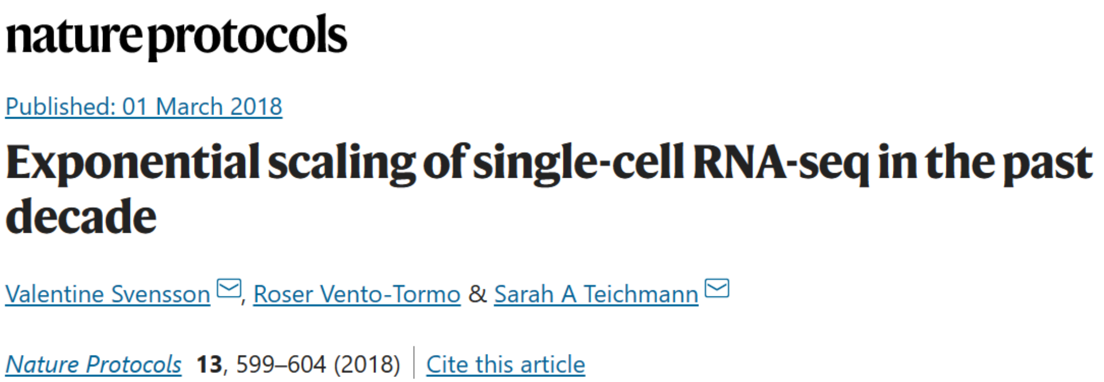
Utilization of single-cell RNA-sequencing data in the public domain for cancer research
公共資料庫中單細胞RNA定序資料於癌症研究之應用簡介

國家衛生研究院 癌症研究所
江士昇 副研究員
蔡芳榆研究助理



nature protocols

Published: 01 March 2018

Exponential scaling of single-cell RNA-seq in the past decade

Valentine Svensson, Roser Vento-Tormo & Sarah A Teichmann

Nature Protocols 13, 599–604 (2018) | Cite this article

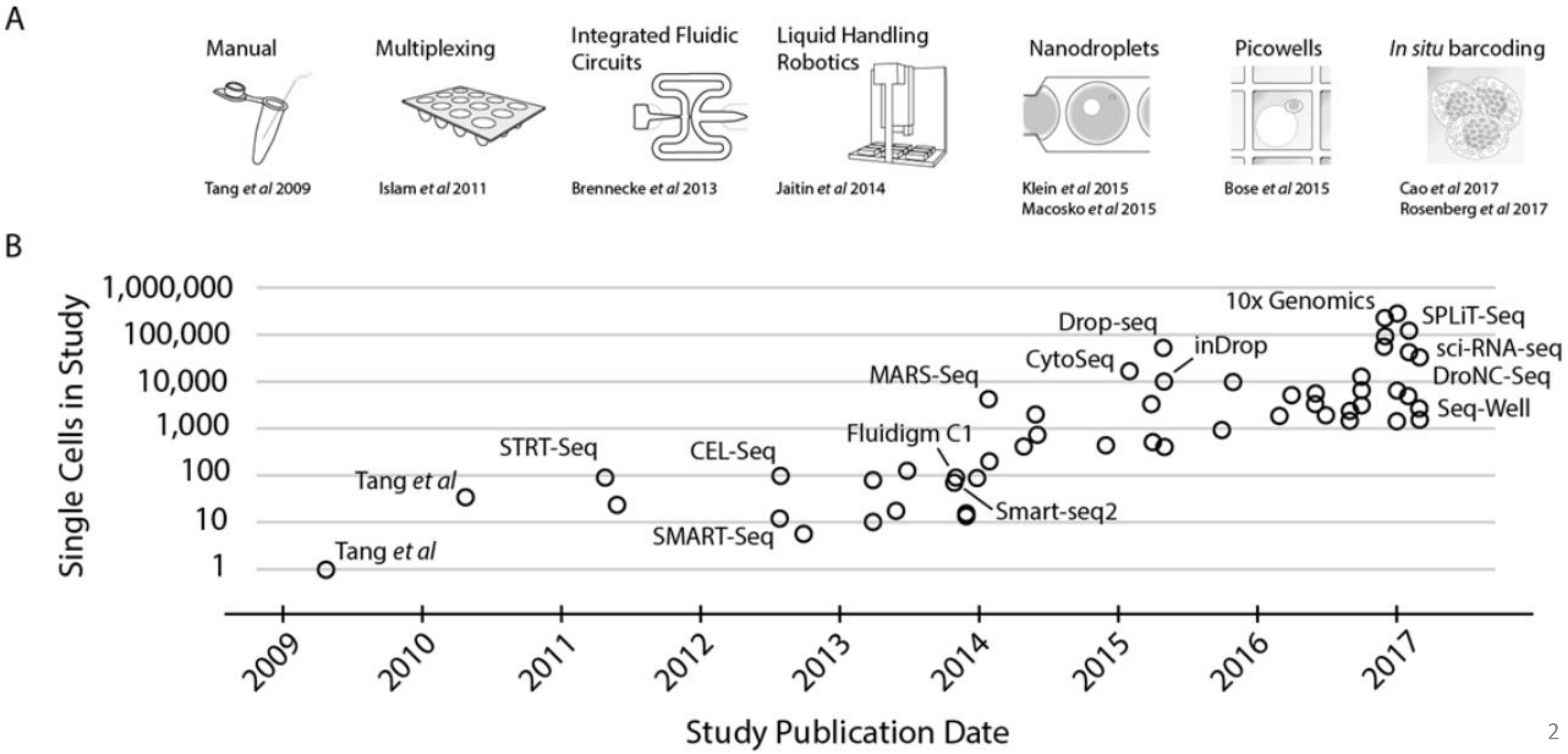
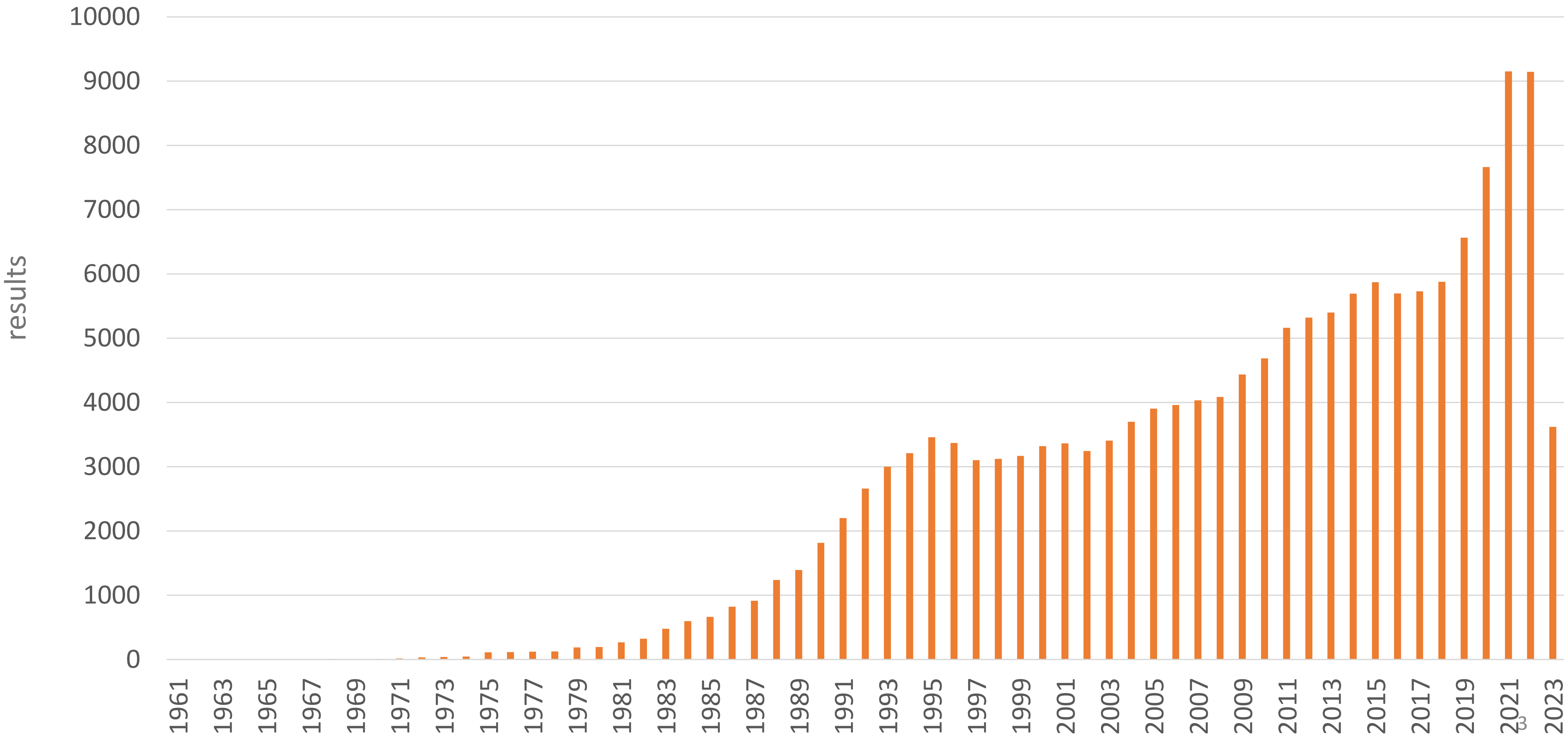
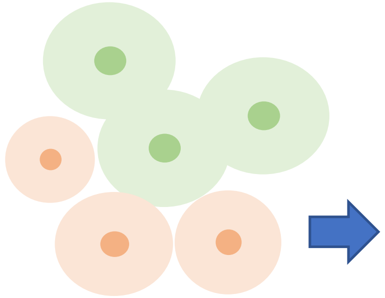


