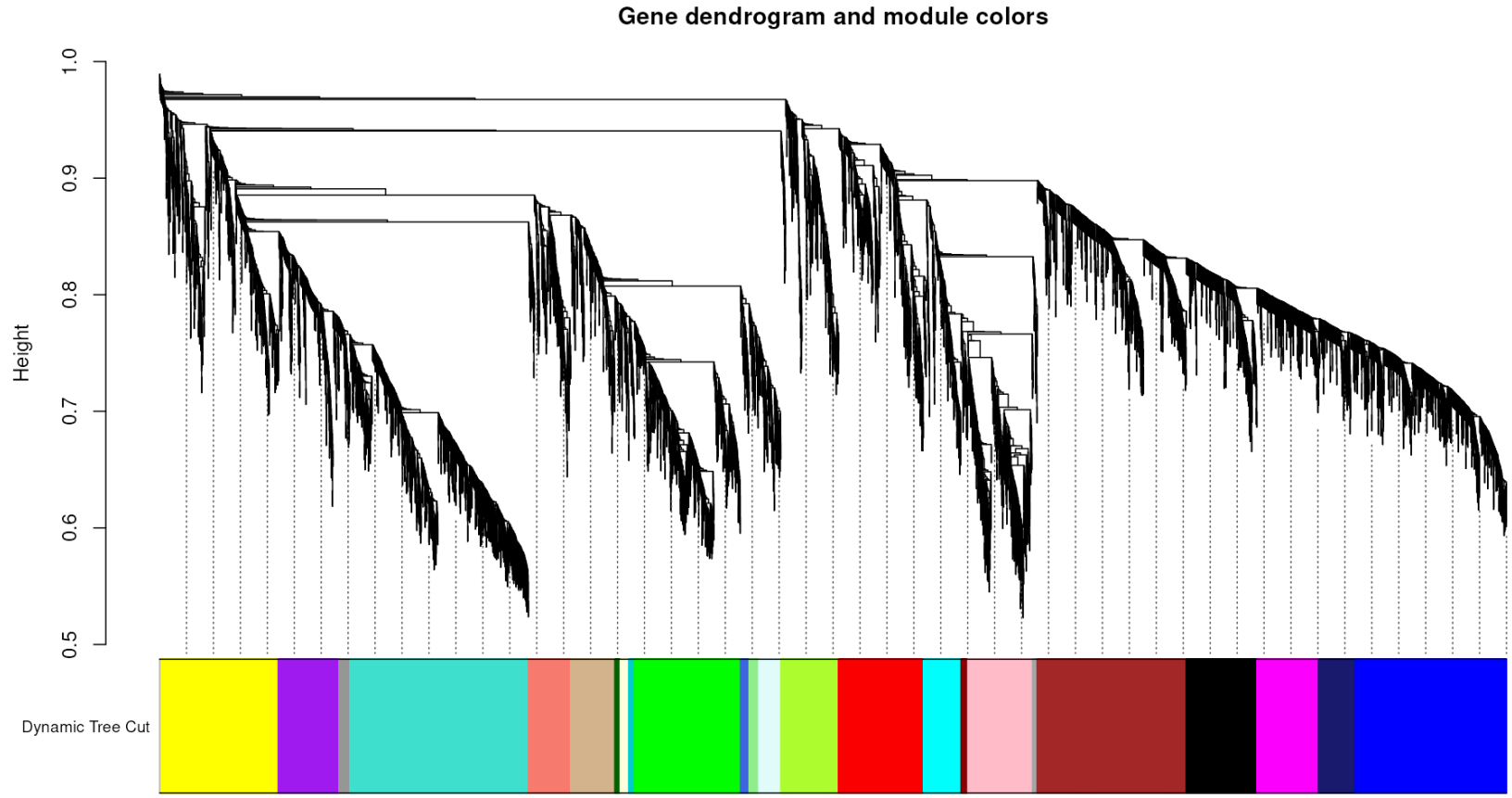
Select a module

1. turquoise (1324 genes)

All modules | Selected module

A network of 9981 genes was divided into 24 modules.

Network Plot | Module Plot | Heatmap | Enrichment | Scale Independence | Mean Connectivity



Single cell RNA-seq data analysis

Contents

- Basic concept of single cell RNA sequencing (scRNA-seq)
  - Bulk vs. single cell RNA-seq
  - Why scRNA-seq?
  - Droplet based scRNA-seq data
  - What dose scRNA-seq data look like?
  - What do we expect to get from the scRNA-seq data?

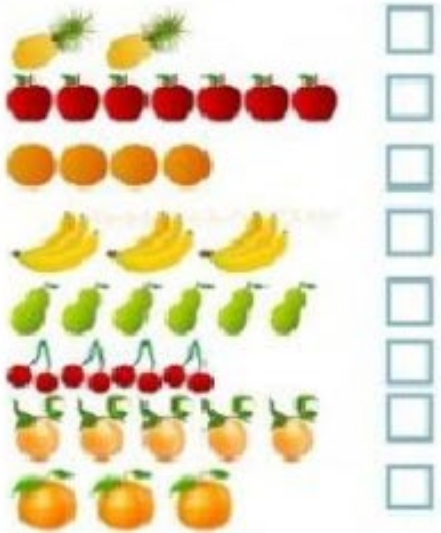
- Standard analysis workflow
  - Import → QC → dimensionality reduction → data correction → clustering → marker genes → cell type annotation → functional enrichment → gene expression dynamics (trajectory prediction)

BULK VS SINGLE CELL RNA-SEQ

RNA-Seq

scRNA-Seq

Average expression level
- Comparative transcriptomics
- Disease biomarker
- Homogenous systems



Separate populations
- Define heterogeneity
- Identify rare cell populations
- Cell population dynamics

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

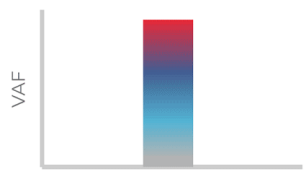
HETEROGENEOUS CELL POPULATION

BULK SEQUENCING

RNA IS MIXED TOGETHER

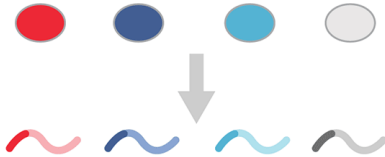
X

SUBPOPULATIONS NOT DEFINED

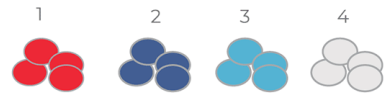


"AVERAGED" READOUT

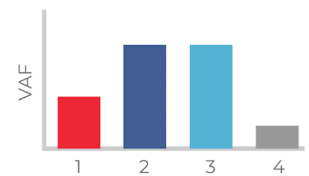
SINGLE-CELL RNA SEQUENCING



RNA FROM EACH CELL IS BARCODED



SUBPOPULATIONS DEFINED

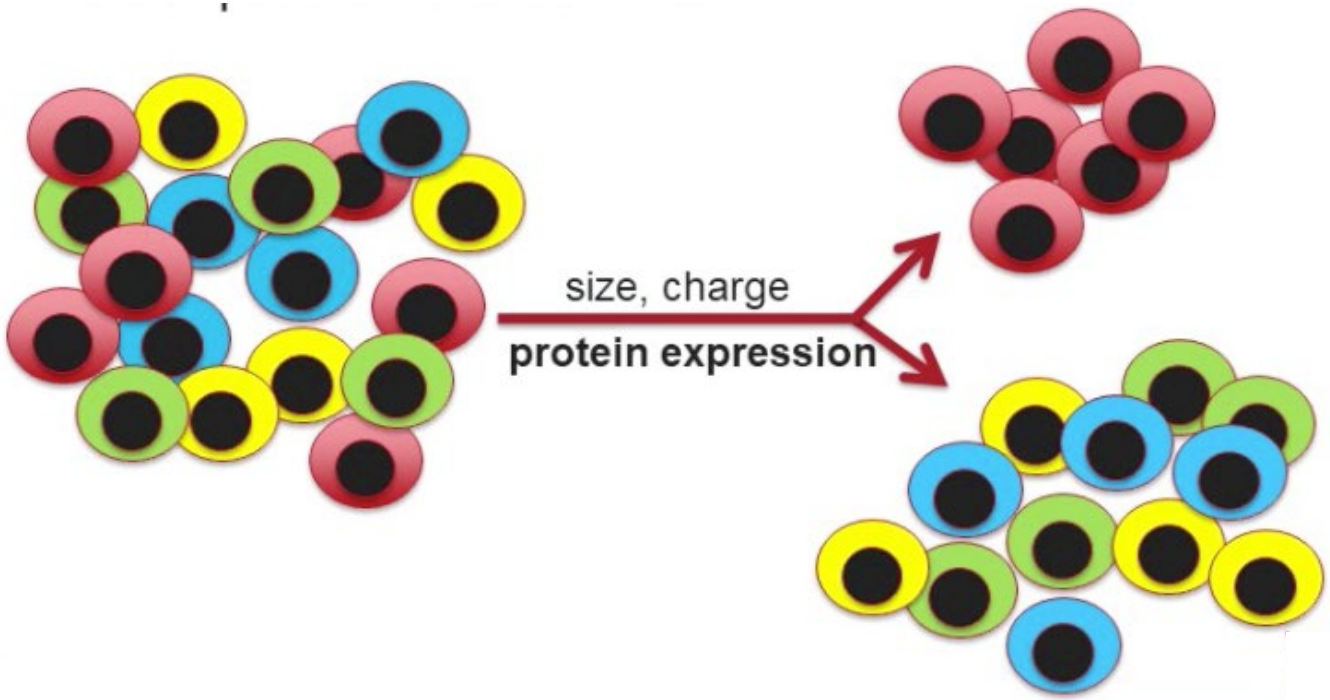


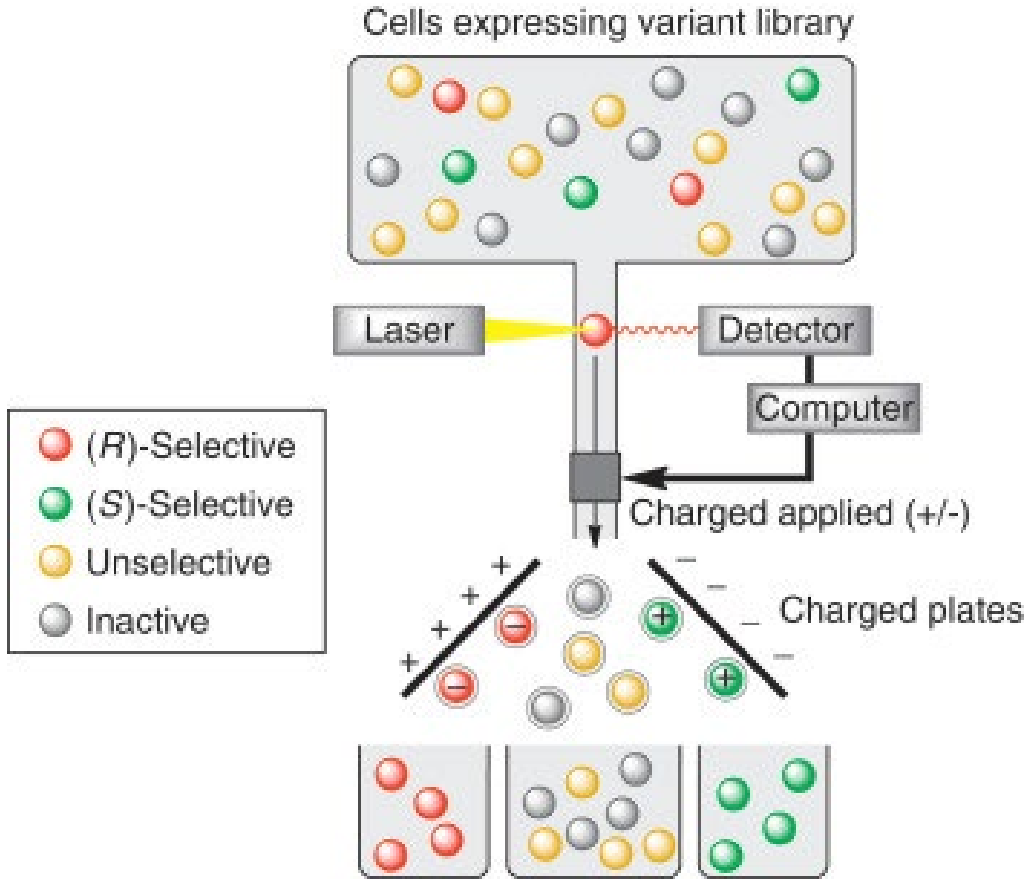
CELL TYPE-SPECIFIC READOUT

Why consider performing scRNA-seq?

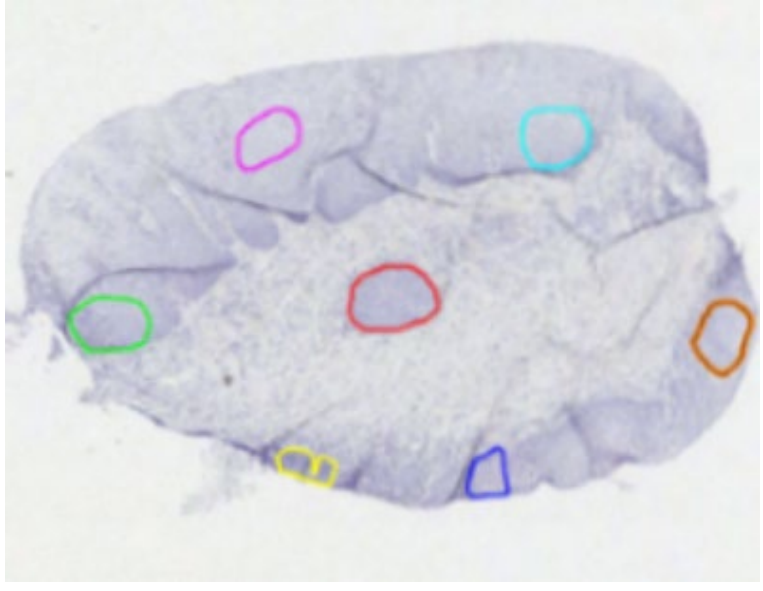
- scRNA-seq permits comparison of the transcriptomes of individual cells. Therefore, a major use of scRNA-seq has been to assess transcriptional similarities and differences within a population of cells, with early reports revealing previously unappreciated levels of heterogeneity, for example in embryonic and immune cells

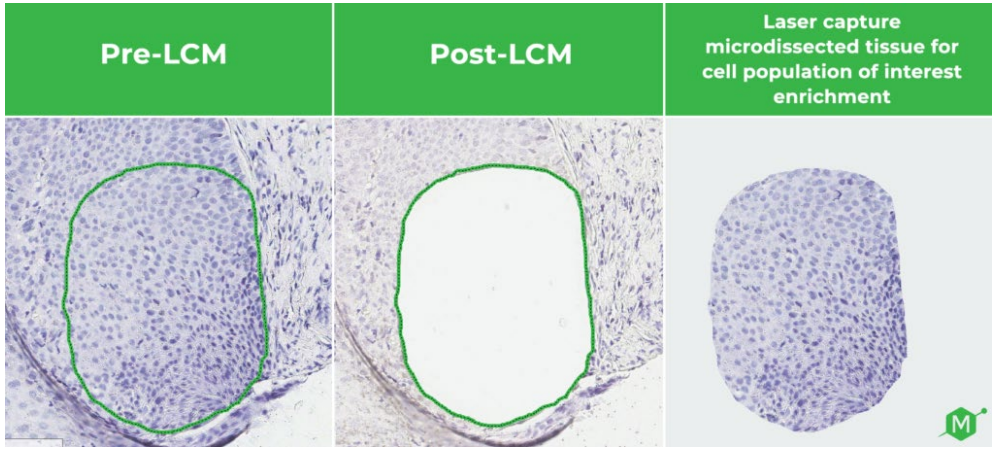
FACS (fluorescence activated cell sorting)

