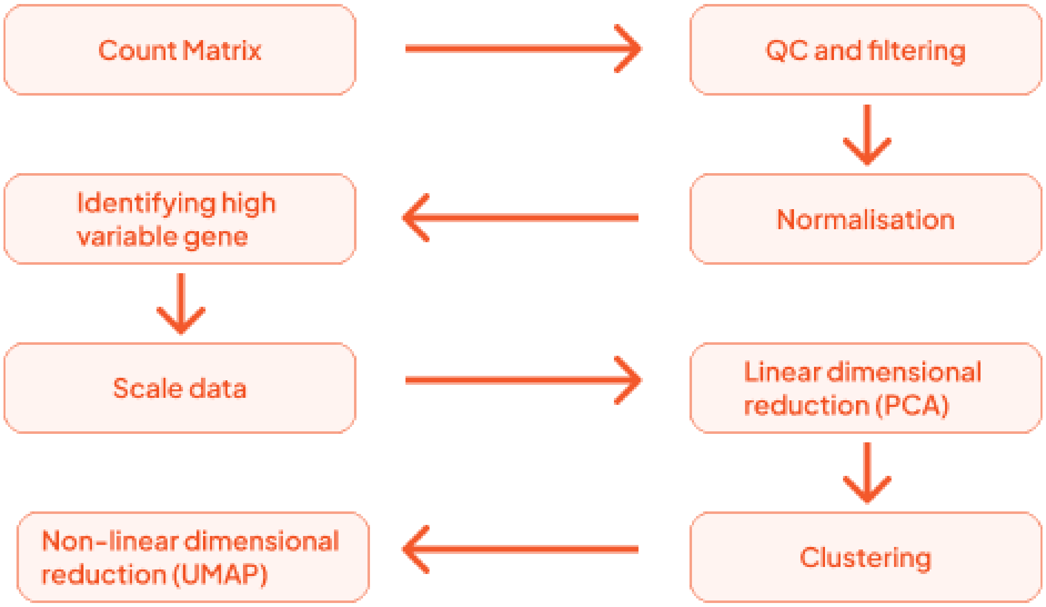
ANALYSIS OF SINGLE-CELL RNA-SEQ DATA FOLLOWING DOWNSTREAM ANALYSIS USING R SEURAT (BY SARAWOOT SOMIN)

DOWNSTREAM ANALYSIS:


Count Matrix
QC and filtering
Normalisation
Identifying high variable gene
Scale data
Linear dimensional reduction (PCA)
Clustering
Non-linear dimensional reduction (UMAP)

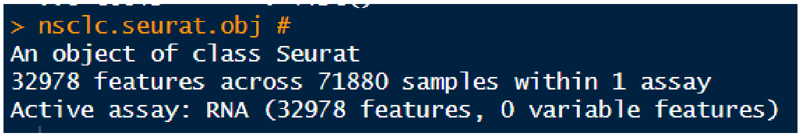
I want to analyse single-cell RNA-seq, the gene expression of lung cancer for the future research.

COUNT MATRIX:

Download the cell matrix of non-small cell lung cancer (NSCLC) dissociated tumor cells from 7 donors (HDF5 file format) from 10X Genomics website (https://www.10xgenomics.com/), description as follows.

Gene Expression and CellPlex libraries were generated from ~33,000 cells as described in the Chromium Single Cell 3' Reagent Kits User Guide (v3.1 Chemistry Dual Index) with Feature Barcode technology for Cell Surface Protein and Cell Multiplexing (CG000390 Rev B) using the Chromium X and sequenced on an Illumina NovaSeq 6000 to a read depth of approximately 70,000 mean reads per cell for Gene Expression and 25,000 mean reads per cell for CellPlex.

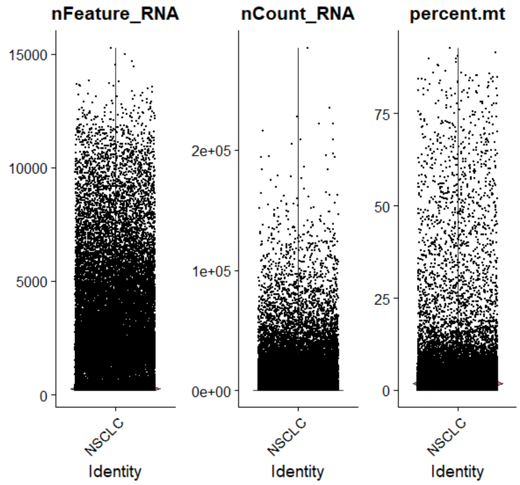
- Convert HDF5 file format to Suerat object.
- There are multiple modalities in Suerat object, we use gene expression only.
- Initial the Seurat oject with the raw data (non-normallise data)
- Set minimum cells and features (genes)
- Count the number of feature and cells across the sample for analysing quality of cells and genese.