Fig1

PubMed:
Search query: single cell sequencing





	cell1	cell2	cell3	
geneA	0	0	0	
geneB				
geneC				

	cell1	cell2	cell3	
geneA	0	0	0	…
geneB	1	0	1	…
geneC	0	1	12	…

Gene Expression Omnibus
GEO(GSExxxx)

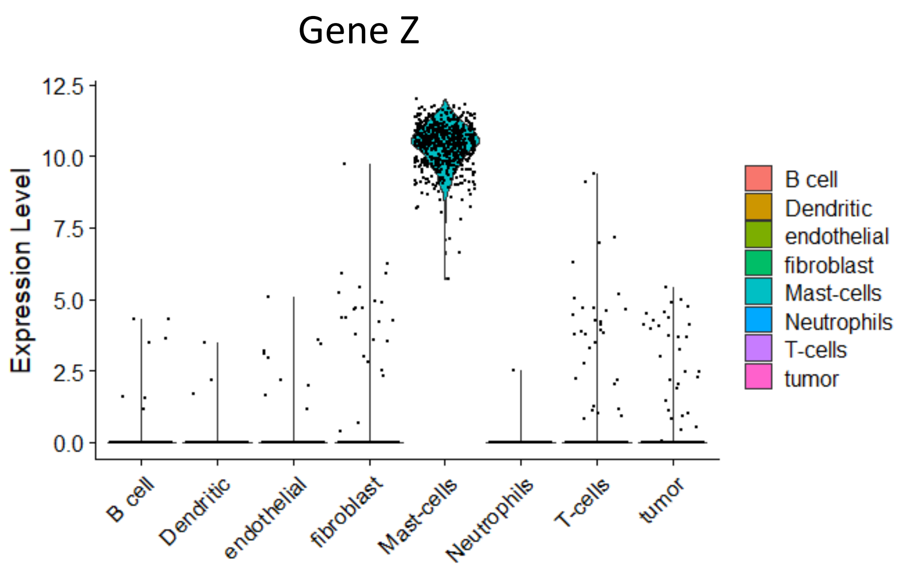
ArrayExpress
(E-MTABxxxx)