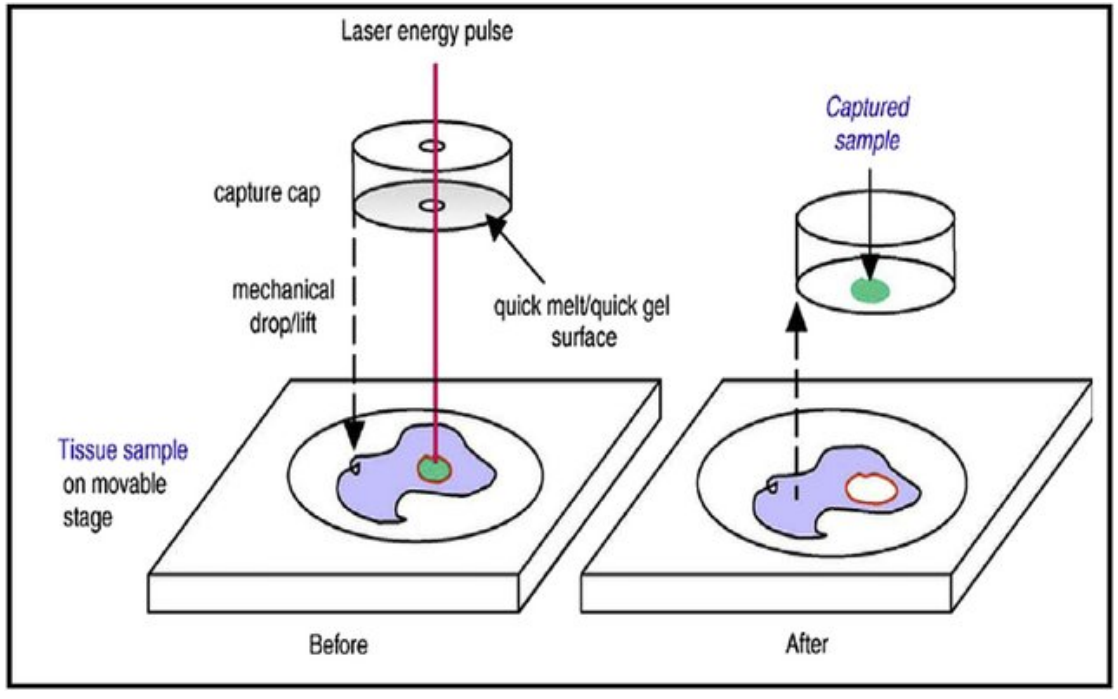




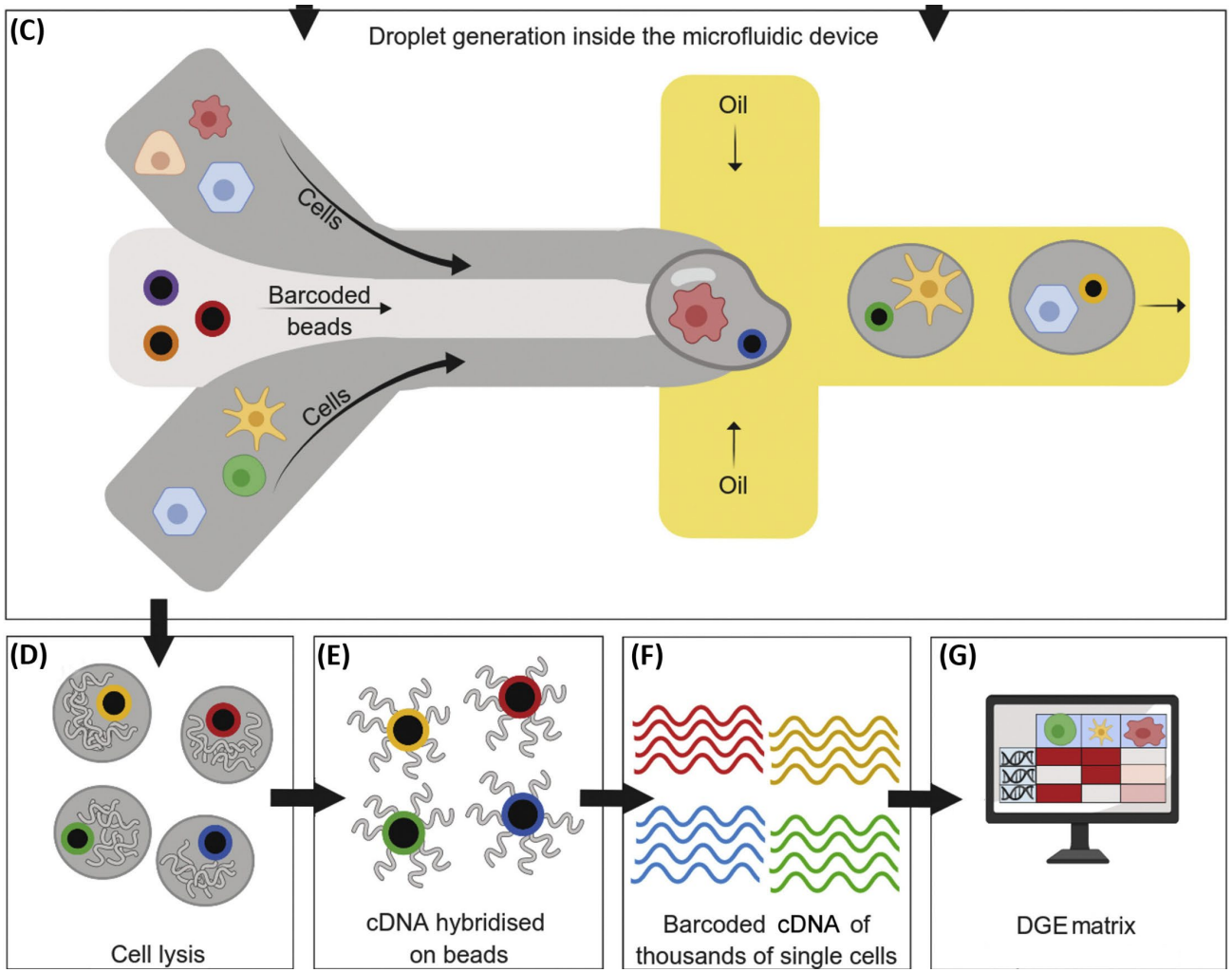
Laser capture microdissection (LCM)

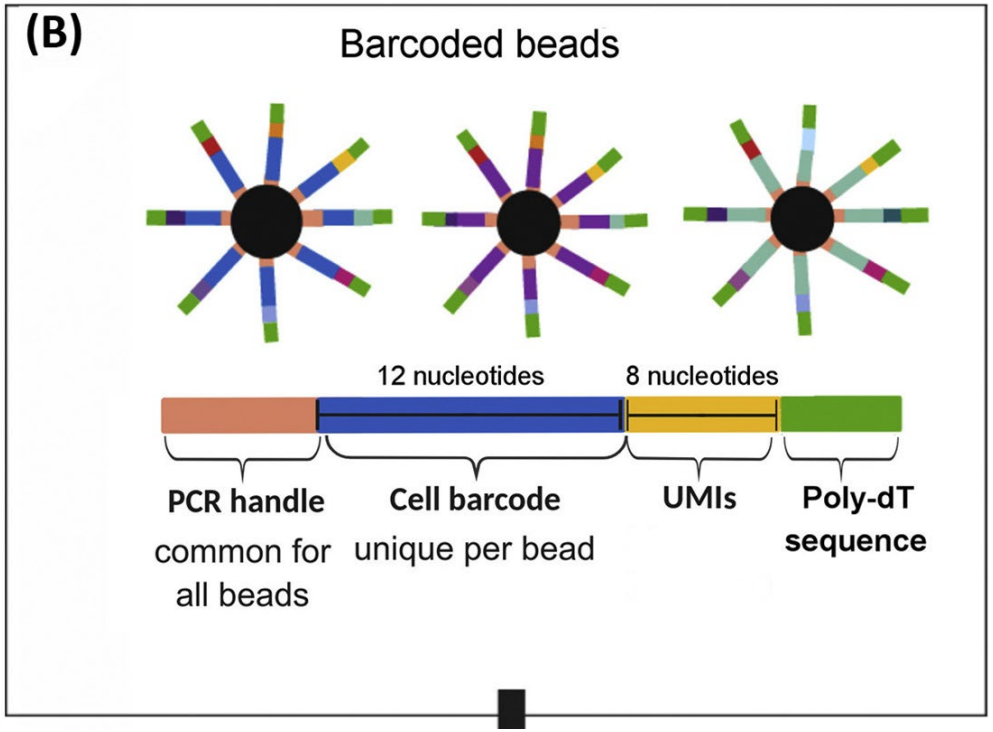




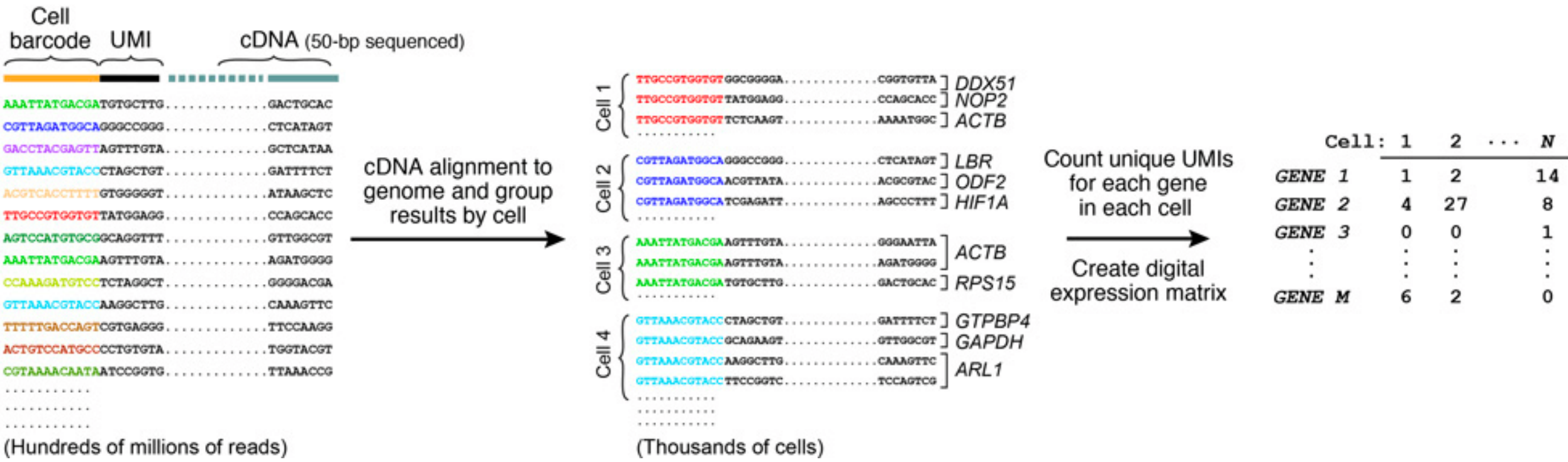


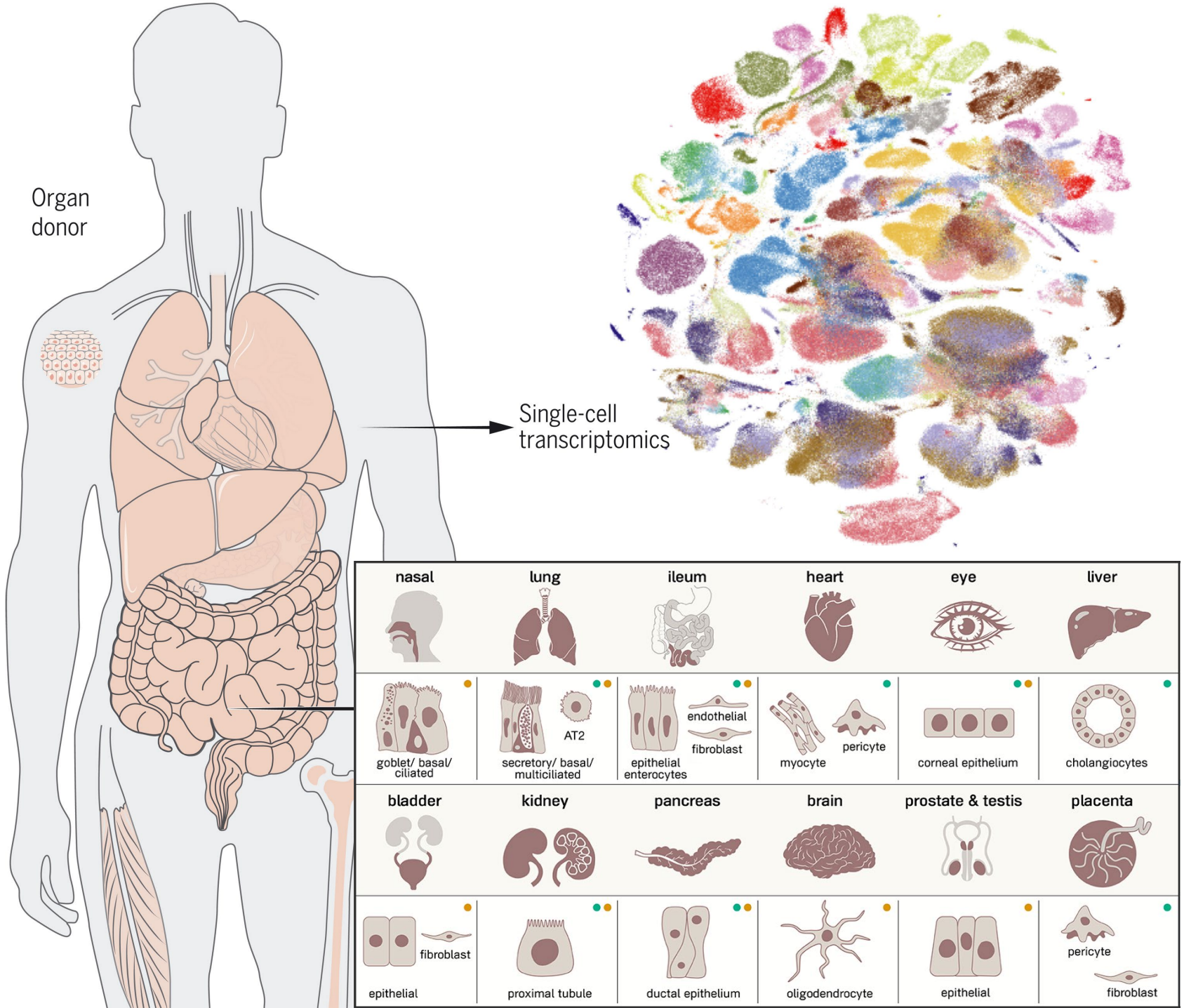
Droplet based single cell RNA sequencing