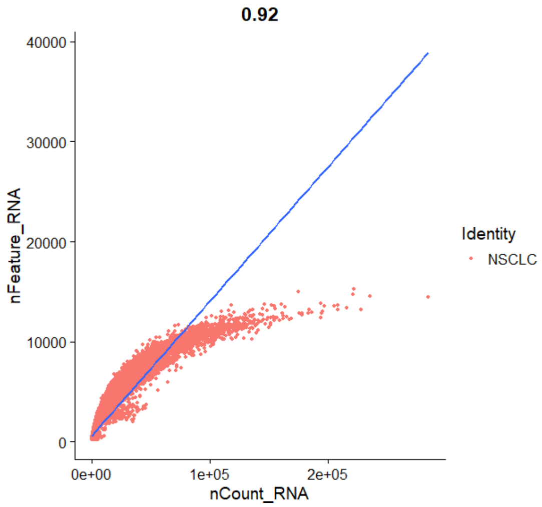
```
> nsclc.seurat.obj #
An object of class Seurat
32978 features across 71880 samples within 1 assay
Active assay: RNA (32978 features, 0 variable features)
```

QC AND FILTERING:

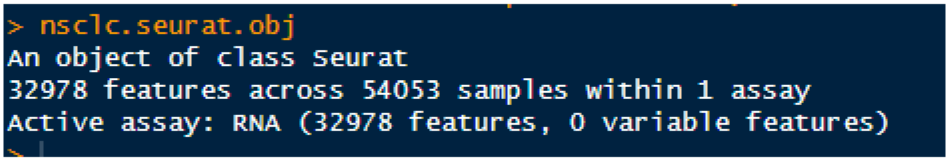
- We calculated percentage of mitochondria for filtering them out later, because the mitochondria molecules won’t be visualised in clustering process.
- The picture below shows that we have a lot of cells having higher number of genes (nFeature_RNA), higher numbers of molecules detected, and also many cells have high percentage of mitochondria (percent.mt)



- I add the straight line to my plot (number of molecule (nCount_RNA) and number of gene (nFeature_RNA)), which quality dataset should follow the straight line.



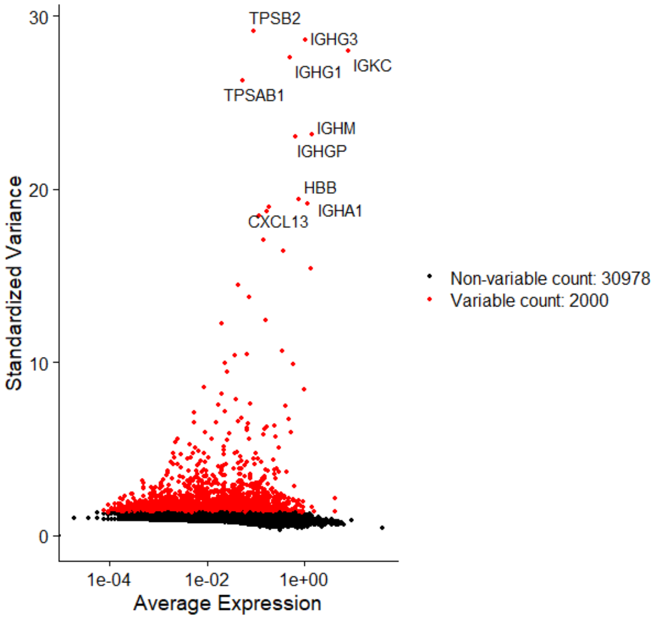
- We filtered the features out 54053



```
> nsclc.seurat.obj
An object of class Seurat
32978 features across 54053 samples within 1 assay
Active assay: RNA (32978 features, 0 variable features)
>
```

IDENTIFYING HIGH VARIABLE GENE:

I identified the 10 most highly variable genes.