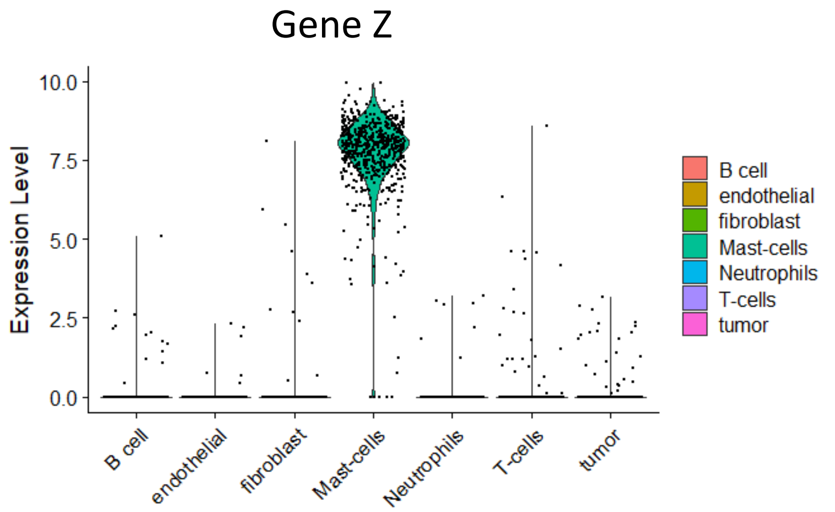
Other

自己已經有single-cell
RNA-sequencing data

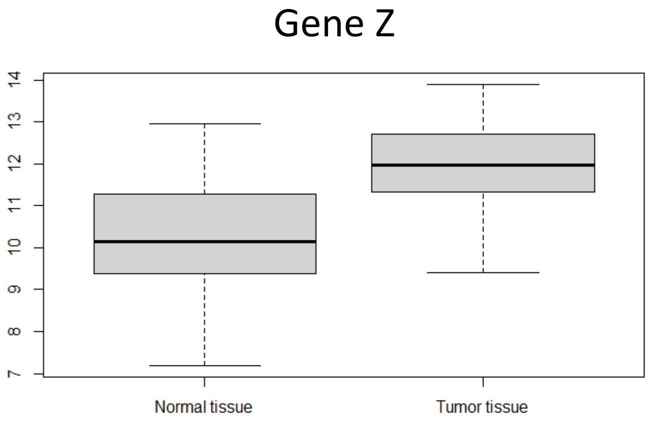
自己還沒有single-cell
RNA-sequencing data

自己已經有 single-cell RNA-sequencing data





自己還沒有 single-cell RNA-sequencing data

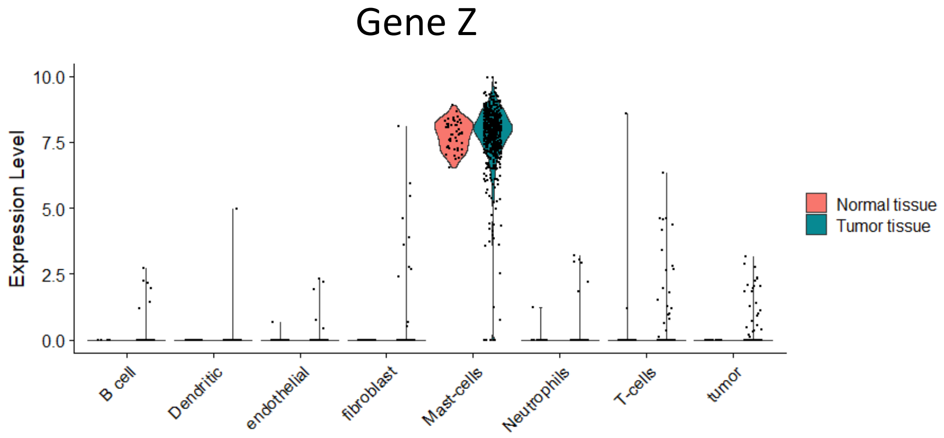


Bulk-tissue

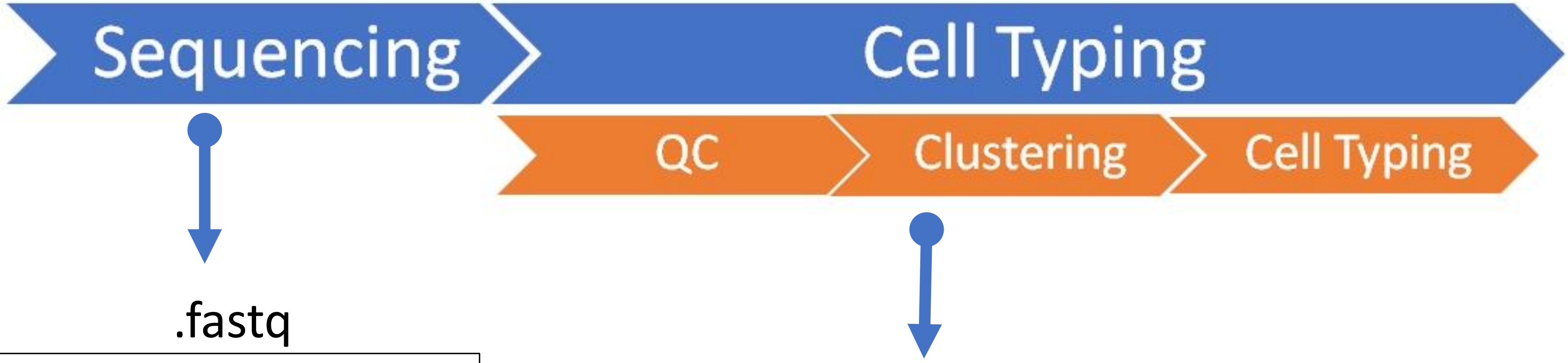
Reference

Gene Z

Mast cell



Single cell RNA sequence data



Sequencing

Cell Typing

QC

Clustering

Cell Typing

.fastq

Ex 1:
sample1_S16_L001_R1_001.fastq.gz
sample1_S16_L001_R2_001.fastq.gz

Gene-cell matrix, Cell type, tSNE / UMAP

Ex 2:
pt1.barcodes.txt.gz
pt1.genes.txt.gz
pt1. matrix.mtx.gz

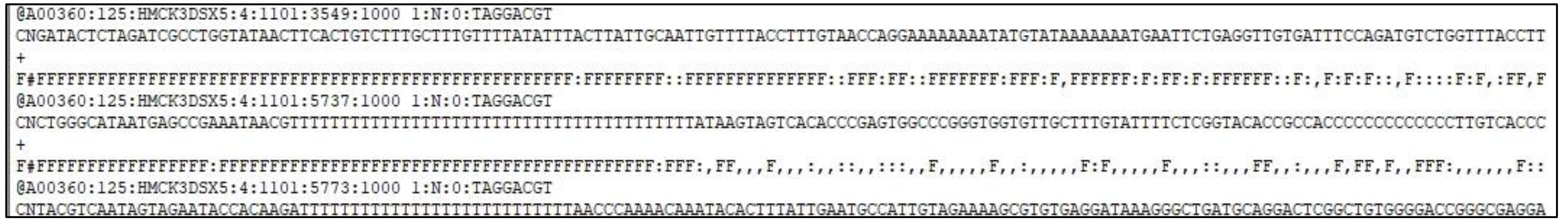
pt2.barcodes.txt.gz
pt2.genes.txt.gz
pt2.matrix.mtx.gz

Ex 3:
GSExxxx_normalized_counts.txt

Ex 4:
GSExxxx_normalized_log2TPM_matrix.rds
GSExxxx_cell_annotation.txt.gz

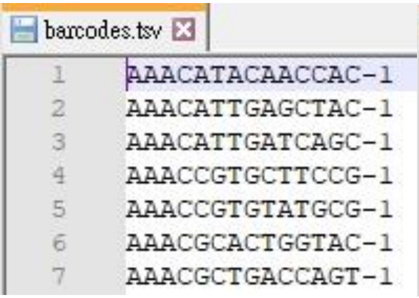
Ex 5:
ALL_cells.Rdata / ALL_cells.rda /ALL_cells.rds