Droplet based single cell RNA sequencing





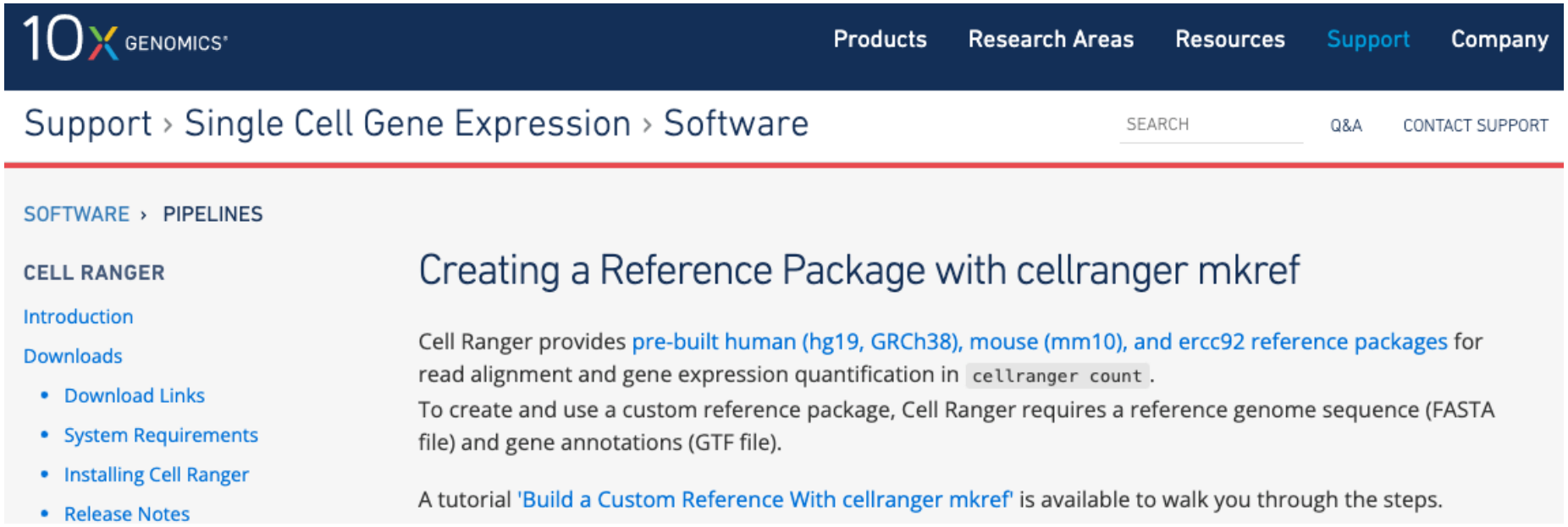
doi:10.1126/science.abl4896

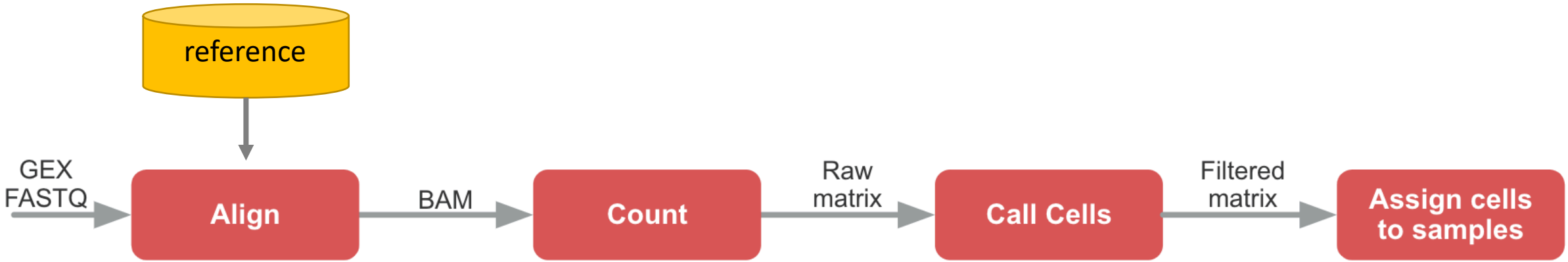
However, there are limitations in scRNA-seq:

- low capture efficiency (~8% mRNAs in a cell were captured)
- higher level of technical noise than bulk RNA-seq data
- multi-cells in a droplet (doublet)
- Cell stress and bias: dead cells (high proportion of mitochondrial RNAs), not every cell type can use droplet based scRNA-seq
- Cost and throughput

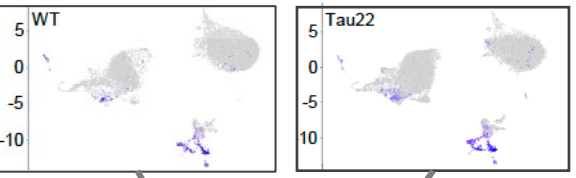
Standard scRNA-seq data analysis

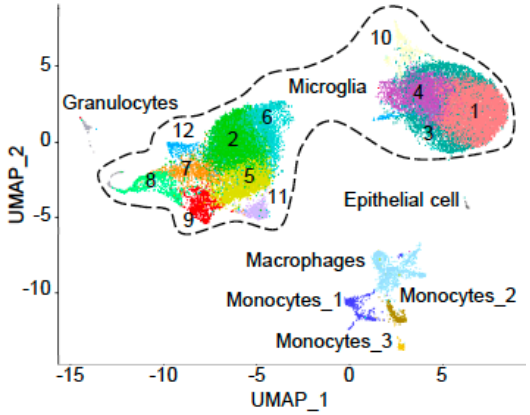
Read alignment (similar to bulk RNA-seq)

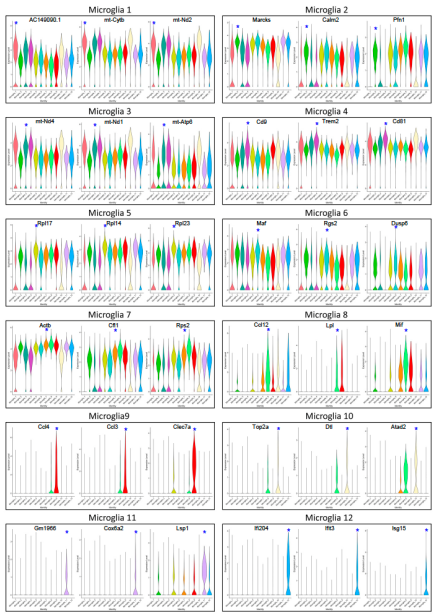




Asc-Seurat – Analytical single-cell Seurat-based web application







SEURAT

Sample
Sample
Sample
Sample
Sample
Integration pipeline
Quality Control
PCA
Clustering
Exclude or select clusters
Yes
Select clusters of interest?
No
Lists of markers
Markers identification
Differential expression analysis
Lists of DE genes
Expression visualization
High resolution plots

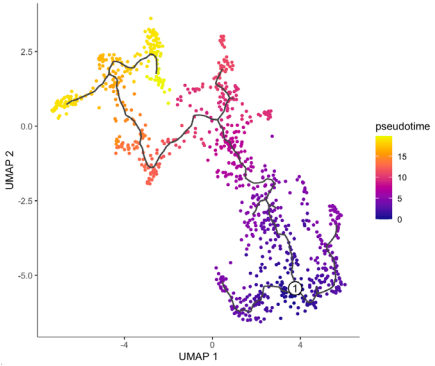
v2.1 May 26, 2021

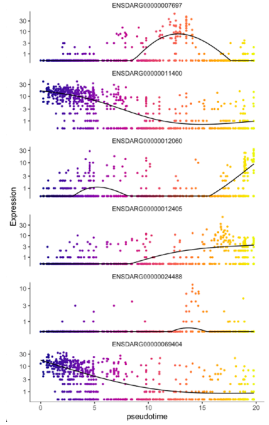
Dynverse

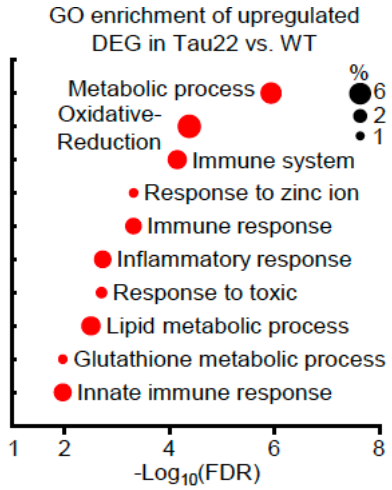
Clustered data
Trajectory inference
Identification of the most predictive genes
Expression visualization within the trajectory
High resolution plots
Lists of DE genes

BioMart

BioMart annotation
GO terms enrichment analysis
List of enriched GO terms
Annotated genes







Pereira et al., bioRxiv, 2021

RStudio

Go to file/function | Addins | Project: (None)

ingleR.R | Basic_analysis.R | Biomart.R | NOIseq_no_replicate_example_2.r | NOIseq_with_replicate_example.r | DEseq2_PCA_e

Source on Save | Run | Source

```
1  library(dplyr)
2  library(Seurat)
3  library(patchwork)
4
5  ##import data from 10x output folder, QC and selecting cells for further analysis
6  input.data <- Read10X(data.dir = "/Volumes/CONTAX/DR_RB_Yang_Scube_genes_expression_profiles_in_endothelial_scRNA-seq_da
7  W0 <- CreateSeuratObject(counts = input.data, project = "W0", min.cells = 3, min.features = 200)
8  W0
9  W0$source <- "W0"
10 W0[["percent.mt"]] <- PercentageFeatureSet(W0, pattern = "^mt-")
11 VlnPlot(W0, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
12 #W0 <- subset(W0, subset = nFeature_RNA > 200 & nFeature_RNA < 6000 & percent.mt < 15)
13 W0 <- NormalizeData(W0, verbose = FALSE)
14 W0 <- FindVariableFeatures(W0, selection.method = "vst")
15
16 ##import data from 10x output folder, QC and selecting cells for further analysis
17 input.data <- Read10X(data.dir = "/Volumes/CONTAX/DR_RB_Yang_Scube_genes_expression_profiles_in_endothelial_scRNA-seq_da
18 W1 <- CreateSeuratObject(counts = input.data, project = "W1", min.cells = 3, min.features = 200)
19 W1
20 W1$source <- "W1"
21 W1[["percent.mt"]] <- PercentageFeatureSet(W1, pattern = "^mt-")
22 VlnPlot(W1, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
23 W1 <- subset(W1, subset = nFeature_RNA > 200 & nFeature_RNA < 6000 & percent.mt < 15)
24 W1 <- NormalizeData(W1, verbose = FALSE)
25 W1 <- FindVariableFeatures(W1, selection.method = "vst")
26
```

16:50 (Top Level) R Script

Console | Terminal | Jobs

/Volumes/Extented-Data/My Drive/temp/LSL_course_sample_data/

```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Environment is empty

Files | Plots | Packages | Help | Viewer

R: Save a ggplot (or other grid object) with sensible defaults | Find in Topic

ggsave {ggplot2} R Documentation

Save a ggplot (or other grid object) with sensible defaults

Description

`ggsave()` is a convenient function for saving a plot. It defaults to saving the last plot that you displayed, using the size of the current graphics device. It also guesses the type of graphics device from the extension.

Usage

