

# Seurat tutorial

Anne Brussaard

2025-05-12

Deelvraag: Kan ik door het volgen van een tutorial met Seurat data preprocessing en visualisatie uitvoeren? Om deze deelvraag te beantwoorden zal het volgende flowschema aangehouden worden.

1. De data wordt geladen
2. Er wordt een Seurat object gemaakt
3. De filterstappen worden uitgevoerd
4. Clusters worden visueel gemaakt

De commands worden uitgevoerd in het environment `project_brie2`. `project_brie2` geeft de mogelijkheid om te werken met `seurat-4.4.0`. Dit script wordt vervolgens gemaakt via de terminal waarin R wordt geopend met het command `rmarkdown::render()`

Eerst worden de packages geladen. Seurat voor analyse, dplyr voor filteren en selecteren en patchwork voor maken van plots.

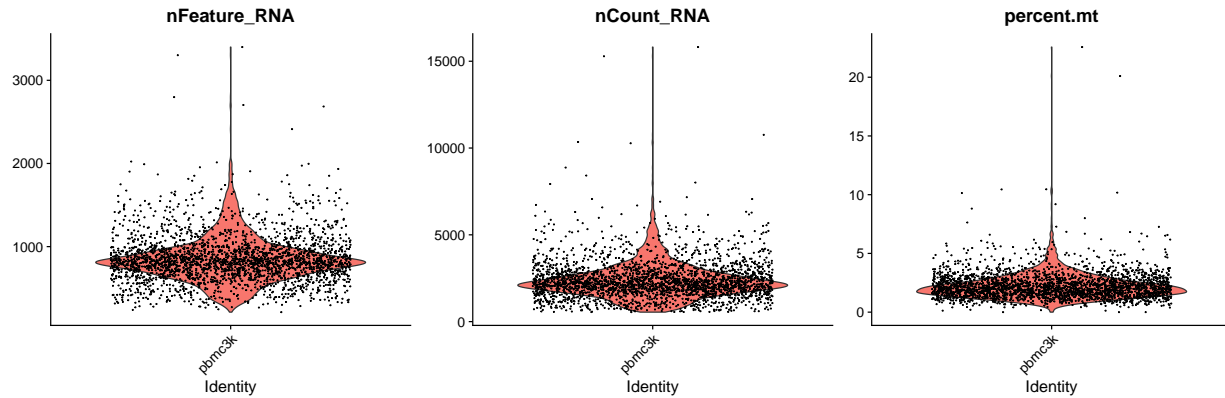
Stap 1: De data wordt ingeladen de data is gedownload van `vignettes/pbmc3k_tutorial`. Het is een data set van PMBC cellen waarbij 2700 losse cellen zijn gesequenced met Illumina Next Seq 500. Deze data wordt ingeladen.

Stap 2.1: Er wordt een Seurat object gemaakt. De data wordt vervolgens met behulp van de packages seurat opgeslagen als object e zodat dit gebruikt kan worden voor preprocessing en data analyse.

Stap 2.2: De 0 values worden weg gefilterd en een compactere versie van het object wordt opgeslagen. Deze worden namelijk niet gebruikt en nemen wel veel geheugen in beslag.