Ex 1:
sample1_S16_L001_R1_001.fastq.gz
sample1_S16_L001_R2_001.fastq.gz

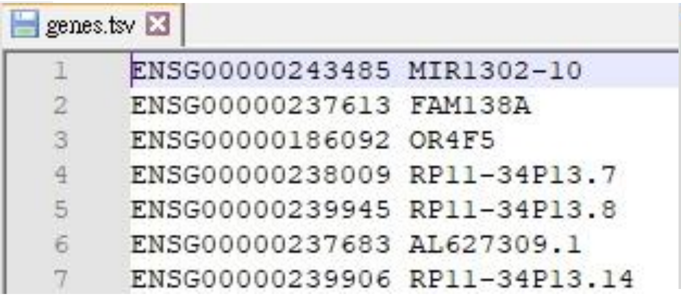


```
@A00360:125:HMCK3DSX5:4:1101:3549:1000 1:N:0:TAGGACGT
CNGATACTCTAGATCGCCTGGTATAACTTCACTGTCTTTGCTTTGTTTTATATTTACTTATTGCAATTGTTTTACCTTTGTAACCAGGAAAAAAAAATATGTATAAAAAAAATGAATTCTGAGGTTGTGATTTCCAGATGTCTGGTTTACCTT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFF::FFFFFFFFFFFFFF::FFF:FF::FFFFFFF:FFF:F,FFFFFF:F:FF:F:FFFFFF::F:,F:F:F::,F::::F:F,:FF,F
@A00360:125:HMCK3DSX5:4:1101:5737:1000 1:N:0:TAGGACGT
CNCTGGGCATAATGAGCCGAAATAACGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTATAAGTAGTCACACCCGAGTGGCCCGGGTGGTGTTGCTTTGTATTTTCTCGGTACACCGCCACCCCCCCCCCCCCTTGTCACCC
+
F#FFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFF:,FF,,,F,,,:,,::,,:::,,F,,,,,F,,:,,,,,F:F,,,,F,,,::,,,FF,,:,,,F,FF,F,,FFF:,,,,,,F::
@A00360:125:HMCK3DSX5:4:1101:5773:1000 1:N:0:TAGGACGT
CNTACGTCAATAGTAGAATACCACAAGATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAACCCAAAACAAATACACTTTATTGAATGCCATTGTAGAAAAGCGTGTGAGGATAAAGGGCTGATGCAGGACTCGGCTGTGGGGACCGGGCGAGGA
```

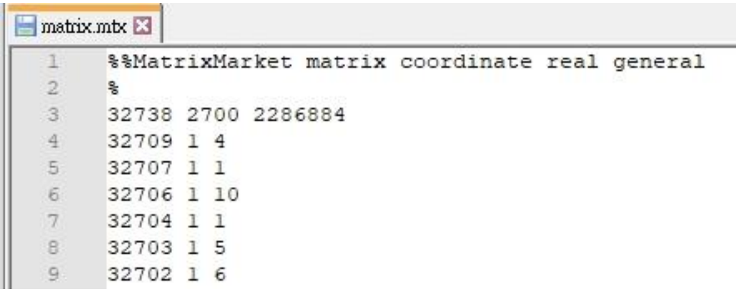
Ex 2:
pt1.barcodes.txt.gz
pt1.genes.txt.gz (pt1.feature.txt.gz
pt1. matrix.mtx.gz



```
AAACATACAACCAC-1
AAACATTGAGCTAC-1
AAACATTGATCAGC-1
AAACCGTGCTTCCG-1
AAACCGTGTATGCG-1
AAACGCACTGGTAC-1
AAACGCTGACCAGT-1
```

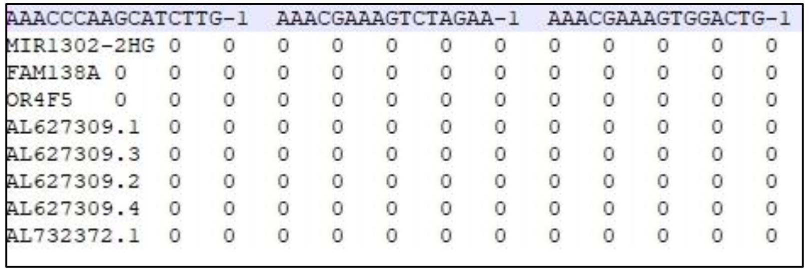


```
ENSG00000243485 MIR1302-10
ENSG00000237613 FAM138A
ENSG00000186092 OR4F5
ENSG00000238009 RP11-34P13.7
ENSG00000239945 RP11-34P13.8
ENSG00000237683 AL627309.1
ENSG00000239906 RP11-34P13.14
```

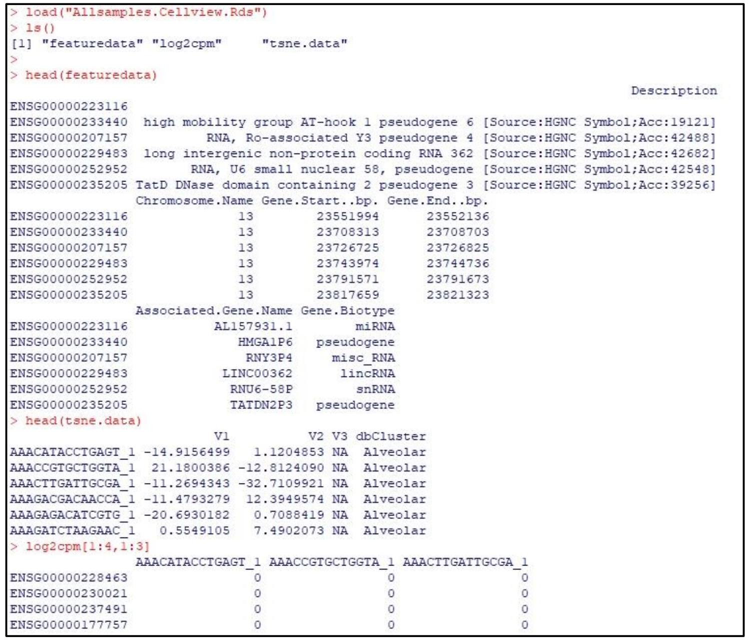


```
%%MatrixMarket matrix coordinate real general
%
32738 2700 2286884
32709 1 4
32707 1 1
32706 1 10
32704 1 1
32703 1 5
32702 1 6
```