SCALE DATA:

- Technical noise might occur, we need to arrange the Seurat object as the coding below:

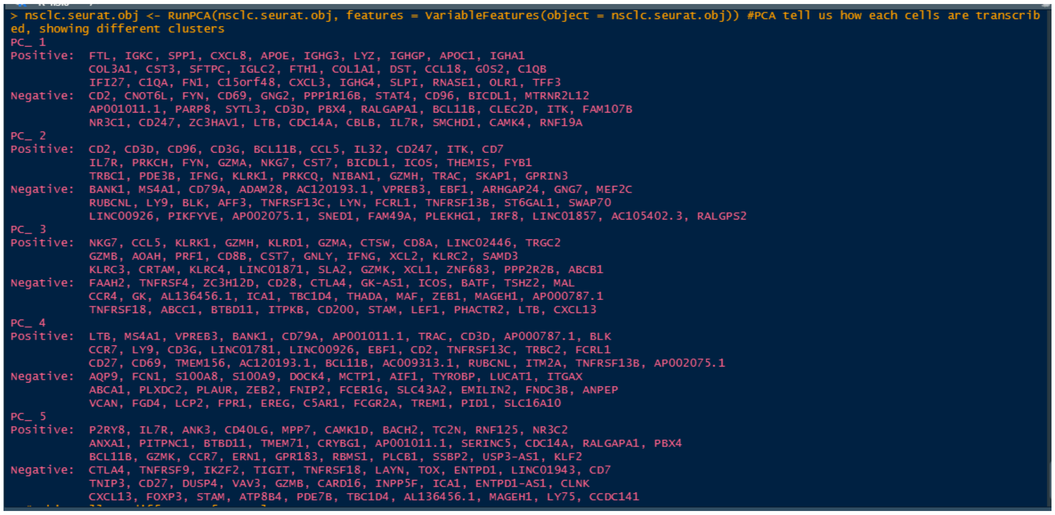
```
all.genes <- rownames(nsclc.seurat.obj)

nsclc.seurat.obj <- ScaleData(nsclc.seurat.obj, features = all.genes)

str(nsclc.seurat.obj)
```

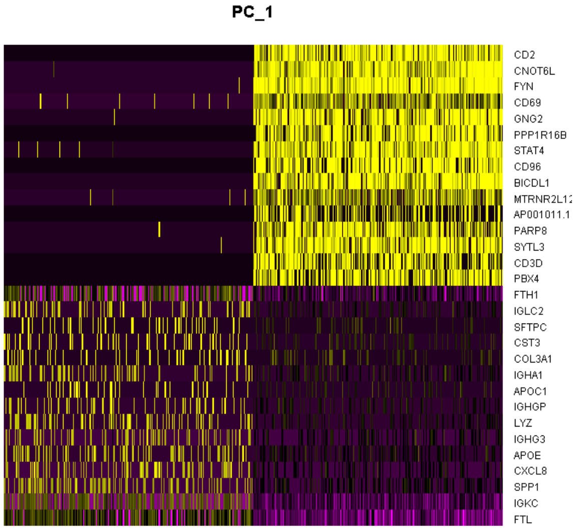
LINEAR DIMENSIONAL REDUCTION (PCA):

- After we scaled the data, we performed linear dimensional reduction.
- We show positive and negative PCA scores.