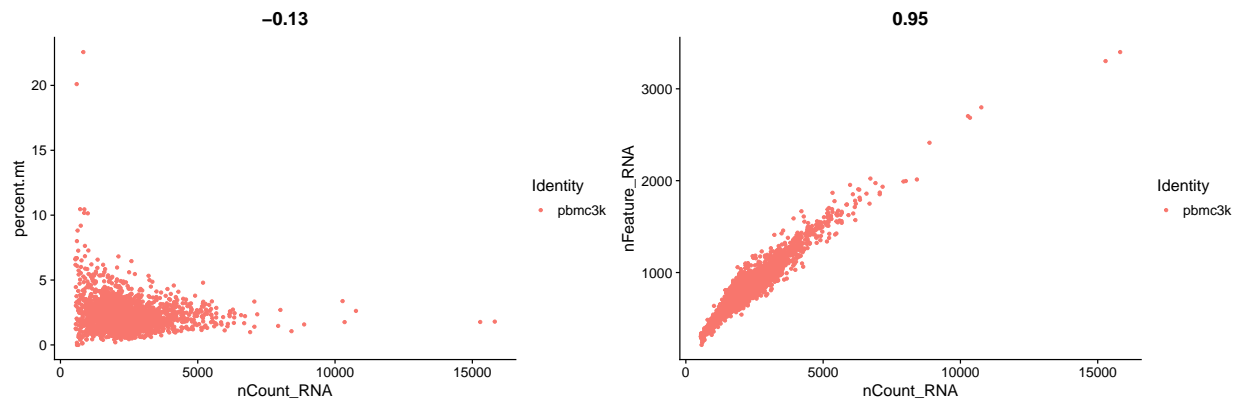
Stap 3.1: Kwaliteits check uitvoeren en cellen selecteren voor analyse. Om de betrouwbaarheid van de analyse te waarborgen wordt eerst gekeken naar welke cellen meegenomen worden voor analyse. Daarvoor wordt eerst het percentage mitochondriale RNA toegevoegd aan het seurat object

Stap 3.2: Om de kwaliteit van de data te beoordelen wordt deze gevisualiseerd in een violin plot van `nFeature_RNA` (unieke genen per cel), `nCount_RNA` (totaal aantal moleculen), en `percent.mt` (mitochondriale expressie). Een laag aantal unieke genen of moleculen kan namelijk wijzen op lege of slechte kwaliteit van de cellen. En te hoog kan wijzen op dubbele metingen. Het mitochondriaal percentage wordt gebruikt om te kijken naar dode of slechte kwaliteit van cellen, een te hoog percentage wijst hier namelijk op.



In deze violin plot wordt dit weergegeven en is te zien hoe de data verdeeld is. Optimaal is om zo veel mogelijk van deze data mee te nemen.

Stap 3.3: Visualisatie in Feature Scatter. Om de relatie tussen de nCount en percent.mt of nFeature aan te geven wordt de Pearson correlatie geanalyseerd. Deze is voor nCount en nFeature 0.95 wat aangeeft dat het een sterke correlatie heeft. Voor nCount en percent.mt is deze -0.13 wat wijst op geen directe correlatie.