Ex 3:
GSExxxx_normalized_counts.txt

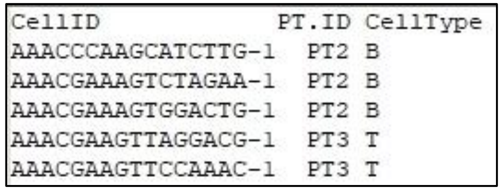


```
AAACCCAAGCATCTTG-1  AAACGAAAGTCTAGAA-1  AAACGAAAGTGGACTG-1
MIR1302-2HG 0   0   0   0   0   0   0   0   0   0   0   0
FAM138A 0   0   0   0   0   0   0   0   0   0   0   0   0
OR4F5   0   0   0   0   0   0   0   0   0   0   0   0   0
AL627309.1  0   0   0   0   0   0   0   0   0   0   0   0
AL627309.3  0   0   0   0   0   0   0   0   0   0   0   0
AL627309.2  0   0   0   0   0   0   0   0   0   0   0   0
AL627309.4  0   0   0   0   0   0   0   0   0   0   0   0
AL732372.1  0   0   0   0   0   0   0   0   0   0   0   0
```



```
> load("Allsamples.Cellview.Rds")
> ls()
[1] "featuredata" "log2cpm"     "tsne.data"
>
> head(featuredata)
                                                                               Description
ENSG00000223116
ENSG00000233440  high mobility group AT-hook 1 pseudogene 6 [Source:HGNC Symbol;Acc:19121]
ENSG00000207157          RNA, Ro-associated Y3 pseudogene 4 [Source:HGNC Symbol;Acc:42488]
ENSG00000229483  long intergenic non-protein coding RNA 362 [Source:HGNC Symbol;Acc:42682]
ENSG00000252952        RNA, U6 small nuclear 58, pseudogene [Source:HGNC Symbol;Acc:42548]
ENSG00000235205 TatD DNase domain containing 2 pseudogene 3 [Source:HGNC Symbol;Acc:39256]
                Chromosome.Name Gene.Start..bp. Gene.End..bp.
ENSG00000223116              13        23551994      23552136
ENSG00000233440              13        23708313      23708703
ENSG00000207157              13        23726725      23726825
ENSG00000229483              13        23743974      23744736
ENSG00000252952              13        23791571      23791673
ENSG00000235205              13        23817659      23821323
                Associated.Gene.Name Gene.Biotype
ENSG00000223116           AL157931.1        miRNA
ENSG00000233440              HMGA1P6   pseudogene
ENSG00000207157               RNY3P4     misc_RNA
ENSG00000229483            LINC00362      lincRNA
ENSG00000252952             RNU6-58P        snRNA
ENSG00000235205             TATDN2P3   pseudogene
> head(tsne.data)
                          V1          V2 V3 dbCluster
AAACATACCTGAGT_1 -14.9156499   1.1204853 NA  Alveolar
AAACCGTGCTGGTA_1  21.1800386 -12.8124090 NA  Alveolar
AAACTTGATTGCGA_1 -11.2694343 -32.7109921 NA  Alveolar
AAAGACGACAACCA_1 -11.4793279  12.3949574 NA  Alveolar
AAAGAGACATCGTG_1 -20.6930182   0.7088419 NA  Alveolar
AAAGATCTAAGAAC_1   0.5549105   7.4902073 NA  Alveolar
> log2cpm[1:4,1:3]
                AAACATACCTGAGT_1 AAACCGTGCTGGTA_1 AAACTTGATTGCGA_1
ENSG00000228463                0                0                0
ENSG00000230021                0                0                0
ENSG00000237491                0                0                0
ENSG00000177757                0                0                0
```

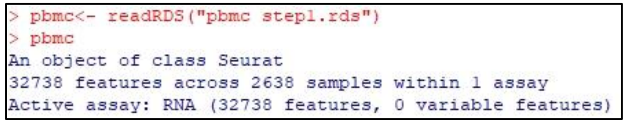
Ex 4:
GSExxxx_normalized_log2TPM_matrix.rds
GSExxxx_cell_annotation.txt.gz



```
CellID             PT.ID CellType
AAACCCAAGCATCTTG-1   PT2 B
AAACGAAAGTCTAGAA-1   PT2 B
AAACGAAAGTGGACTG-1   PT2 B
AAACGAAGTTAGGACG-1   PT3 T
AAACGAAGTTCCAAAC-1   PT3 T
```

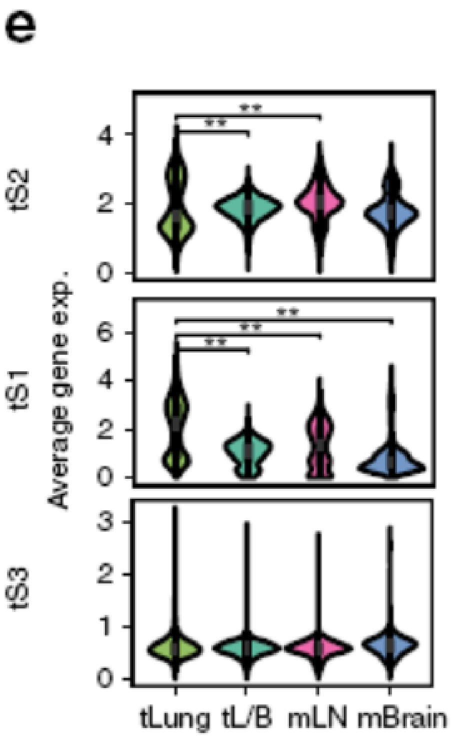
Ex 5:
ALL_cells.Rdata / ALL_cells.rda /ALL_cells.rds

?

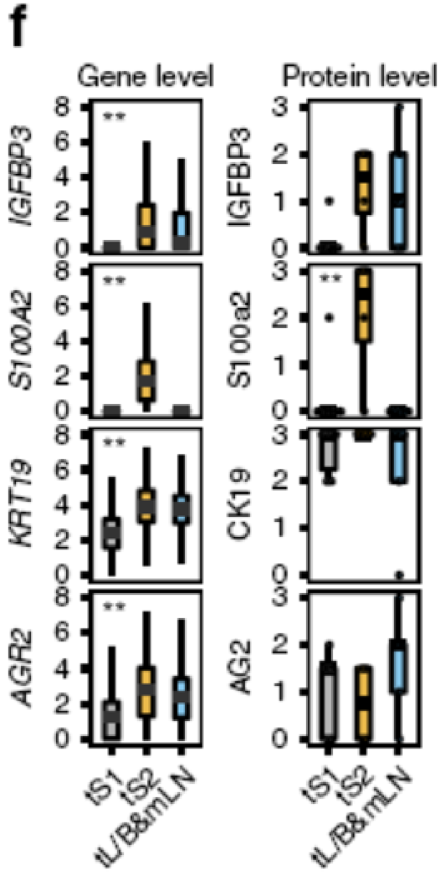


```
> pbmc<- readRDS("pbmc step1.rds")
> pbmc
An object of class Seurat
32738 features across 2638 samples within 1 assay
Active assay: RNA (32738 features, 0 variable features)
```