```
ggsave(
  filename,
  plot = last_plot(),
  device = NULL,
  path = NULL,
  scale = 1,
  width = NA,
  height = NA,
  units = c("in", "cm", "mm", "px"),
  dpi = 300,
  limitsize = TRUE,
  bg = NULL,
  ...
)
```

Arguments

Web-based scRNA-seq data analysis tool

Tool name	Published on	Journal name	Data	Need registration
SingleCAnalyzer	23 May 2022	Frontiers in Bioinformatics	FASTQ	yes
ICARUS	10 May 2022	Nucleic Acids Research	gene-cell count matrix	no
SC1	5 Aug 2021	Journal of Computational Biology	gene-cell count matrix	no

Published online 10 May 2022 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* ***W427–W433***
https://doi.org/10.1093/nar/gkac322

ICARUS, an interactive web server for single cell RNA-seq analysis

Andrew Jiang[1,*], Klaus Lehnert[1], Linya You[2] and Russell G. Snell[1]

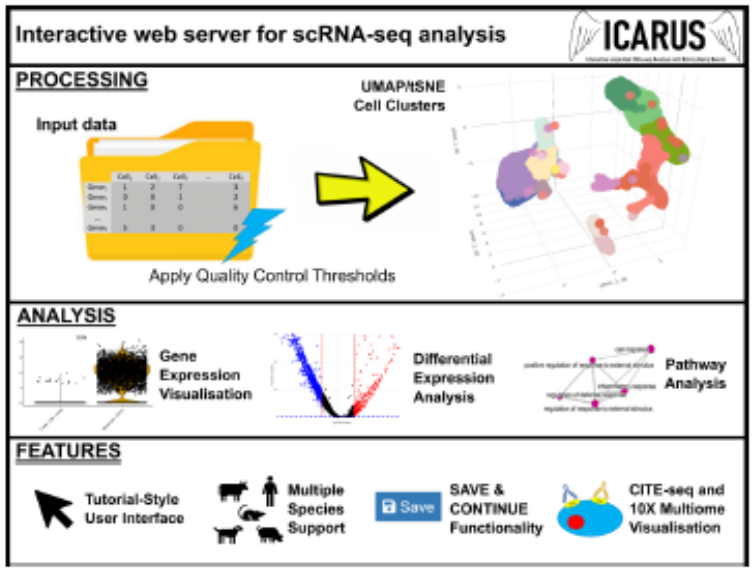
[1]Applied Translational Genetics Group, School of Biological Sciences, The University of Auckland, Auckland, New Zealand and [2]Department of Human Anatomy & Histoembryology, School of Basic Medical Sciences, Fudan University, Shanghai, China

Received March 24, 2022; Revised April 14, 2022; Editorial Decision April 19, 2022; Accepted April 21, 2022

ABSTRACT

Here we present ICARUS, a web server to enable users without experience in R to undertake single cell RNA-seq analysis. The focal point of ICARUS is its intuitive tutorial-style user interface, designed to guide logical navigation through the multitude of pre-processing, analysis and visualization steps. ICARUS is easily accessible through a dedicated web server (https://launch.icarus-scrnaseq.cloud.edu.au/) and avoids installation of software on the user's computer. Notable features include the facility to apply quality control thresholds and adjust dimensionality reduction and cell clustering parameters. Data is visualized through 2D/3D UMAP and t-SNE plots and may be curated to remove potential confounders such as cell cycle heterogeneity.

GRAPHICAL ABSTRACT



Welcome to ICARUS (**I**nteractive single **C**ell RNA-seq **A**nalysis with **R** shiny **U**sing **S**eurat)

This application was designed to guide the user through single cell RNA-seq analysis using the Seurat scRNA-seq analysis toolkit via a tutorial style interface. It offers user control over each of the steps to personalise analysis based on the dataset of interest. Graphical outputs at each analysis step ensures easy and logical interpretation.

The purpose of this application is to allow the user to interactively visualize single cell RNA-seq data without the requirement of previous R programming knowledge.

Features include:

1. Tutorial inspired user interface!
2. Support for 11 common species!
3. Adjust your own quality control thresholds!
4. Adjust your own dimensionality reduction and clustering parameters!
5. 3D UMAP and t-SNE plots!
6. Data correction for cell cycle effects!
7. Removal of cell doublets (multiplets) with DoubletFinder!
8. Labelling of cell clusters with sctype and SingleR!
9. Gene expression and gene pathway visualisation!
10. Trajectory analysis with Monocle3!
11. Differential expression analysis and gene set enrichment analysis with ClusterProfiler and ReactomePA!
12. Custom differential expression analysis with user selected cell groups to compare!
13. Integration with second dataset and adjustment for batch effects!
14. Support for multimodal analysis (i.e. CITE-seq, 10X multiome kit)!
15. Save and continue functionality!
16. Downloadable tables and plots!

Please refer to the "Help" tab on the sidebar menu for troubleshooting.

NEW

Choose from the following species

HOMO SAPIENS

START

PRESS START TO BEGIN

CONTINUE

Load previously saved enviroment

BROWSE... No file selecte

Applied Translational Genetics

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | CENTRE FOR BRAIN RESEARCH

Created by: Andrew Jiang
Email: ajia169@aucklanduni.ac.nz
Applied Translational Genetics Group
Centre for Brain Research
The University of Auckland

Step 1: Load your data

ICARUS

Interactive single Cell RNA-seq Analysis with Rshiny Using Seurat

DATA INPUT

A

Single sample

	$Cell_1$	$Cell_2$	$Cell_3$	...	$Cell_X$
$Gene_1$	1	2	7		
$Gene_2$	0	0	1		
$Gene_3$	1	0	0		6
...					
$Gene_X$	5	3	0		0

Column names must not contain any underscores

B

Multiple samples

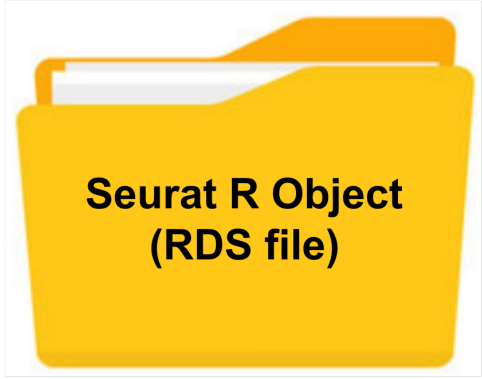
	$S1\_Cell_1$	$S1\_Cell_2$	$S2\_Cell_3$	...	$S2\_Cell_X$
$Gene_1$	1	2	7		
$Gene_2$	0	0	1		
$Gene_3$	1	0	0		6
...					
$Gene_X$	5	3	0		0

Each sample can be denoted by an identifier separated by an underscore

C

10x GENOMICS®

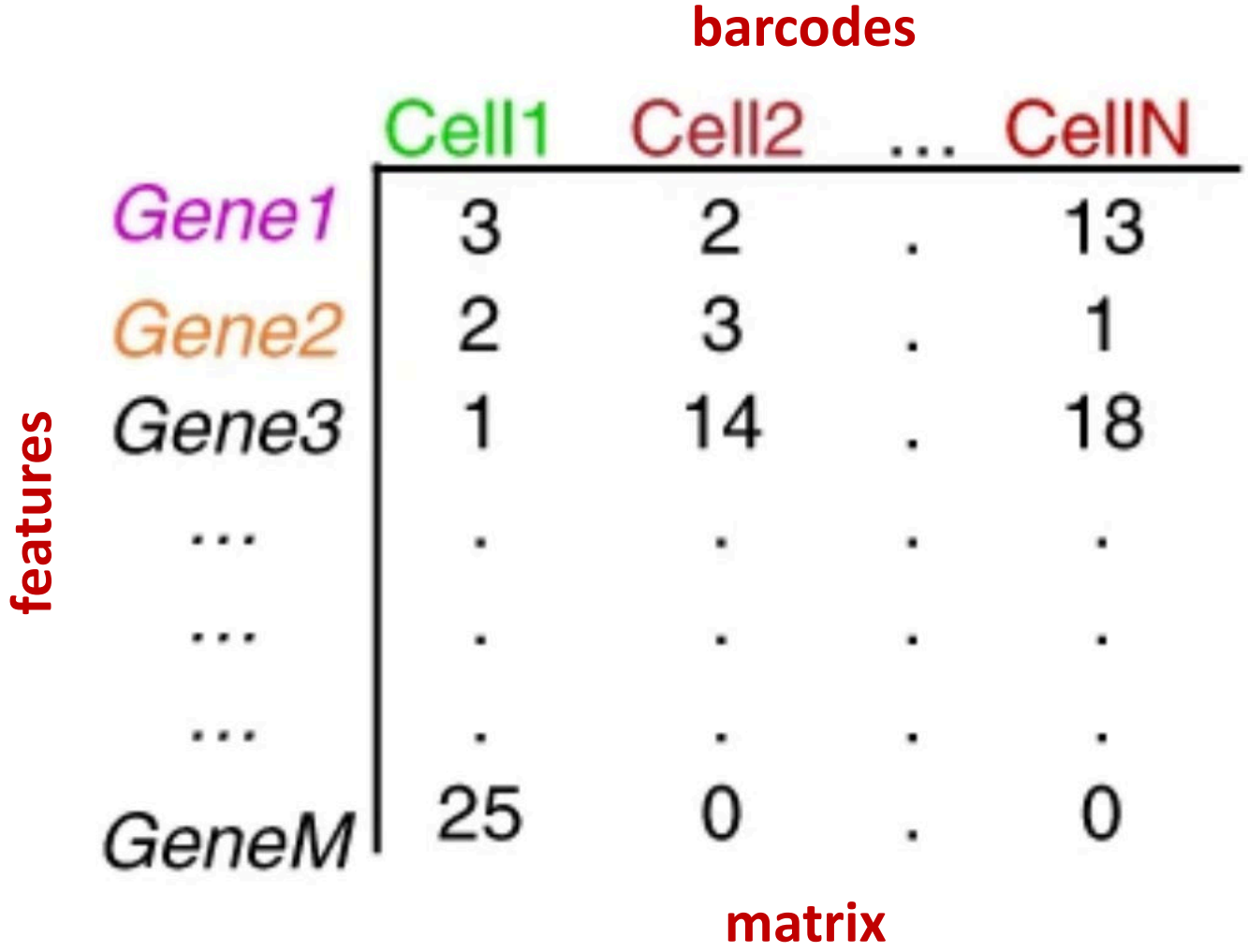
D



R

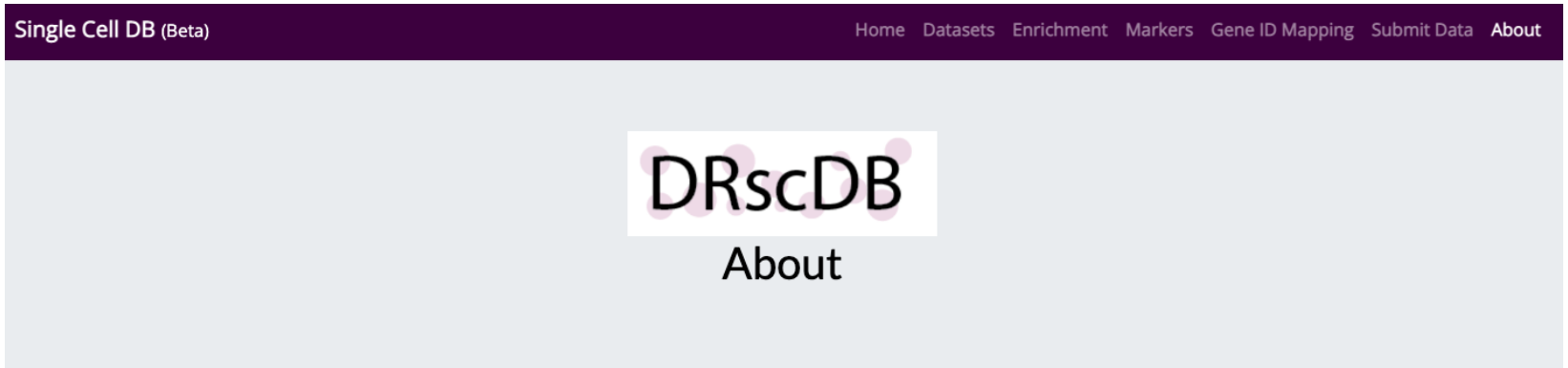
Processed data from 10X cellranger

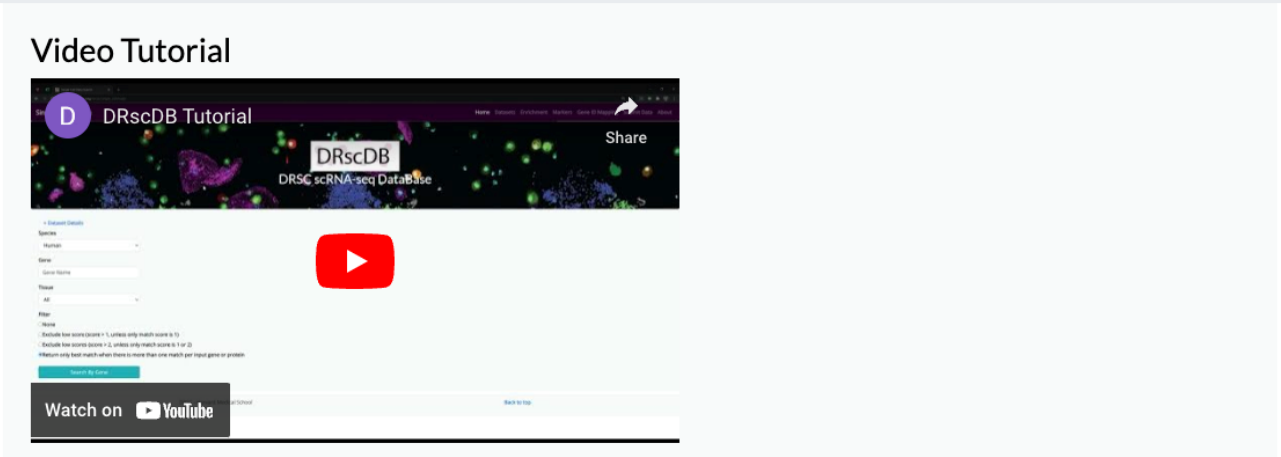
```
$ cd /home/jdoe/runs/sample345/outs
$ tree filtered_feature_bc_matrix
filtered_feature_bc_matrix
├── barcodes.tsv.gz
├── features.tsv.gz
└── matrix.mtx.gz