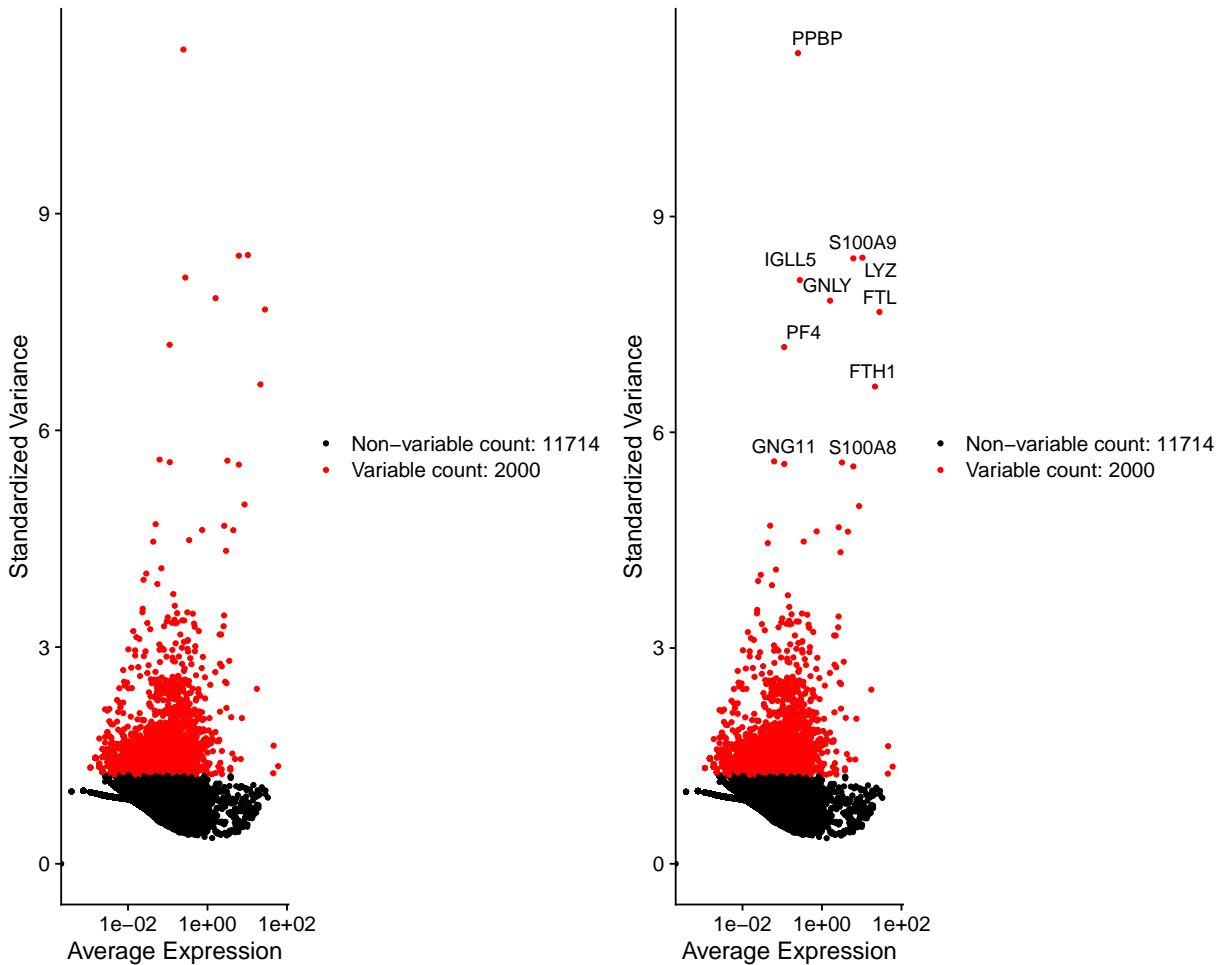
Stap 3.4. Filtering Voor de filtering is bij de tutorial is gekozen voor >200 gene expression (lage of lege droplets hebben vaak weinig genen), <2500 cellen met veel genen (dubbel getelde droplets hebben vaak hoge genen), <5 mitochondriale expressie (hogere MT expressie komt vaak door lage kwaliteit van de cel)

Stap 3.5. Normalisatie van data volgens standaard normalisatie. Er is een veel gebruikte schaling methode gebruikt waarbij de feature expressie van iedere cel genormaliseerd wordt voor de normale expressie. Zodat bij analyse gerekend wordt met het zelfde aantal RNA moleculen voor iedere cel. Dit wordt gedaan door daarna te schalen met factor 10000 en log te transformeren. De log transformatie is nodig om te zorgen dat genen met hoge counts vergelijkbaar is met genen met lage counts. Ook helpt het bij het normaal verdelen van de data voor PCA analyse.

Stap 3.6. Feature selection. De genen die veel verschil in expressie hebben per cel worden geselecteerd. Uit eerder onderzoek is namelijk gebleken dat focus op deze genen helpt bij downstream analyse van biologisch signaal in single cell datasets. Er zijn hier 2000 features voor een data set om de analyse werkbaar te houden voor de server.

Stap 3.6. De 10 meest variabele genen worden geselecteerd.

Stap 3.7. De 10 meest variabele genen worden geplotted. Hierin worden zijn de zwarte stippen die niet variabel zijn en niet meegenomen worden bij verdere analyse. De rode stippen zijn de variabele genen en de top tien meest variabele genen zijn aangegeven met een label.