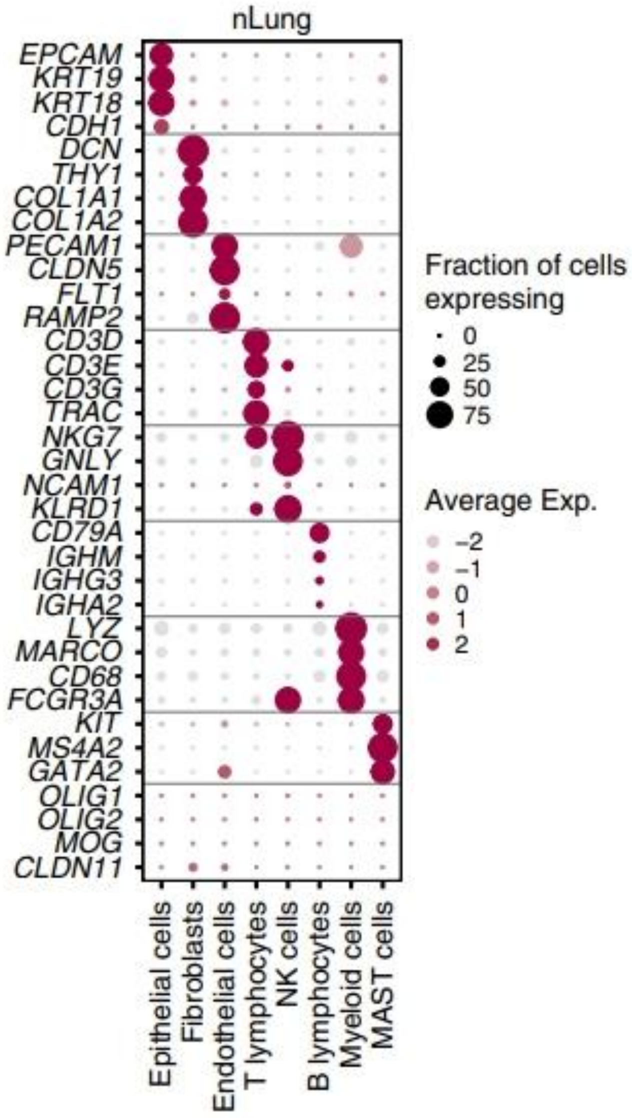
.fastq	X
Gene-cell matrix	X
Gene-cell matrix + Cell type	O



Violin plot



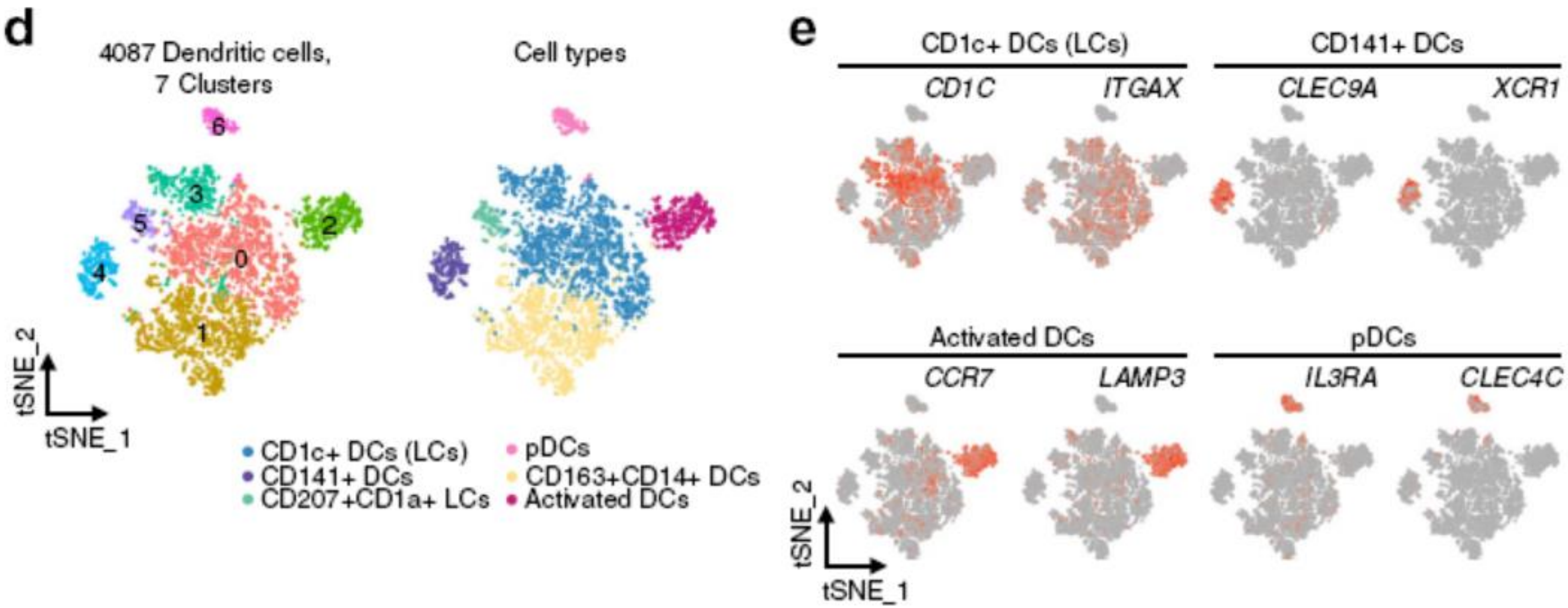
Box plot



Dot plot

(Nat Commun. 2020, Fig. 1d, Fig. 2e, Fig. 2f)