0 directories, 3 files
```



Download gene-cell matrix from database

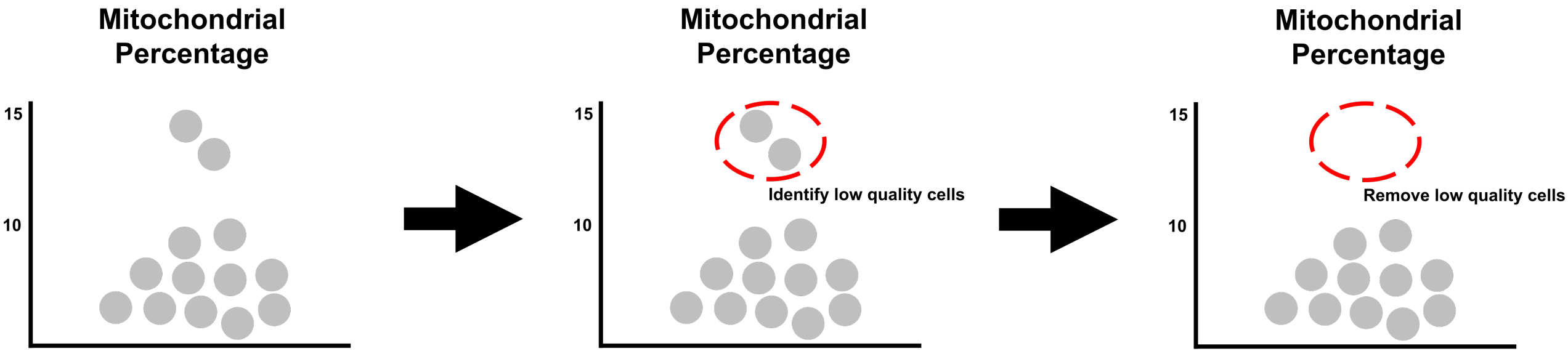
Single Cell DB

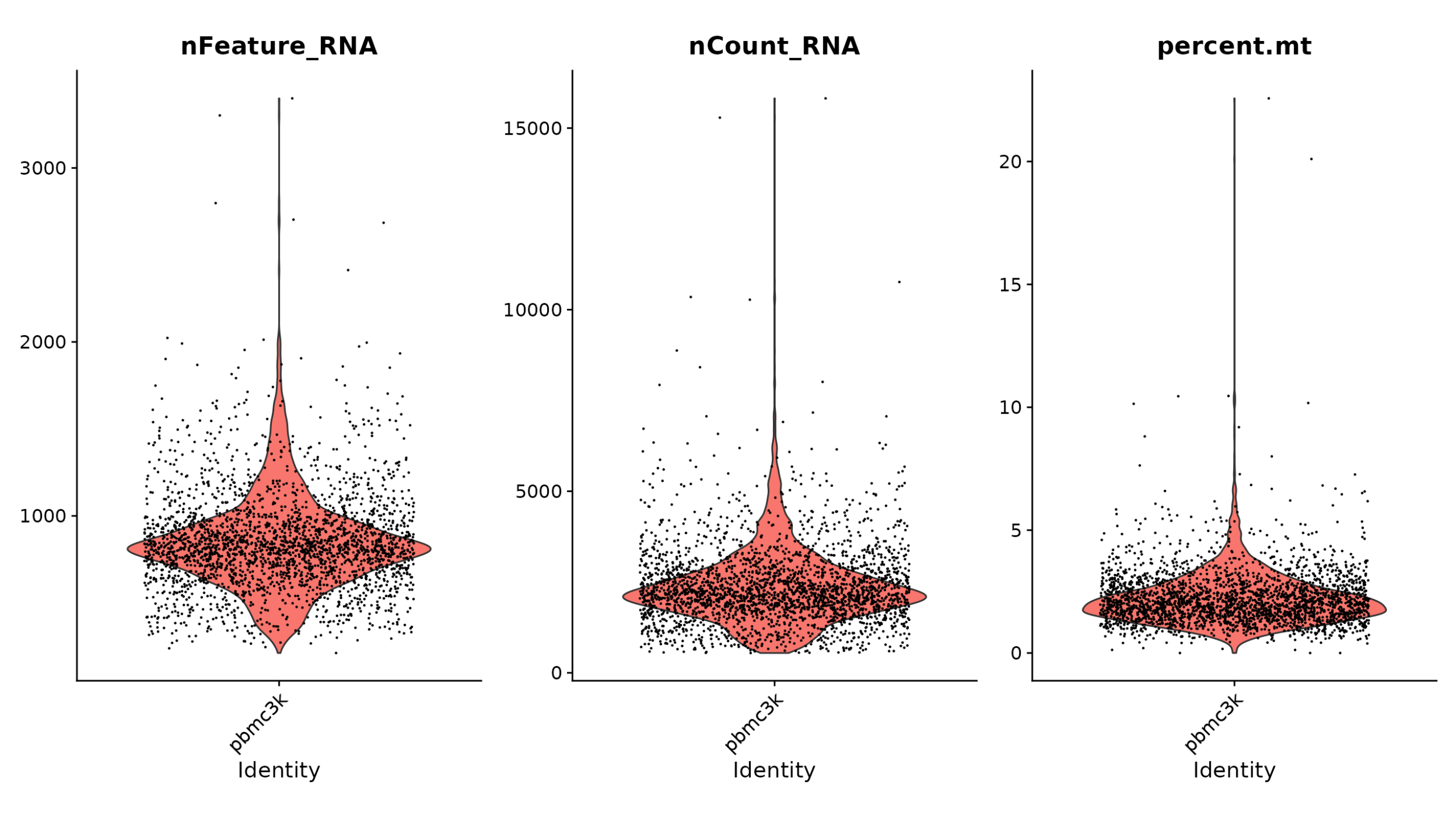


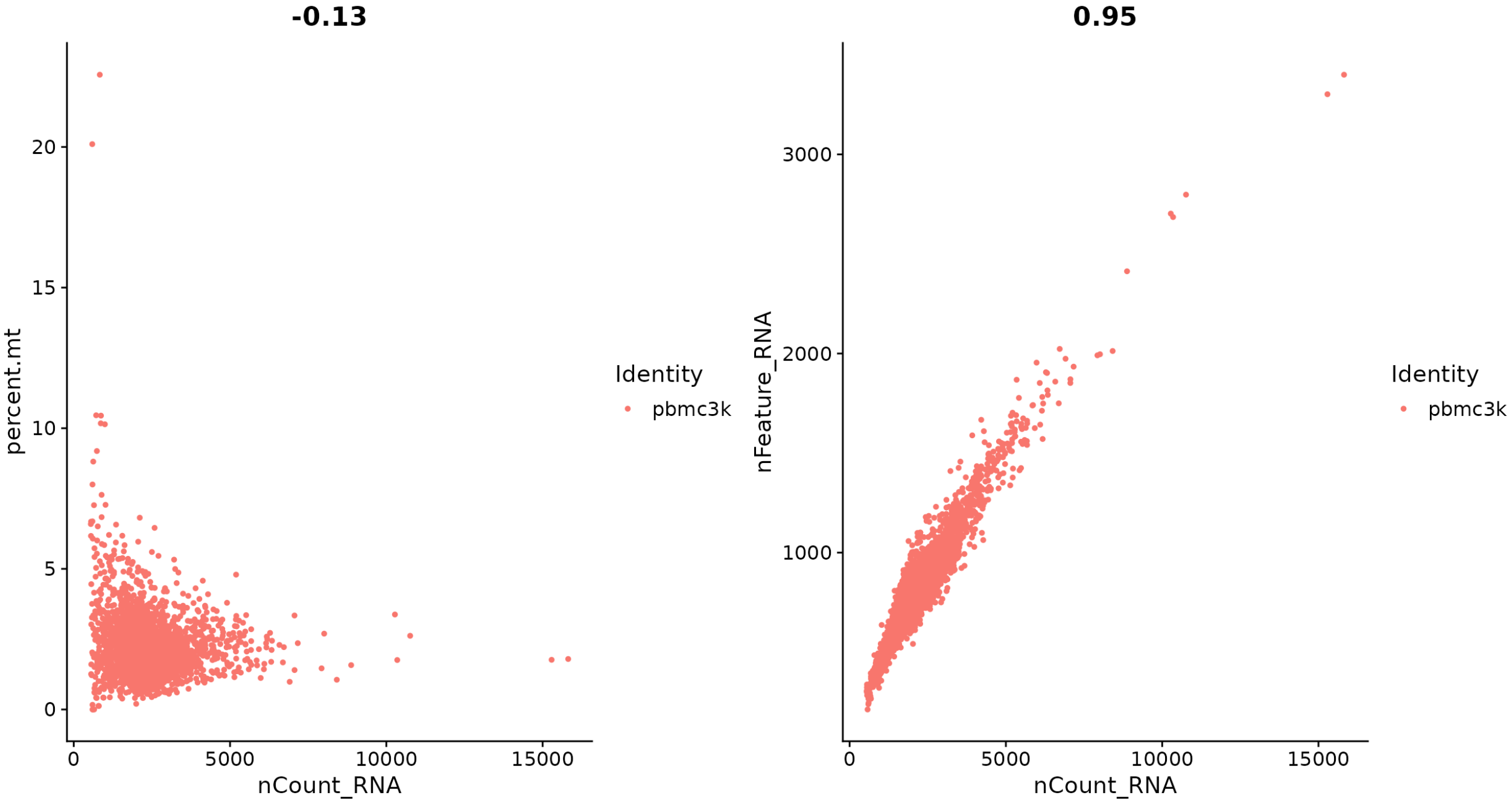


Step 2: Quality Control

QUALITY CONTROL WORKFLOW

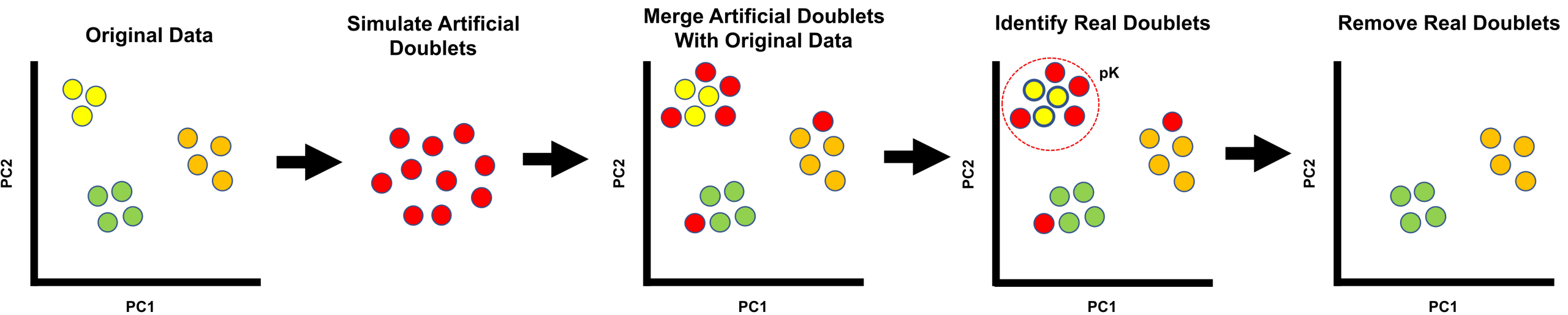



nFeature_RNA
3000
2000
1000
pbmc3k
Identity
nCount_RNA
15000
10000
5000
0
pbmc3k
Identity
percent.mt
20
15
10
5
0
pbmc3k
Identity

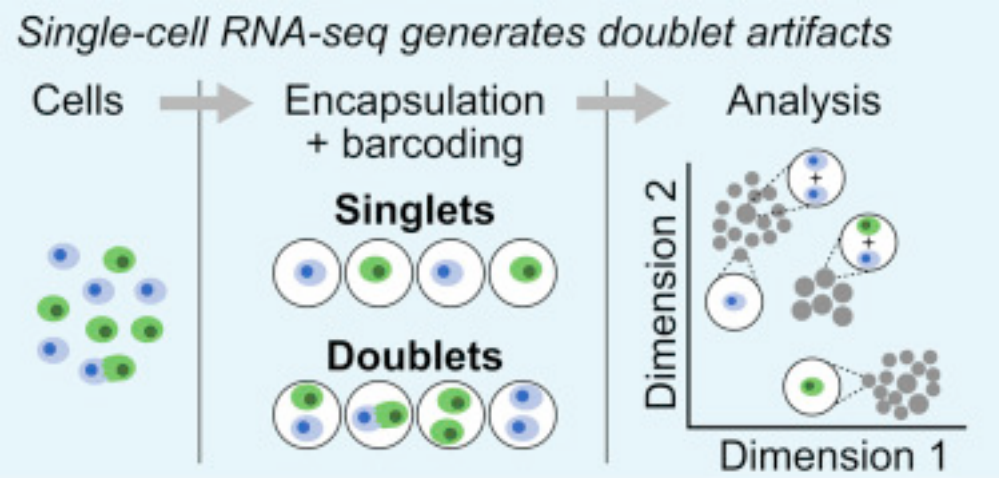

-0.13
percent.mt
20
15
10
5
0
0
5000
10000
15000
nCount_RNA
Identity
pbmc3k
0.95
nFeature_RNA
3000
2000
1000
0
5000
10000
15000
nCount_RNA
Identity
pbmc3k

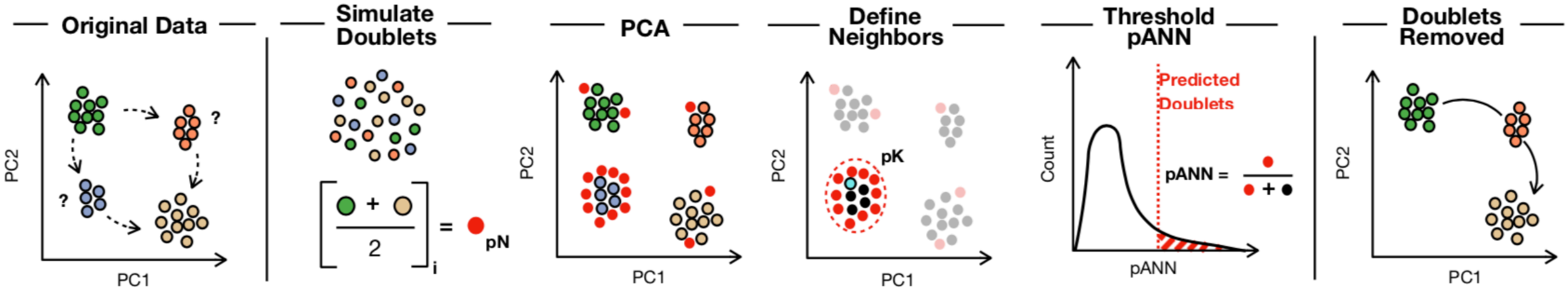
Step 3: Doublet Removal

DOUBLETFINDER WORKFLOW

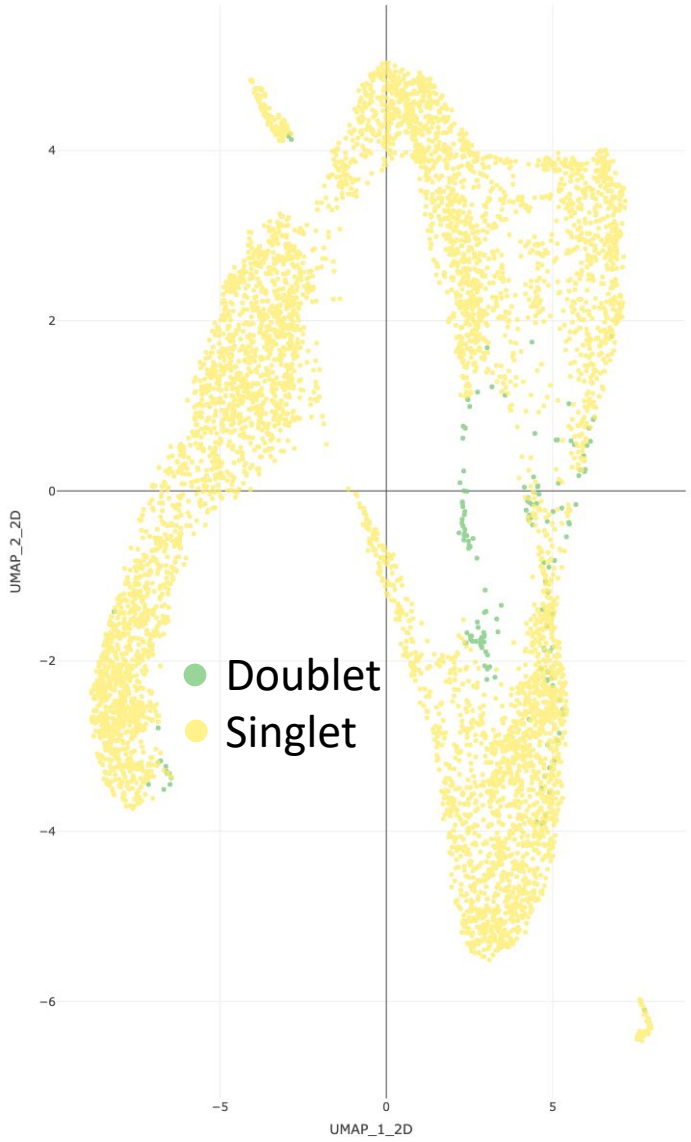


Droplet based single cell RNA sequencing

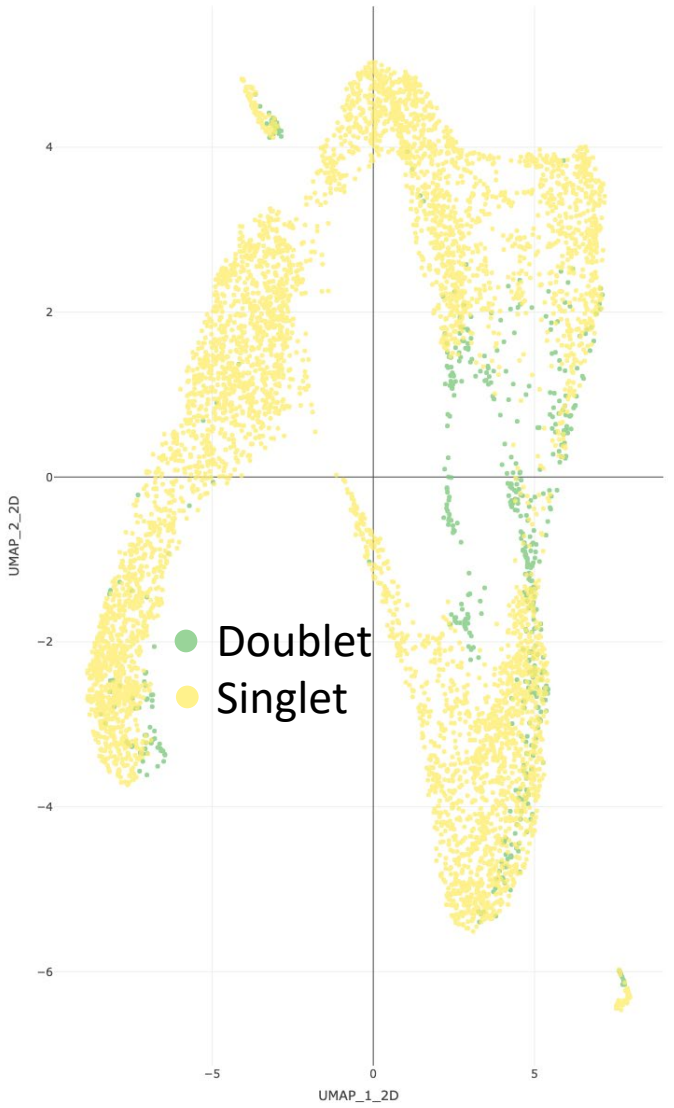




3% of doublets



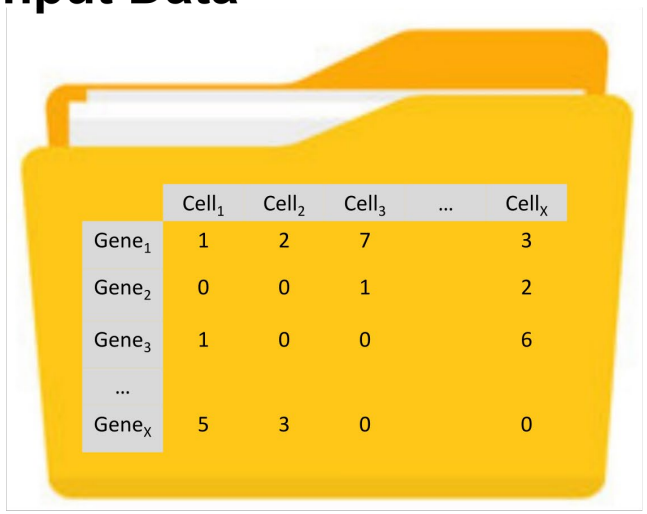
10% of doublets



Step 4: Dimensionality Reduction

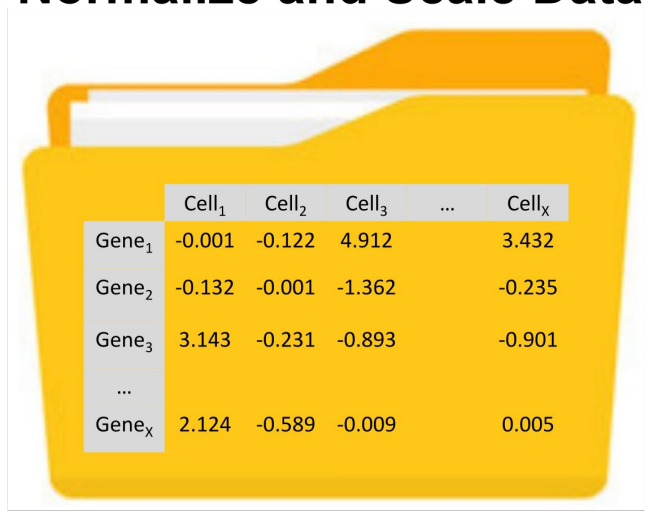
DIMENSIONALITY REDUCTION WORKFLOW

Input Data



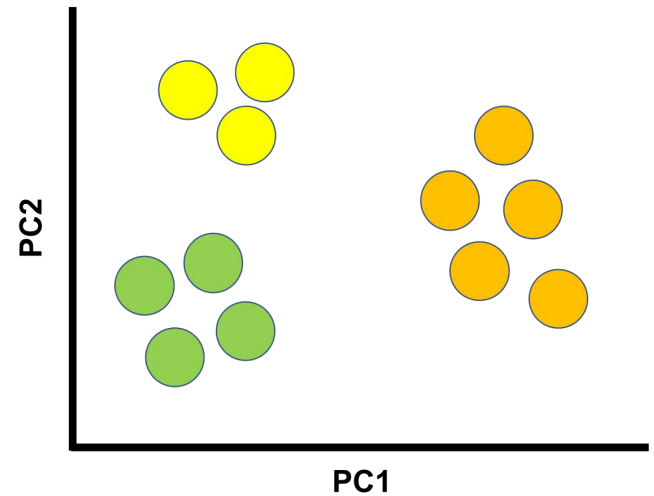
	$Cell_1$	$Cell_2$	$Cell_3$	...	$Cell_X$
$Gene_1$	1	2	7		3
$Gene_2$	0	0	1		2
$Gene_3$	1	0	0		6
...					
$Gene_X$	5	3	0		0

Normalize and Scale Data

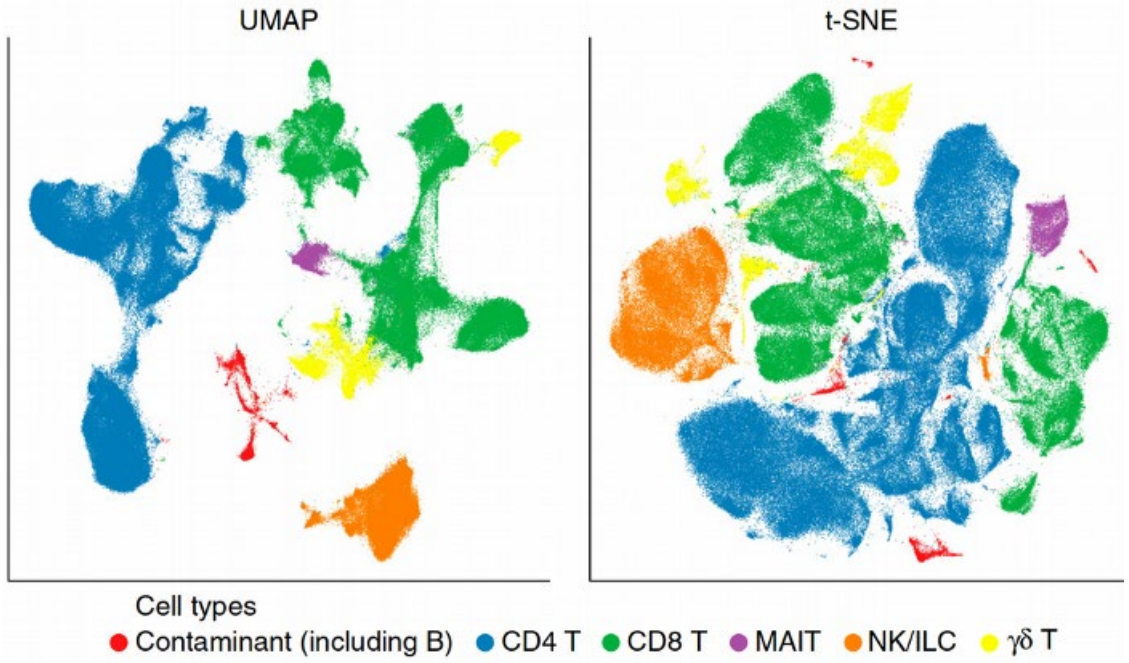


	$Cell_1$	$Cell_2$	$Cell_3$	...	$Cell_X$
$Gene_1$	-0.001	-0.122	4.912		3.432
$Gene_2$	-0.132	-0.001	-1.362		-0.235
$Gene_3$	3.143	-0.231	-0.893		-0.901
...					
$Gene_X$	2.124	-0.589	-0.009		0.005