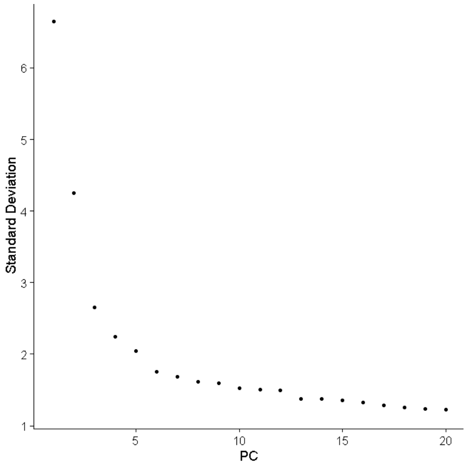


```
> nsclc.seurat.obj <- RunPCA(nsclc.seurat.obj, features = VariableFeatures(object = nsclc.seurat.obj)) #PCA tell us how each cells are transcrib
ed, showing different clusters
PC_ 1
Positive:  FTL, IGKC, SPP1, CXCL8, APOE, IGHG3, LYZ, IGHGP, APOC1, IGHA1
	   COL3A1, CST3, SFTPC, IGLC2, FTH1, COL1A1, DST, CCL18, G0S2, C1QB
	   IFI27, C1QA, FN1, C15orf48, CXCL3, IGHG4, SLPI, RNASE1, OLR1, TFF3
Negative:  CD2, CNOT6L, FYN, CD69, GNG2, PPP1R16B, STAT4, CD96, BICDL1, MTRNR2L12
	   AP001011.1, PARP8, SYTL3, CD3D, PBX4, RALGAPA1, BCL11B, CLEC2D, ITK, FAM107B
	   NR3C1, CD247, ZC3HAV1, LTB, CDC14A, CBLB, IL7R, SMCHD1, CAMK4, RNF19A
PC_ 2
Positive:  CD2, CD3D, CD96, CD3G, BCL11B, CCL5, IL32, CD247, ITK, CD7
	   IL7R, PRKCH, FYN, GZMA, NKG7, CST7, BICDL1, ICOS, THEMIS, FYB1
	   TRBC1, PDE3B, IFNG, KLRK1, PRKCQ, NIBAN1, GZMH, TRAC, SKAP1, GPRIN3
Negative:  BANK1, MS4A1, CD79A, ADAM28, AC120193.1, VPREB3, EBF1, ARHGAP24, GNG7, MEF2C
	   RUBCNL, LY9, BLK, AFF3, TNFRSF13C, LYN, FCRL1, TNFRSF13B, ST6GAL1, SWAP70
	   LINC00926, PIKFYVE, AP002075.1, SNED1, FAM49A, PLEKHG1, IRF8, LINC01857, AC105402.3, RALGPS2
PC_ 3
Positive:  NKG7, CCL5, KLRK1, GZMH, KLRD1, GZMA, CTSW, CD8A, LINC02446, TRGC2
	   GZMB, AOAH, PRF1, CD8B, CST7, GNLY, IFNG, XCL2, KLRC2, SAMD3
	   KLRC3, CRTAM, KLRC4, LINC01871, SLA2, GZMK, XCL1, ZNF683, PPP2R2B, ABCB1
Negative:  FAAH2, TNFRSF4, ZC3H12D, CD28, CTLA4, GK-AS1, ICOS, BATF, TSHZ2, MAL
	   CCR4, GK, AL136456.1, ICA1, TBC1D4, THADA, MAF, ZEB1, MAGEH1, AP000787.1
	   TNFRSF18, ABCC1, BTBD11, ITPKB, CD200, STAM, LEF1, PHACTR2, LTB, CXCL13
PC_ 4
Positive:  LTB, MS4A1, VPREB3, BANK1, CD79A, AP001011.1, TRAC, CD3D, AP000787.1, BLK
	   CCR7, LY9, CD3G, LINC01781, LINC00926, EBF1, CD2, TNFRSF13C, TRBC2, FCRL1
	   CD27, CD69, TMEM156, AC120193.1, BCL11B, AC009313.1, RUBCNL, ITM2A, TNFRSF13B, AP002075.1
Negative:  AQP9, FCN1, S100A8, S100A9, DOCK4, MCTP1, AIF1, TYROBP, LUCAT1, ITGAX
	   ABCA1, PLXDC2, PLAUR, ZEB2, FNIP2, FCER1G, SLC43A2, EMILIN2, FNDC3B, ANPEP
	   VCAN, FGD4, LCP2, FPR1, EREG, C5AR1, FCGR2A, TREM1, PID1, SLC16A10
PC_ 5
Positive:  P2RY8, IL7R, ANK3, CD40LG, MPP7, CAMK1D, BACH2, TC2N, RNF125, NR3C2
	   ANXA1, PITPNC1, BTBD11, TMEM71, CRYBG1, AP001011.1, SERINC5, CDC14A, RALGAPA1, PBX4
	   BCL11B, GZMK, CCR7, ERN1, GPR183, RBMS1, PLCB1, SSBP2, USP3-AS1, KLF2
Negative:  CTLA4, TNFRSF9, IKZF2, TIGIT, TNFRSF18, LAYN, TOX, ENTPD1, LINC01943, CD7
	   TNIP3, CD27, DUSP4, VAV3, GZMB, CARD16, INPP5F, ICA1, ENTPD1-AS1, CLNK
	   CXCL13, FOXP3, STAM, ATP8B4, PDE7B, TBC1D4, AL136456.1, MAGEH1, LY75, CCDC141
```

- We also show heat-map of 500 cells and features (genes), as follows.