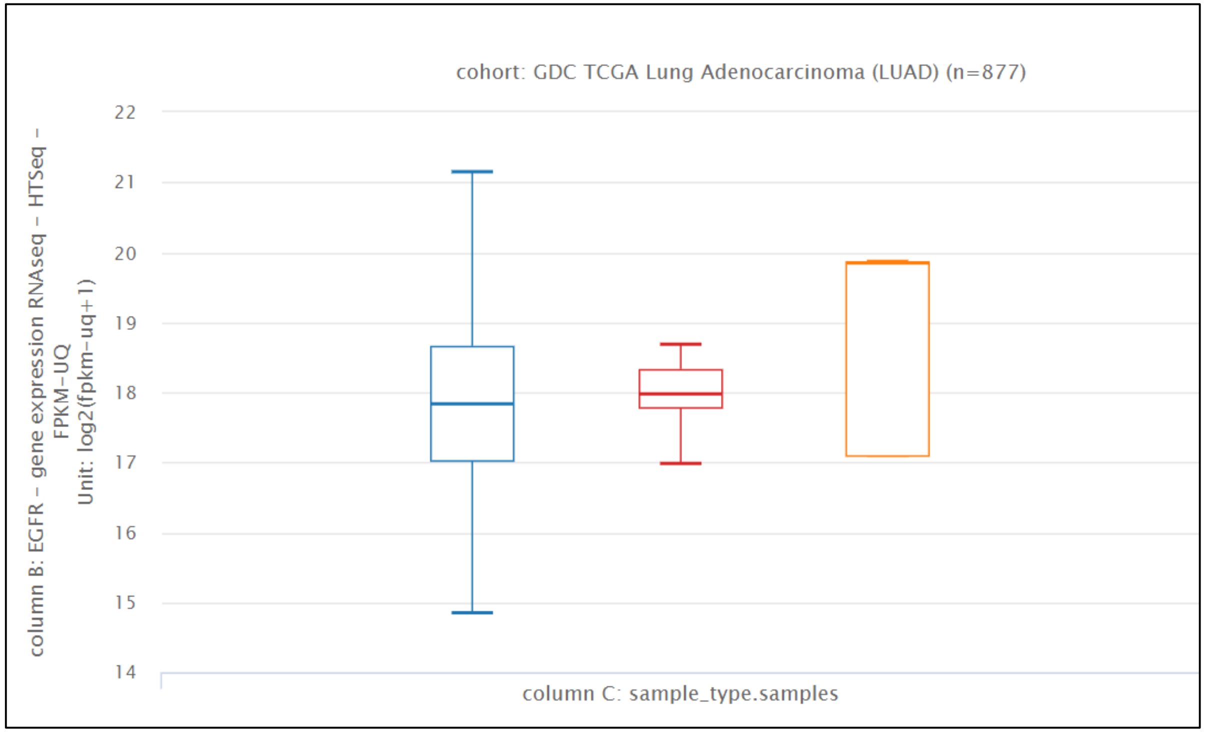
.fastq	X
Gene-cell matrix	X
Gene-cell matrix + Cell type	X
Gene-cell matrix + Cell type+ tSNE / UMAP information	O



(Nat Commun. 2020, Fig. 2)

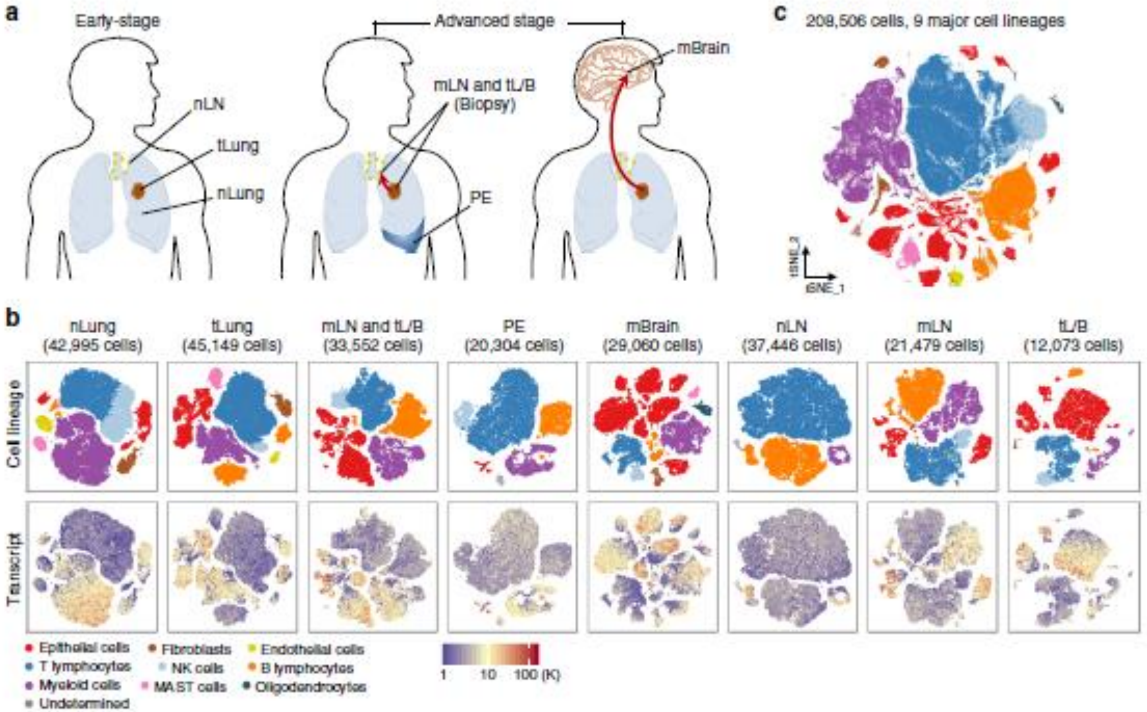
Feature plot

TCGA LUAD (Xena)