Dimensionality Reduction



Dimensionality reduction



- PCA (principal component analysis)

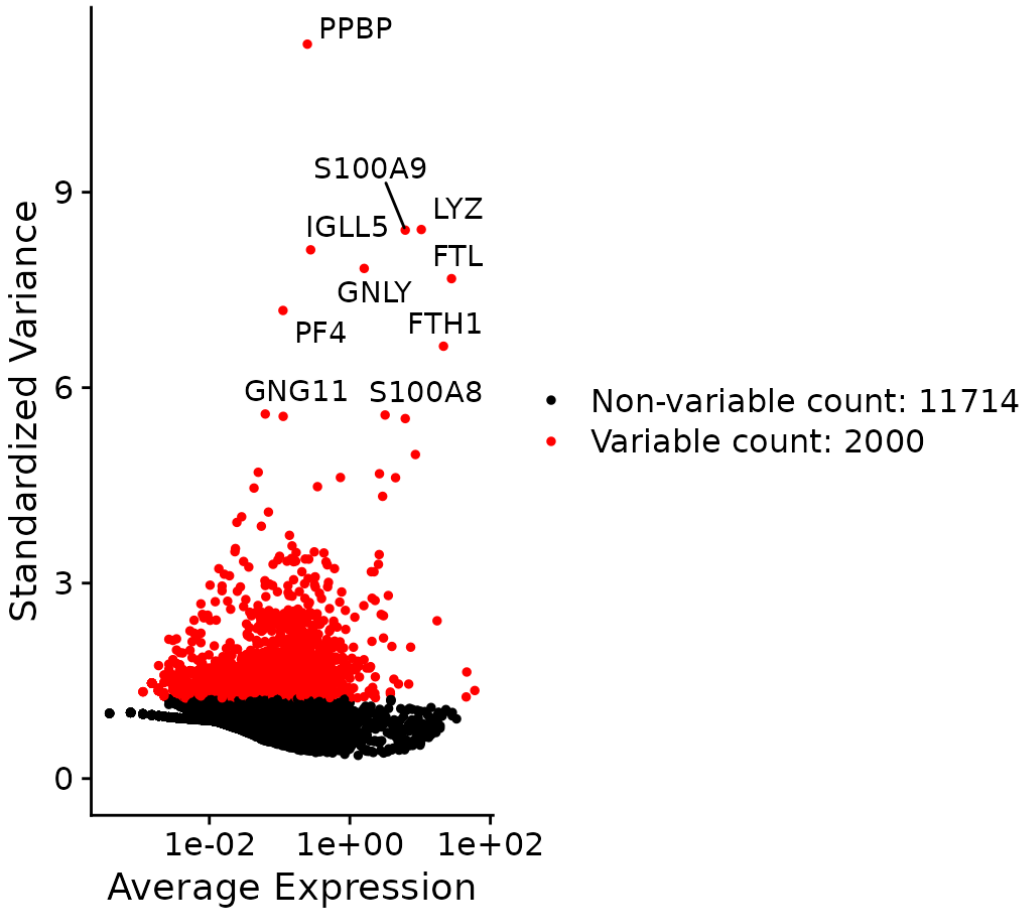
Visualization:

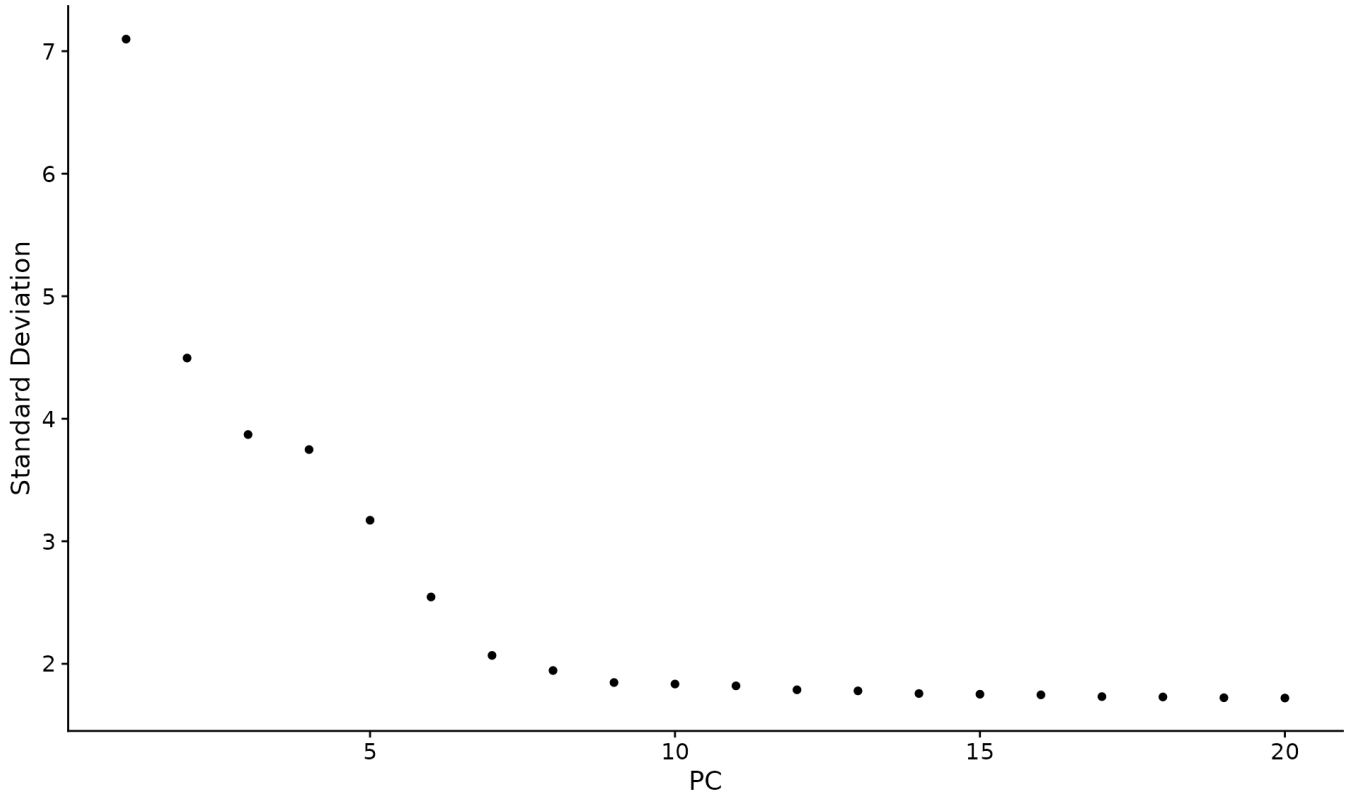
- tSNE (t-distributed Stochastic Neighbor Embedding)

  #2008 #old #slow

- UMAP (Uniform Manifold Approximation and Projection)

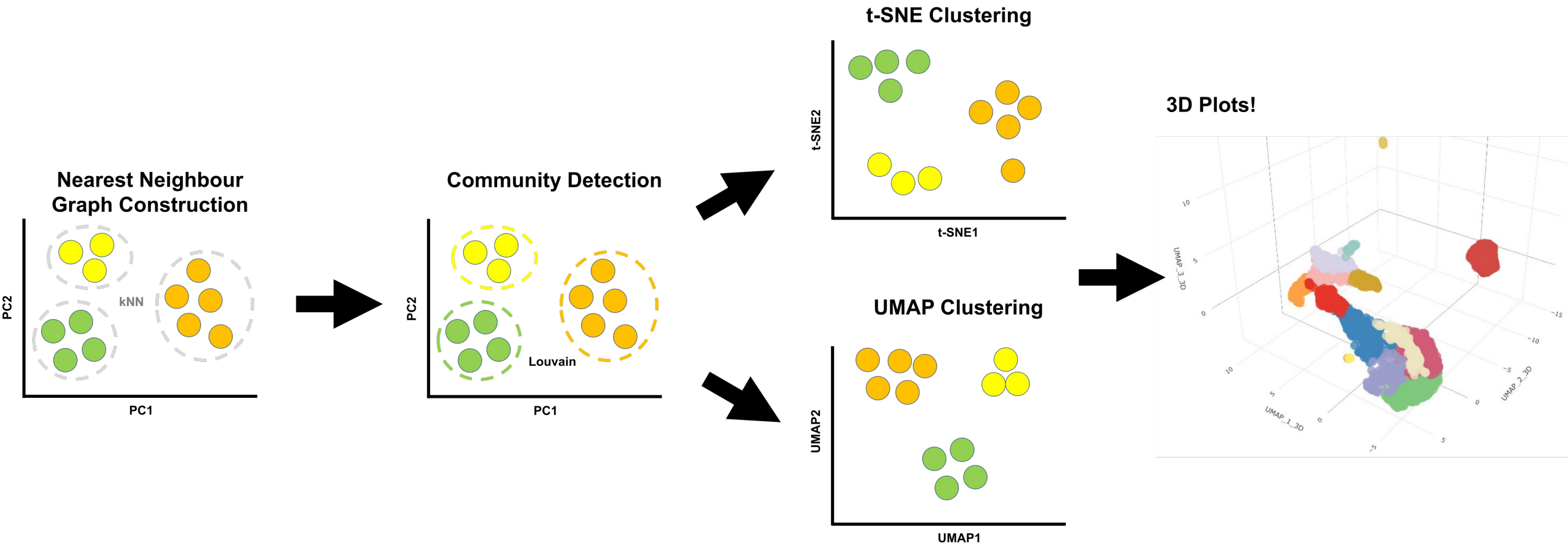
  #2018 #new #fast

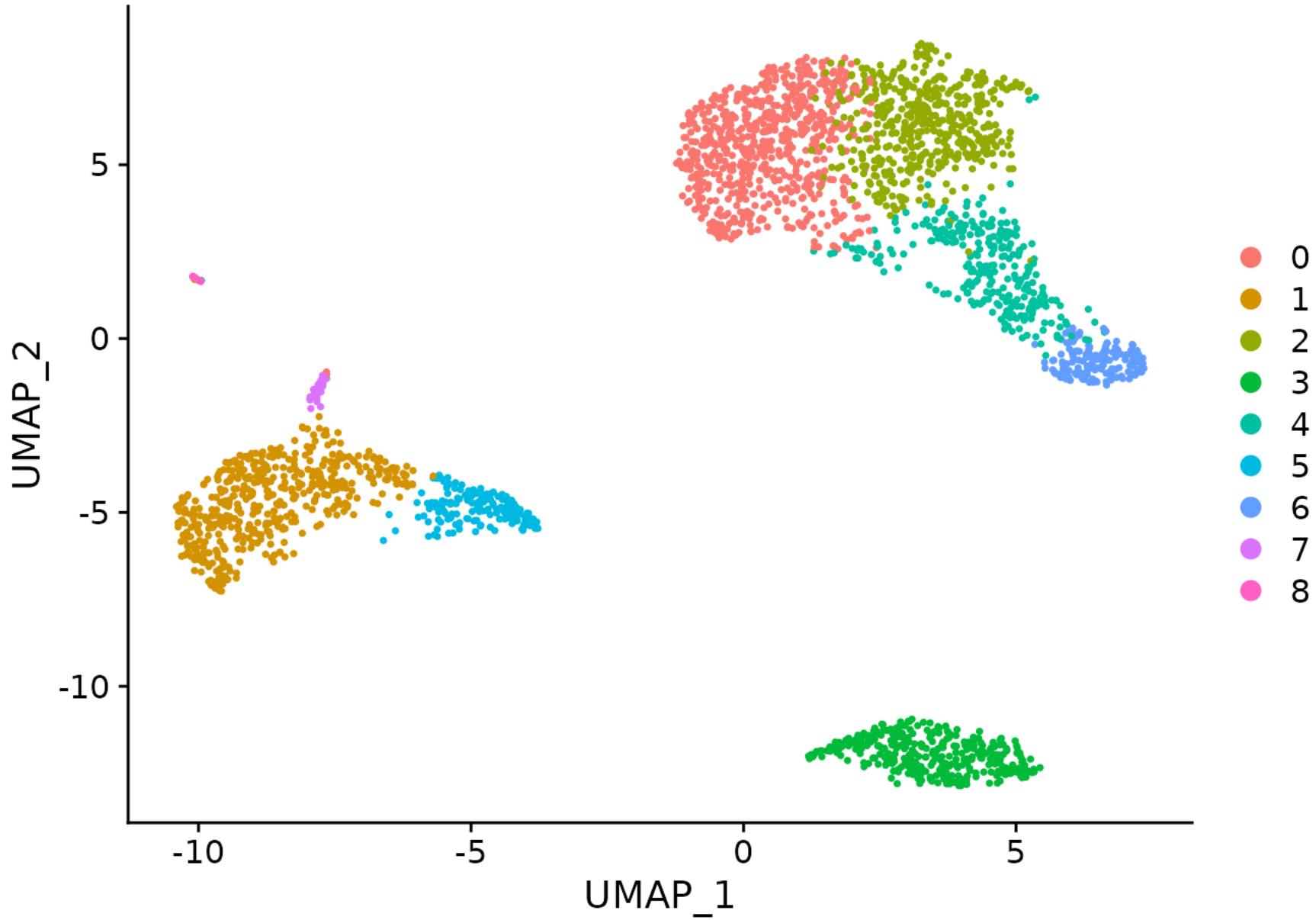

PPBP
S100A9
LYZ
IGLL5
FTL
GNLY
PF4
FTH1
GNG11
S100A8
Standardized Variance
9
6
3
0
1e-02
1e+00
1e+02
Average Expression
Non-variable count: 11714
Variable count: 2000


Standard Deviation
7
6
5
4
3
2
5
10
15
20
PC

Step 5: Clustering

CLUSTERING WORKFLOW



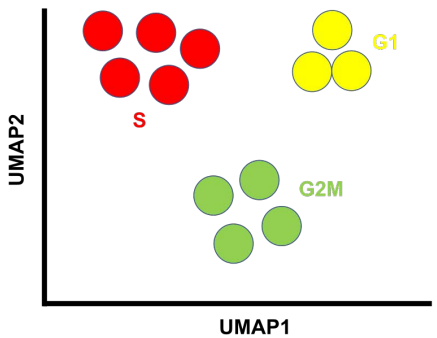

5
0
-5
-10
UMAP_2
-10
-5
0
5
UMAP_1
0
1
2
3
4
5
6
7
8

Step 6: Data Correction

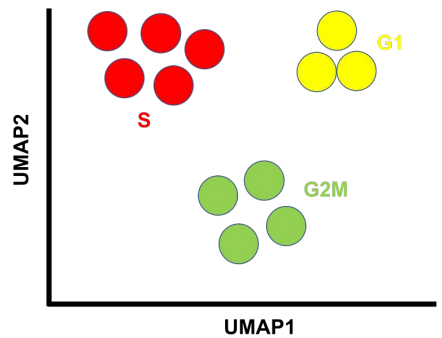
DATA CORRECTION WORKFLOW

Cell Cycle Effect

Cell Cycle Scoring

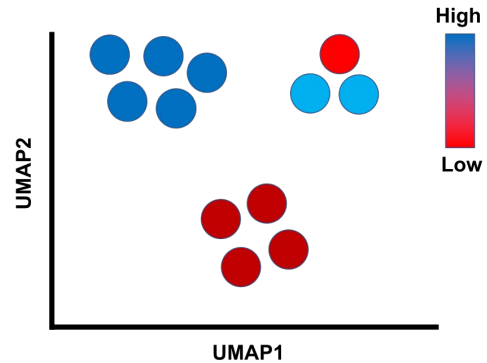


Regress Cell Cycle Effect

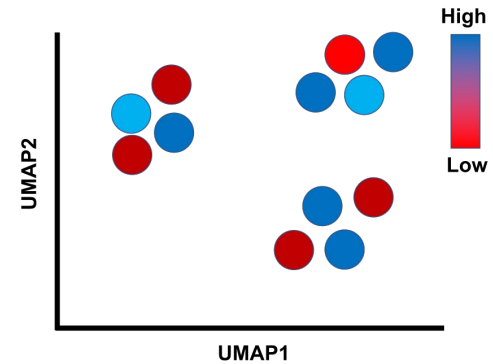


Gene Effect

Gene Expression



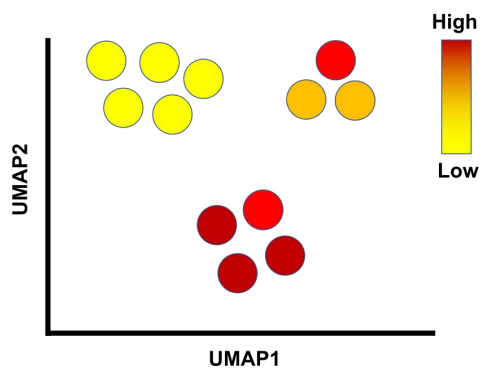
Regress Gene Effect



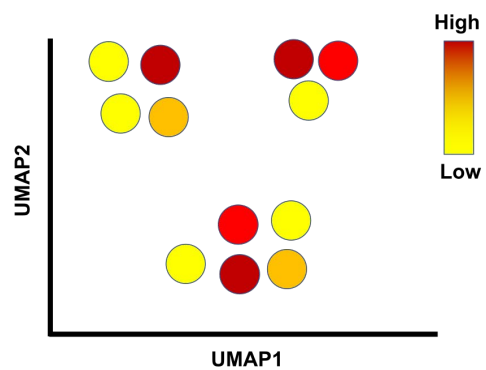
Gene Pathway Effect

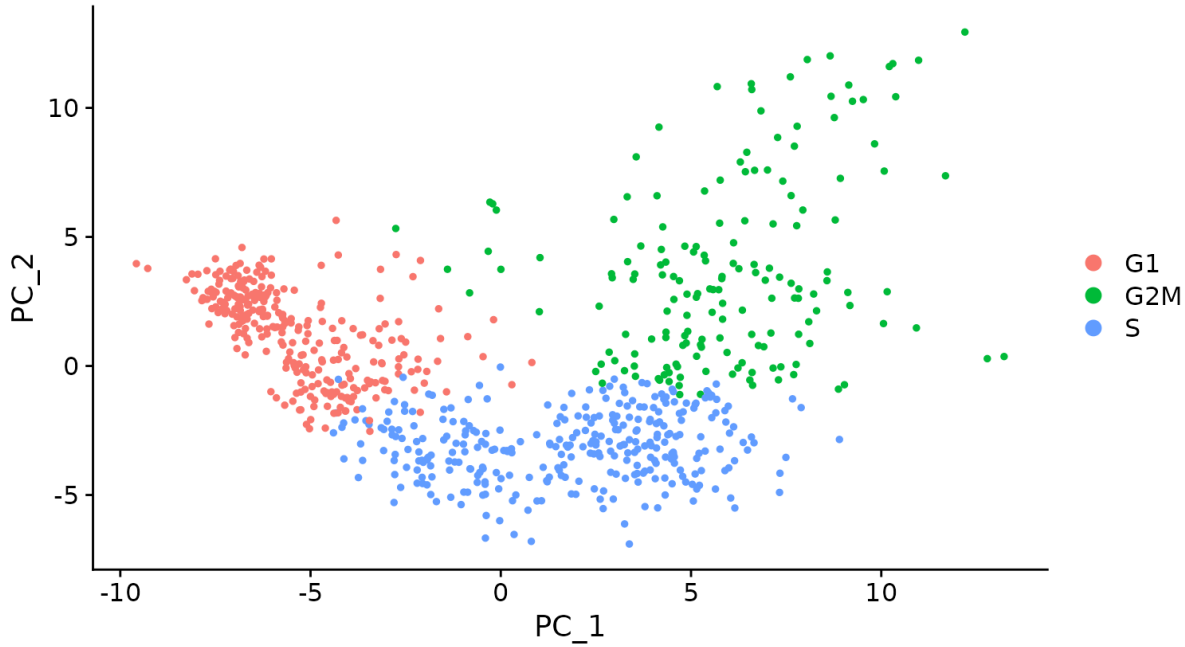
MSigDB Molecular Signatures Database

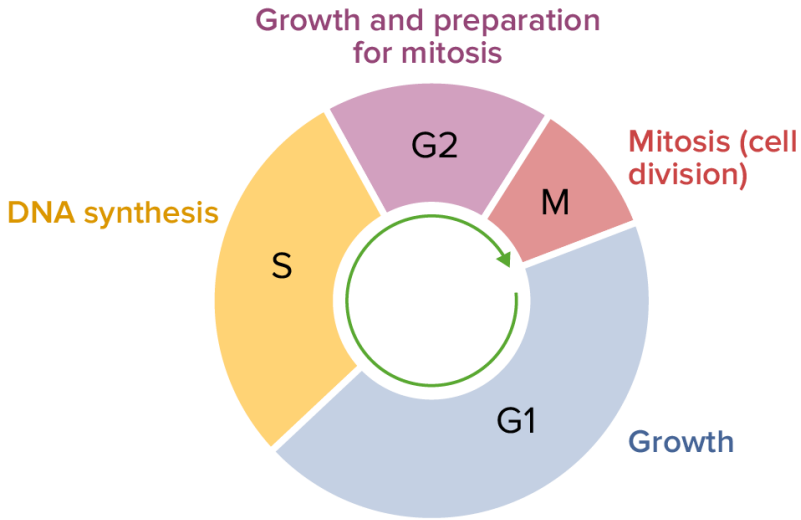
Gene Pathway Expression

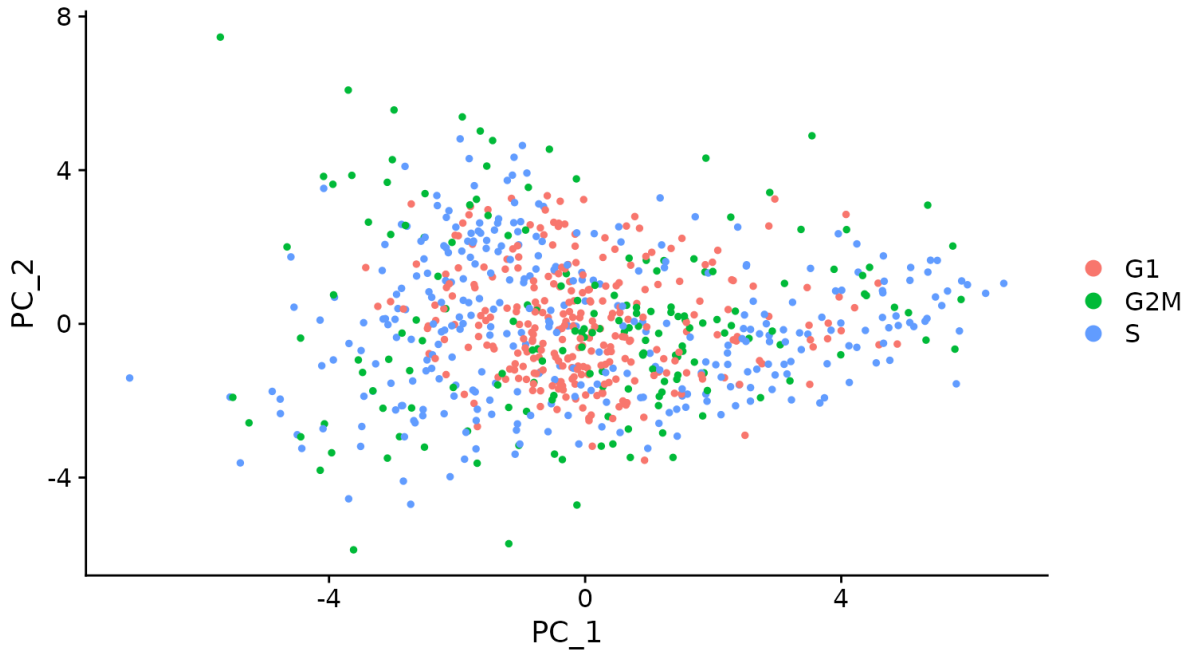


Regress Gene Pathway Effect



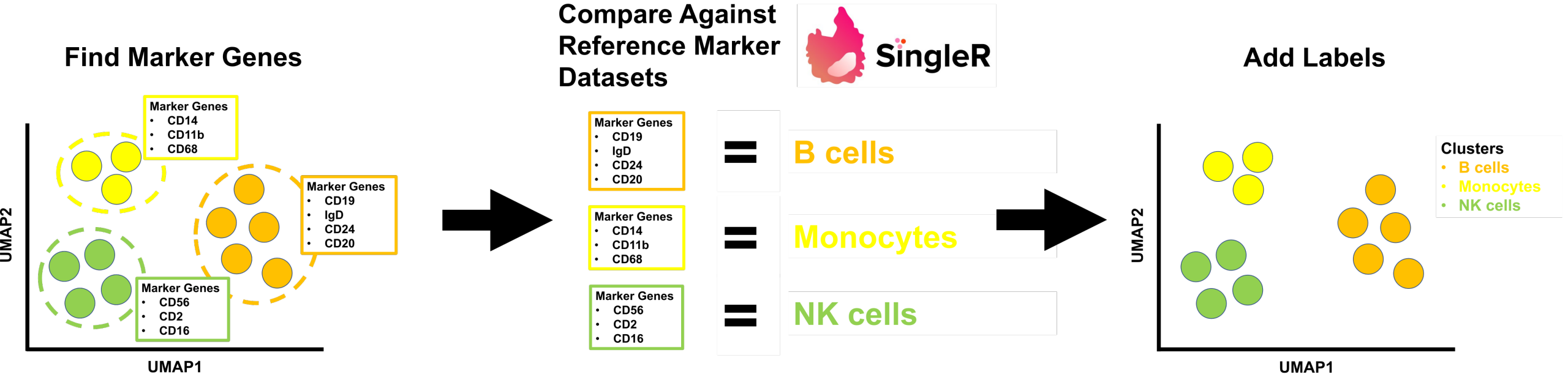

PC_2
PC_1
10
5
0
-5
-10
-5
0
5
10
G1
G2M