Stap 3.8. De data wordt geschaalt volgens standaard procedures voor PCA analyse. Hierbij wordt de expressie van alle genen gelijk zodat de hoeveelheid expressie geen invloed heeft in de verdere analyse.

Stap 3.9. De PCA analyse wordt uitgevoerd en gevisualiseerd.

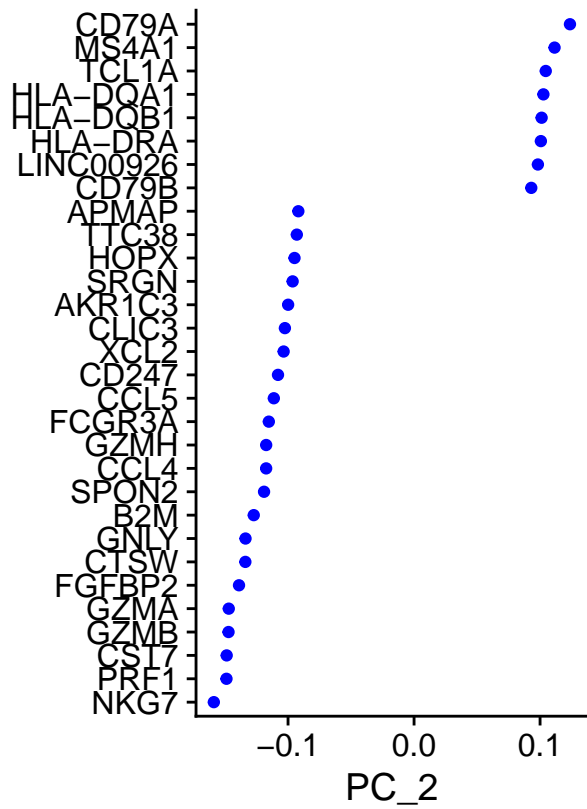
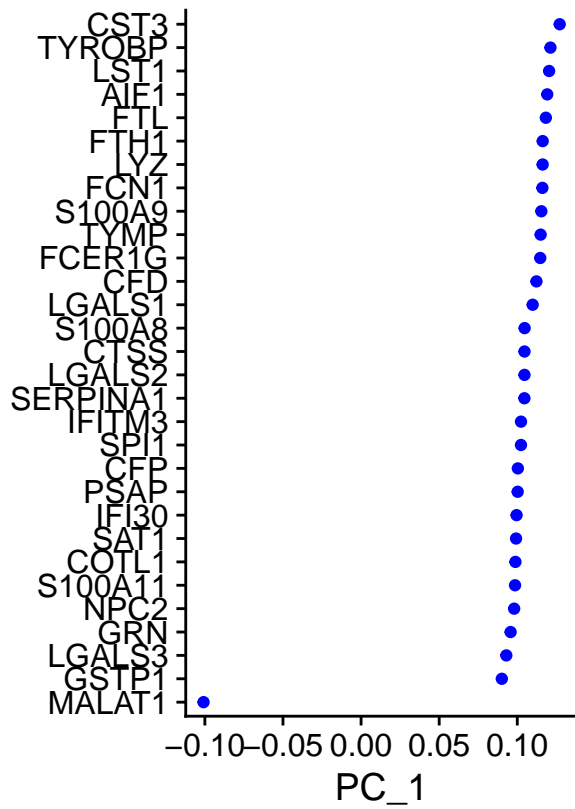
```
## PC_ 1
## Positive: CST3, TYROBP, LST1, AIF1, FTL, FTH1, LYZ, FCN1, S100A9, TYMP
##           FCER1G, CFD, LGALS1, S100A8, CTSS, LGALS2, SERPINA1, IFITM3, SPI1, CFP
##           PSAP, IFI30, SAT1, COTL1, S100A11, NPC2, GRN, LGALS3, GSTP1, PYCARD
## Negative: MALAT1, LTB, IL32, IL7R, CD2, B2M, ACAP1, CD27, STK17A, CTSW
##           CD247, GIMAP5, AQP3, CCL5, SELL, TRAF3IP3, GZMA, MAL, CST7, ITM2A
##           MYC, GIMAP7, HOPX, BEX2, LDLRAP1, GZMK, ETS1, ZAP70, TNFAIP8, RIC3
## PC_ 2
## Positive: CD79A, MS4A1, TCL1A, HLA-DQA1, HLA-DQB1, HLA-DRA, LINC00926, CD79B, HLA-DRB1, CD74
##           HLA-DMA, HLA-DPB1, HLA-DQA2, CD37, HLA-DRB5, HLA-DMB, HLA-DPA1, FCRLA, HVCN1, LTB
##           BLNK, P2RX5, IGLL5, IRF8, SWAP70, ARHGAP24, FCGR2B, SMIM14, PPP1R14A, C16orf74
## Negative: NKG7, PRF1, CST7, GZMB, GZMA, FGFBP2, CTSW, GNLY, B2M, SPON2
##           CCL4, GZMH, FCGR3A, CCL5, CD247, XCL2, CLIC3, AKR1C3, SRGN, HOPX
##           TTC38, APMAP, CTSC, S100A4, IGFBP7, ANXA1, ID2, IL32, XCL1, RHOC
## PC_ 3
## Positive: HLA-DQA1, CD79A, CD79B, HLA-DQB1, HLA-DPB1, HLA-DPA1, CD74, MS4A1, HLA-DRB1, HLA-DRA
```