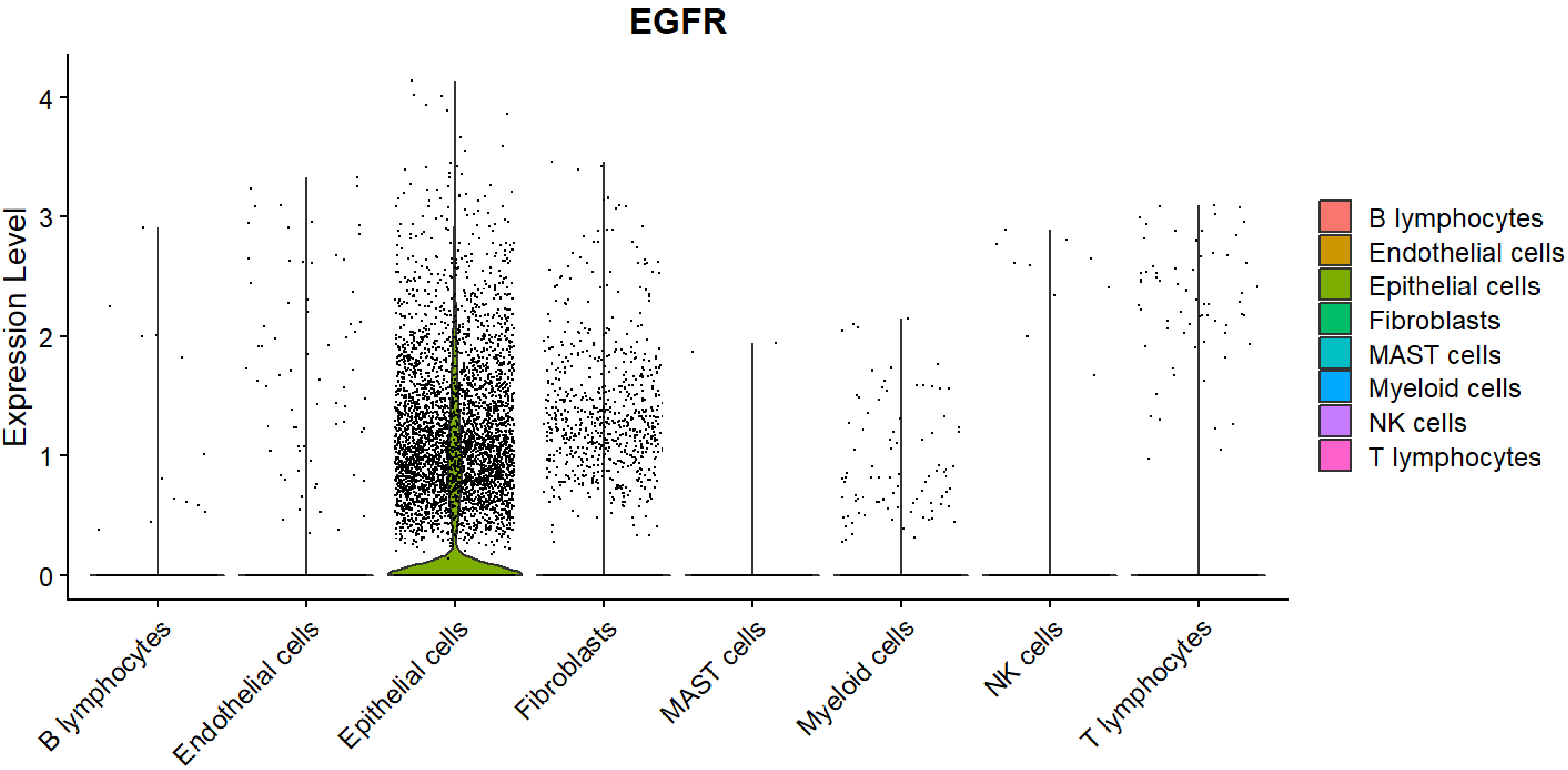


column C: sample_type.samples

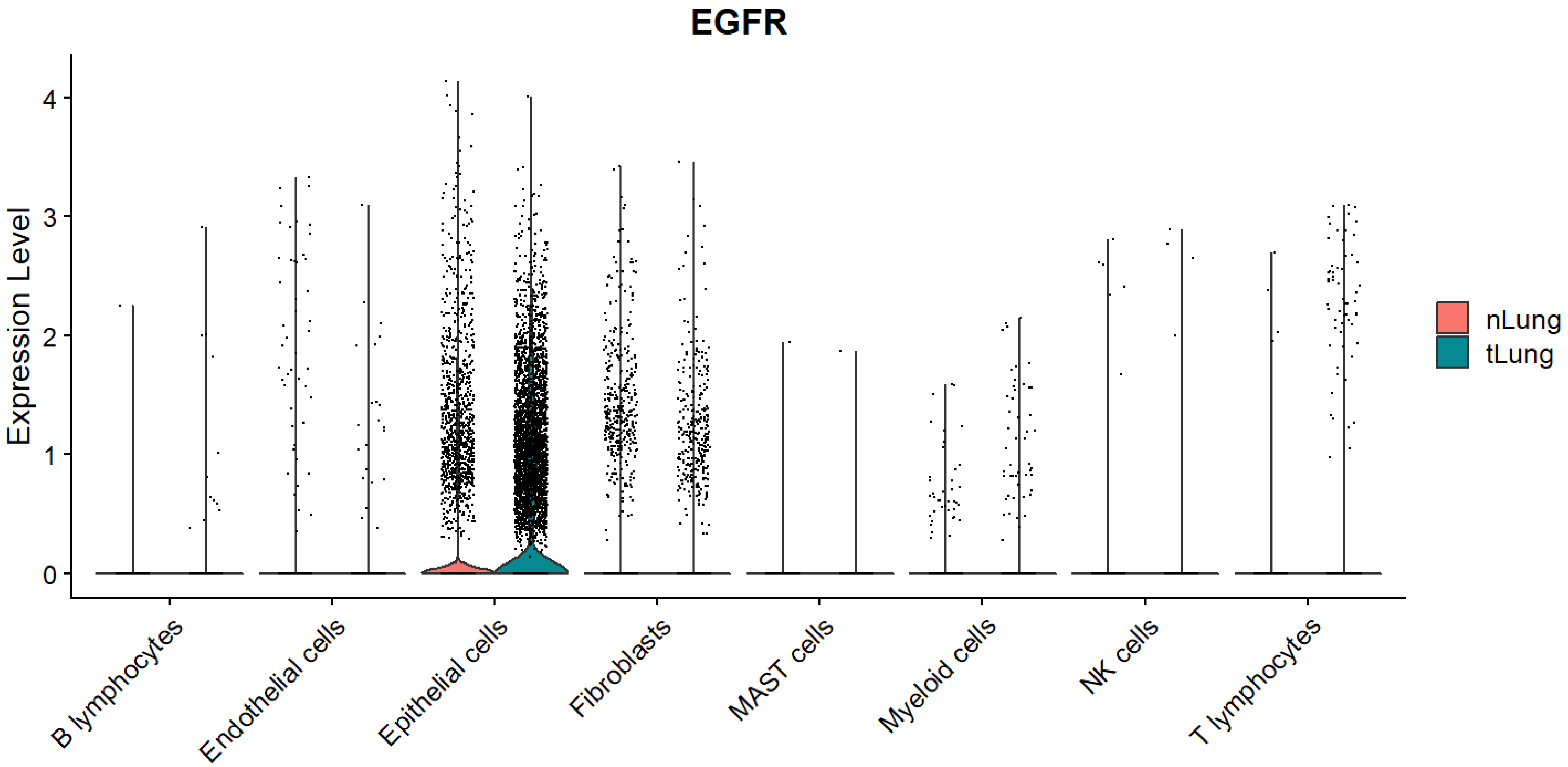
- ● Primary Tumor
- ● Solid Tissue Normal
- ● Recurrent Tumor



(Nat Commun. 2020, Fig. 1,   GSE131907)



GSE131907, tLung+nLung



GSE131907, tLung+nLung

Reference

1. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11:2285. doi: 10.1038/s41467-020-16164-1. (GSE131907)