S

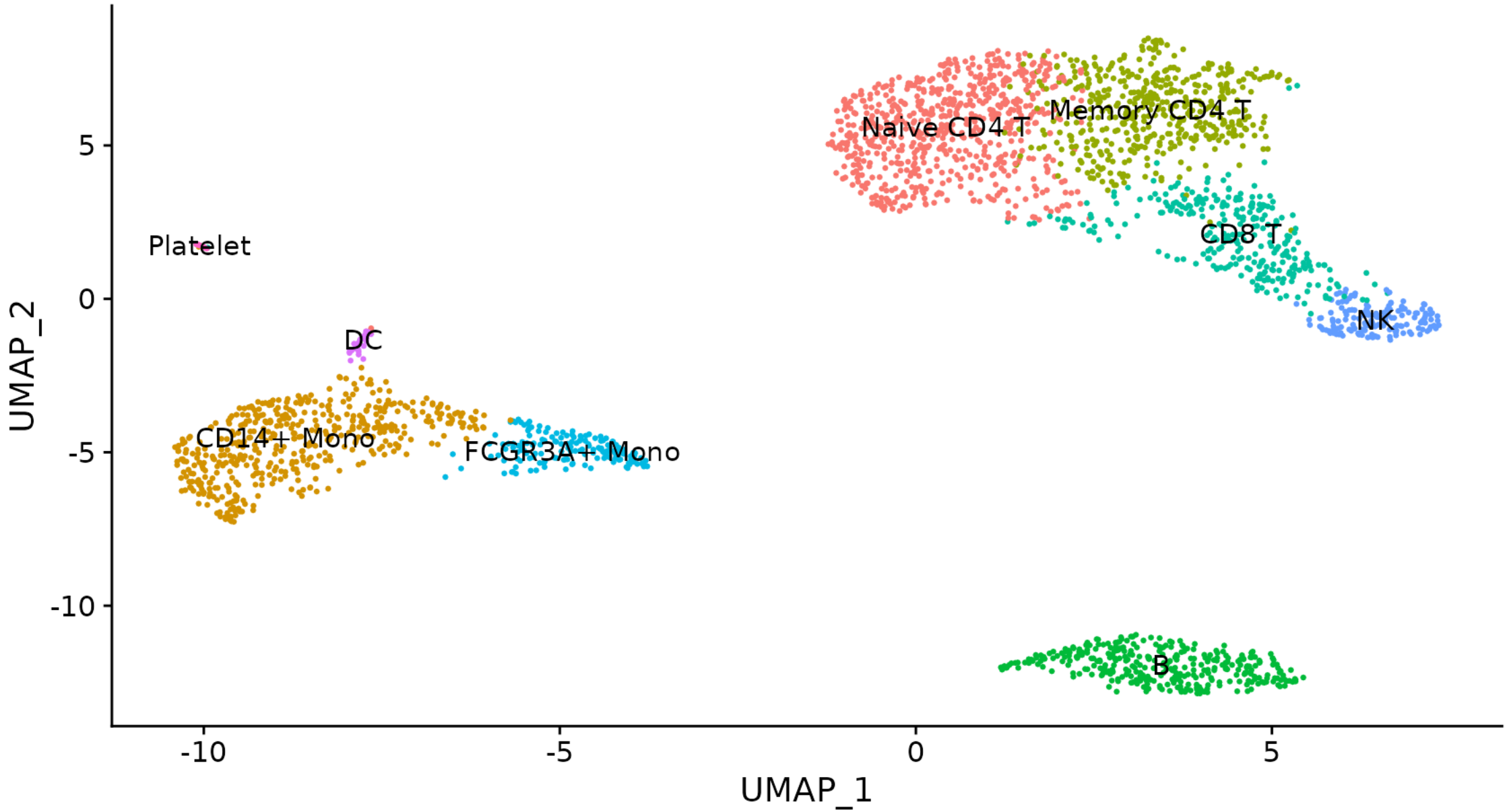

Growth and preparation for mitosis
G2
Mitosis (cell division)
M
DNA synthesis
S
G1
Growth


PC_2
PC_1
8
4
0
-4
-4
0
4
G1
G2M
S

Step 7: Labelling Clusters

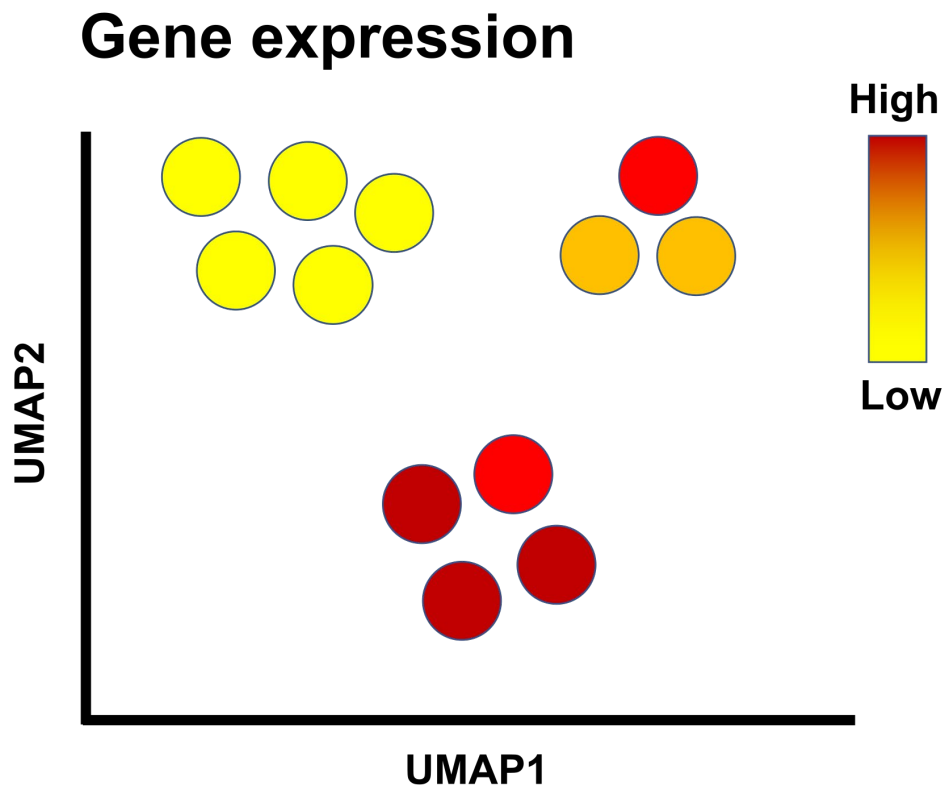
LABELLING WORKFLOW




Naive CD4 T
Memory CD4 T
CD8 T
NK
Platelet
DC
CD14+ Mono
FCGR3A+ Mono
B
UMAP_2
UMAP_1
5
0
-5
-10
-10
-5
0
5

Step 8: Gene Expression

GENE EXPRESSION VISUALISATION



Visualise Gene Pathways From:

MSigDB Molecular Signatures Database

WikiPathways Pathways for the People

GeneOntology Unifying Biology

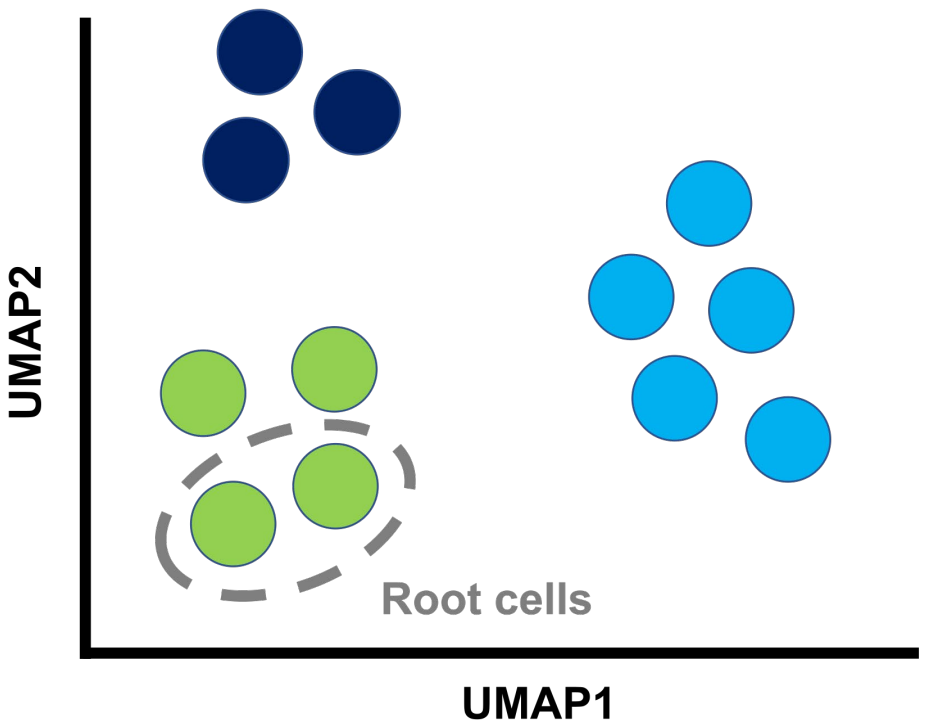
KEGG Kyoto Encyclopedia of Genes and Genomes

reactome

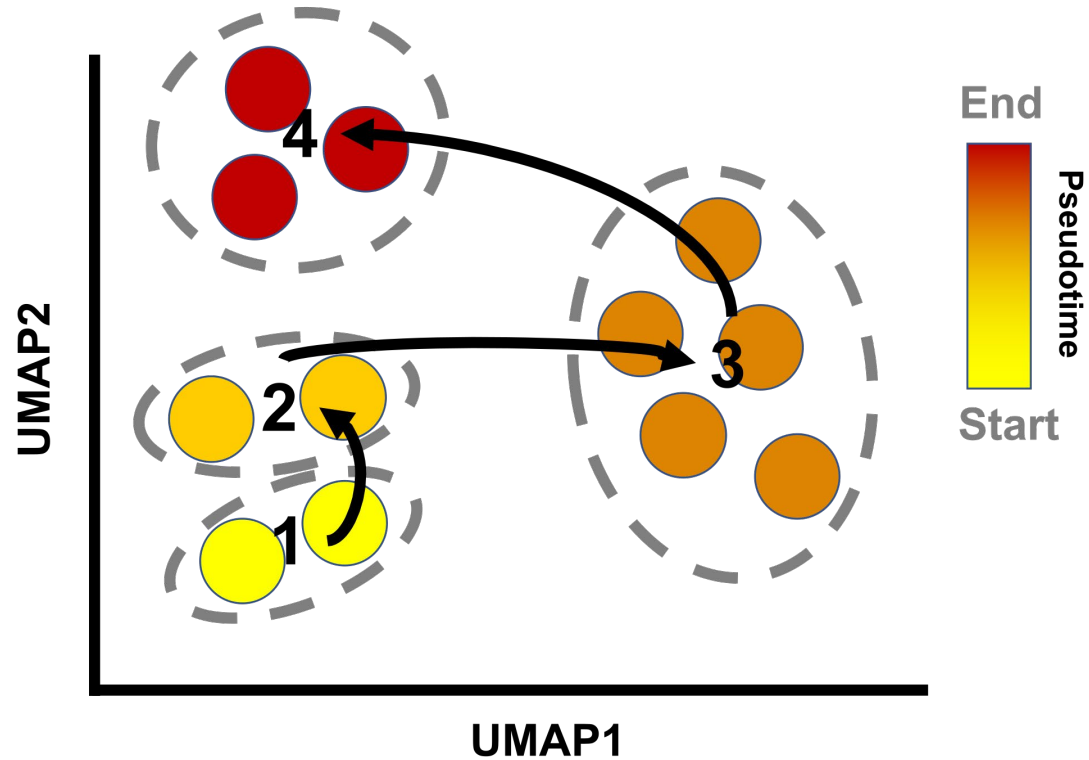
Step 9: Trajectory Analysis

TRAJECTORY ANALYSIS WORKFLOW

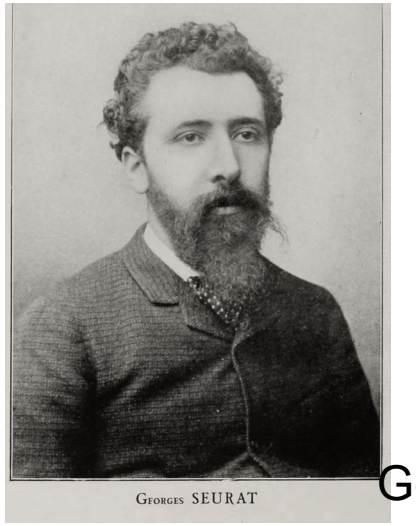
Assign Root Cells



Trajectory Analysis

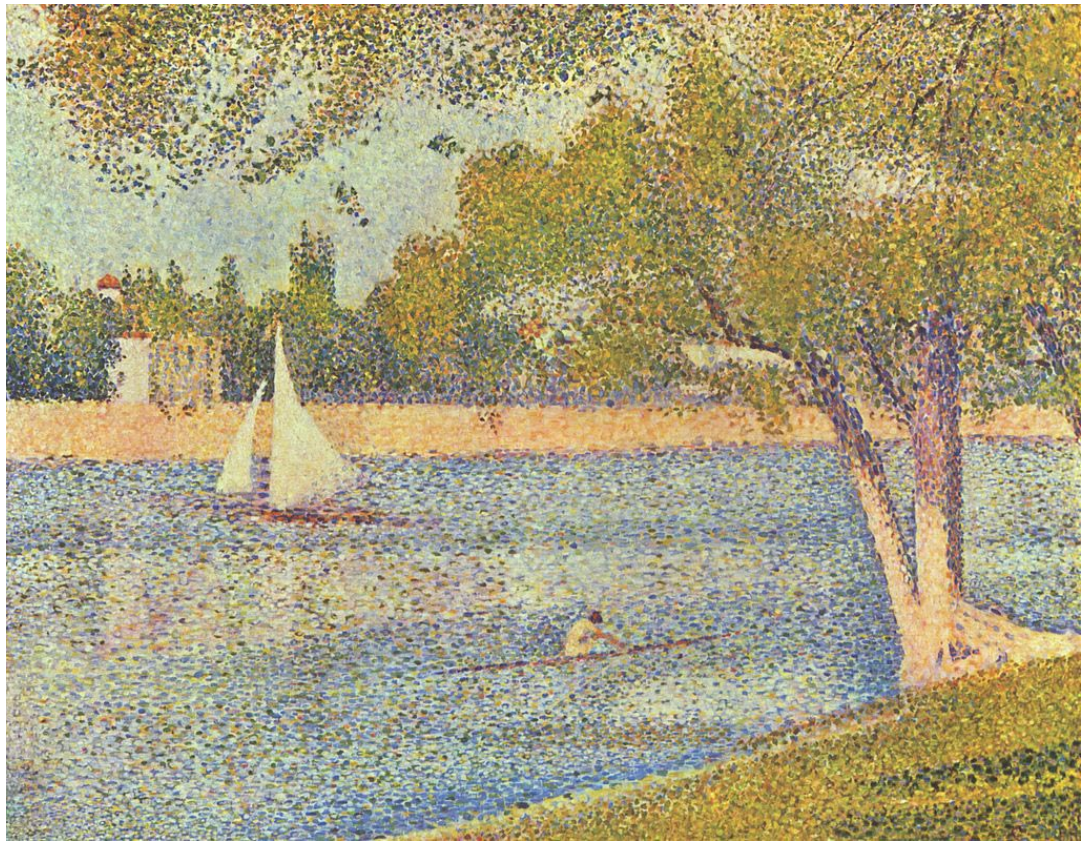


Pointillism and Divisionism 點彩畫派



Georges Pierre Seurat (1859 – 1891)

A Sunday Afternoon on the Island of La Grande Jatte (1884–1886)



The Seine and la Grande Jatte (Springtime 1888)

Seurat v5

We are excited to release Seurat v5! To install, please follow the instructions in our install page. This update brings the following new features and functionality:

- **Integrative multimodal analysis:** The cellular transcriptome is just one aspect of cellular identity, and recent technologies enable routine profiling of chromatin accessibility, histone modifications, and protein levels