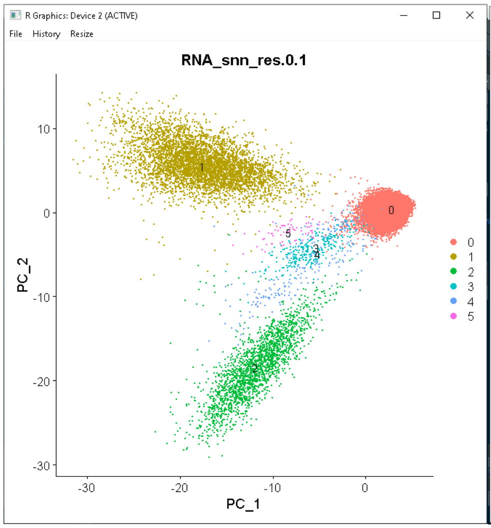


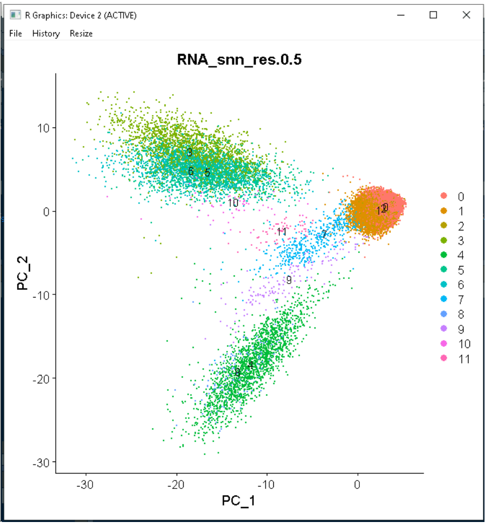
- Standard deviation was also considered for the reduction, as follows.



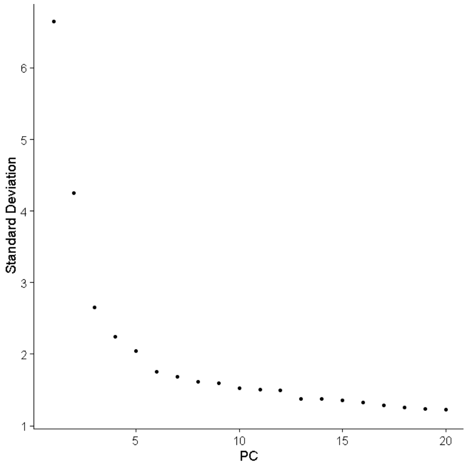
CLUSTERING:

- We tried 0.1 and 0.5 resolution to see which solution is suitable for clustering this data.



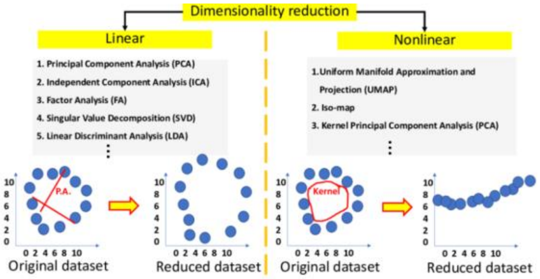


- 0.1 resolution works best because cluster 3, 6 and 5, which are different cells, are grouped together.

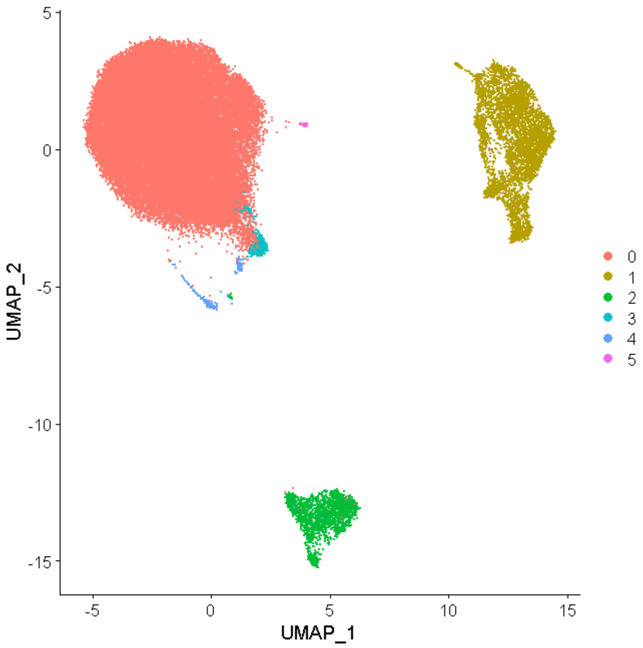


NON-LINEAR DIMENSIONAL REDUCTION (UMAP):

- The difference between linear and non-linear concepts is as follows.


Dimensionality reduction
Linear
Nonlinear
1. Principal Component Analysis (PCA)
2. Independent Component Analysis (ICA)
3. Factor Analysis (FA)
4. Singular Value Decomposition (SVD)
5. Linear Discriminant Analysis (LDA)
1.Uniform Manifold Approximation and Projection (UMAP)
2. Iso-map
3. Kernel Principal Component Analysis (PCA)
P.A.
Kernel
10 8 6 4 2 0
0 2 4 6 8 10
Original dataset
Reduced dataset
Original dataset
Reduced dataset

- We used uniform manifold approximation (UMAP) with 0.1 resolution for lower dimensions space of clustering.


5
0
-5
-10
-15
UMAP_2
-5
0
5
10
15
UMAP_1
● 0
● 1
● 2
● 3
● 4
● 5