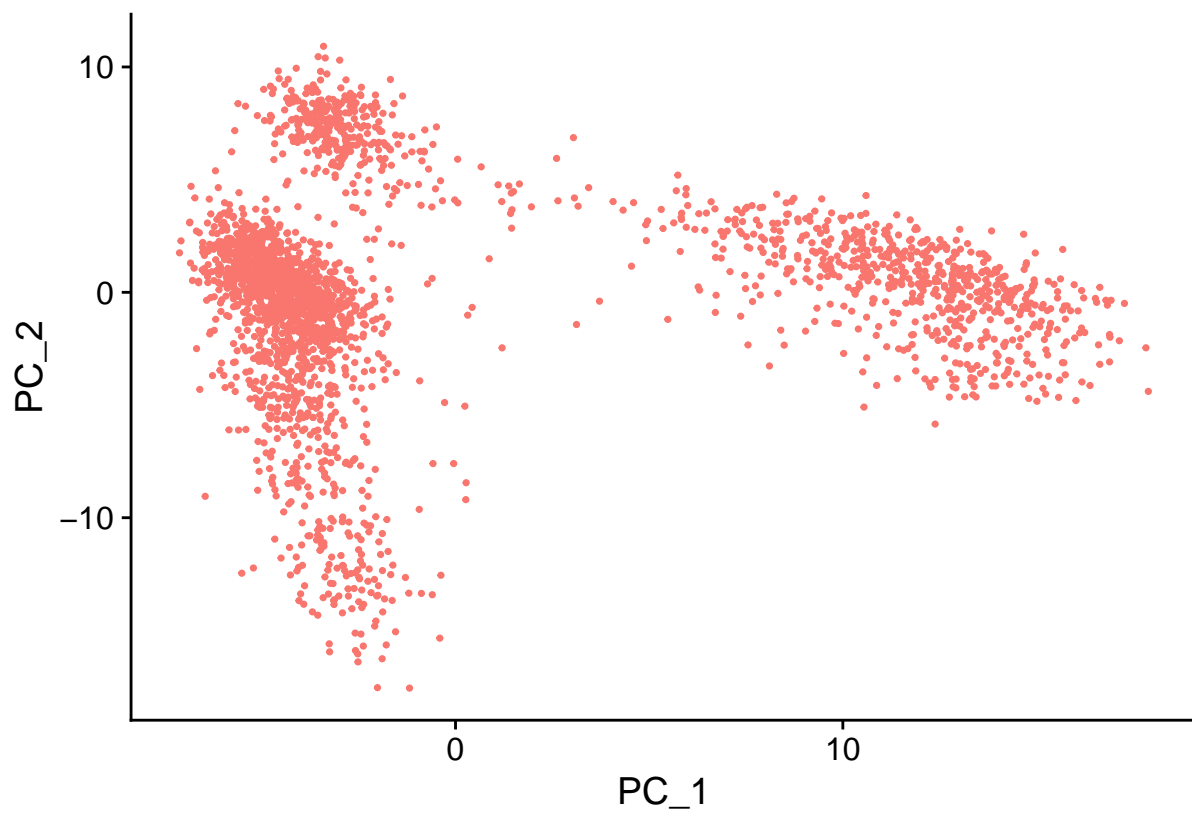
```

##      HLA-DRB5, HLA-DQA2, TCL1A, LINC00926, HLA-DMB, HLA-DMA, CD37, HVCN1, FCRLA, IRF8
##      PLAC8, BLNK, MALAT1, SMIM14, PLD4, P2RX5, IGLL5, LAT2, SWAP70, FCGR2B
## Negative: PPBP, PF4, SDPR, SPARC, GNG11, NRGN, GP9, RGS18, TUBB1, CLU
##      HIST1H2AC, AP001189.4, ITGA2B, CD9, TMEM40, PTCRA, CA2, ACRBP, MMD, TREML1
##      NGFRAP1, F13A1, SEPT5, RUFY1, TSC22D1, MPP1, CMTM5, RP11-367G6.3, MYL9, GP1BA
## PC_ 4
## Positive: HLA-DQA1, CD79B, CD79A, MS4A1, HLA-DQB1, CD74, HIST1H2AC, HLA-DPB1, PF4, SDPR
##      TCL1A, HLA-DRB1, HLA-DPA1, HLA-DQA2, PPBP, HLA-DRA, LINC00926, GNG11, SPARC, HLA-DRB5
##      