from single cells. In Seurat v5, we introduce 'bridge integration', a statistical method to integrate experiments measuring different modalities (i.e. separate scRNA-seq and scATAC-seq datasets), using a separate multiomic dataset as a molecular 'bridge'. For example, we demonstrate how to map scATAC-seq datasets onto scRNA-seq datasets, to assist users in interpreting and annotating data from new modalities.

  We recognize that while the goal of matching shared cell types across datasets may be important for many problems, users may also be concerned about which method to use, or that integration could result in a loss of biological resolution. In Seurat v5, we also introduce flexible and streamlined workflows for the integration of multiple scRNA-seq datasets. This makes it easier to explore the results of different integration methods, and to compare these results to a workflow that excludes integration steps.

  - Paper: Dictionary learning for integrative, multimodal, and scalable single-cell analysis
  - Vignette: Streamlined integration of scRNA-seq data
  - Vignette: Cross-modality bridge integration
  - Website: Azimuth-ATAC, reference-mapping for scATAC-seq datasets

Links

View on CRAN

Browse source code

Report a bug

License

Full license

MIT + file LICENSE

Community

Code of conduct

Citation

Citing Seurat

Developers

Rahul Satija
Author, maintainer

Satija Lab and Collaborators
Funder

More about authors...

NK cell

T cell

Monocyte

B cell