GP9, AP001189.4, CA2, PTCRA, CD9, NRGN, RGS18, CLU, TUBB1, GZMB
## Negative: VIM, IL7R, S100A6, IL32, S100A8, S100A4, GIMAP7, S100A10, S100A9, MAL
##      AQP3, CD2, CD14, FYB, LGALS2, GIMAP4, ANXA1, CD27, FCN1, RBP7
##      LYZ, S100A11, GIMAP5, MS4A6A, S100A12, FOLR3, TRABD2A, AIF1, IL8, IFI6
## PC_ 5
## Positive: GZMB, NKG7, S100A8, FGFBP2, GNLY, CCL4, CST7, PRF1, GZMA, SPON2
##      GZMH, S100A9, LGALS2, CCL3, CTSW, XCL2, CD14, CLIC3, S100A12, RBP7
##      CCL5, MS4A6A, GSTP1, FOLR3, IGFBP7, TYROBP, TTC38, AKR1C3, XCL1, HOPX
## Negative: LTB, IL7R, CKB, VIM, MS4A7, AQP3, CYTIP, RP11-290F20.3, SIGLEC10, HMOX1
##      LILRB2, PTGES3, MAL, CD27, HN1, CD2, GDI2, CORO1B, ANXA5, TUBA1B
##      FAM110A, ATP1A1, TRADD, PPA1, CCDC109B, ABRACL, CTD-2006K23.1, WARS, VMO1, FYB

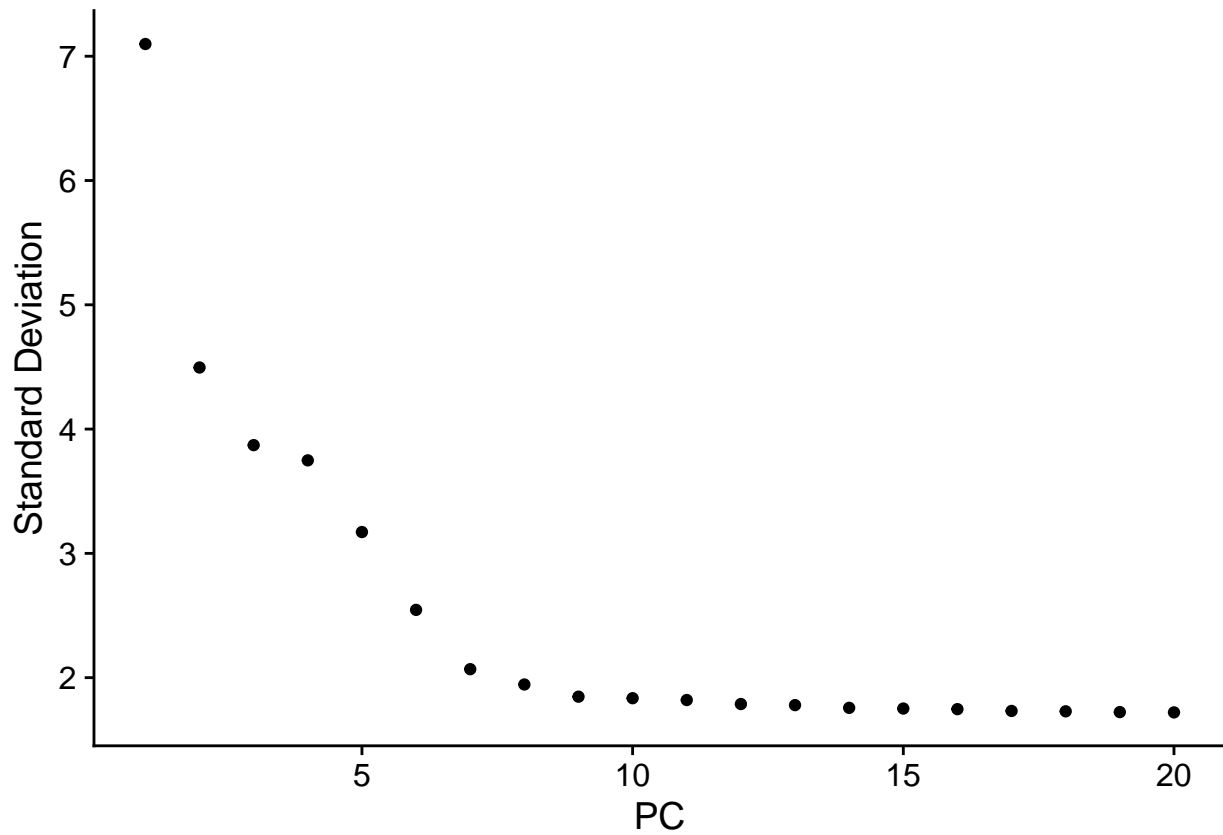
## PC_ 1
## Positive: CST3, TYROBP, LST1, AIF1, FTL
## Negative: MALAT1, LTB, IL32, IL7R, CD2
## PC_ 2
## Positive: CD79A, MS4A1, TCL1A, HLA-DQA1, HLA-DQB1
## Negative: NKG7, PRF1, CST7, GZMB, GZMA
## PC_ 3
## Positive: HLA-DQA1, CD79A, CD79B, HLA-DQB1, HLA-DPB1
## Negative: PPBP, PF4, SDPR, SPARC, GNG11
## PC_ 4
## Positive: HLA-DQA1, CD79B, CD79A, MS4A1, HLA-DQB1
## Negative: VIM, IL7R, S100A6, IL32, S100A8
## PC_ 5
## Positive: GZMB, NKG7, S100A8, FGFBP2, GNLY
## Negative: LTB, IL7R, CKB, VIM, MS4A7

```







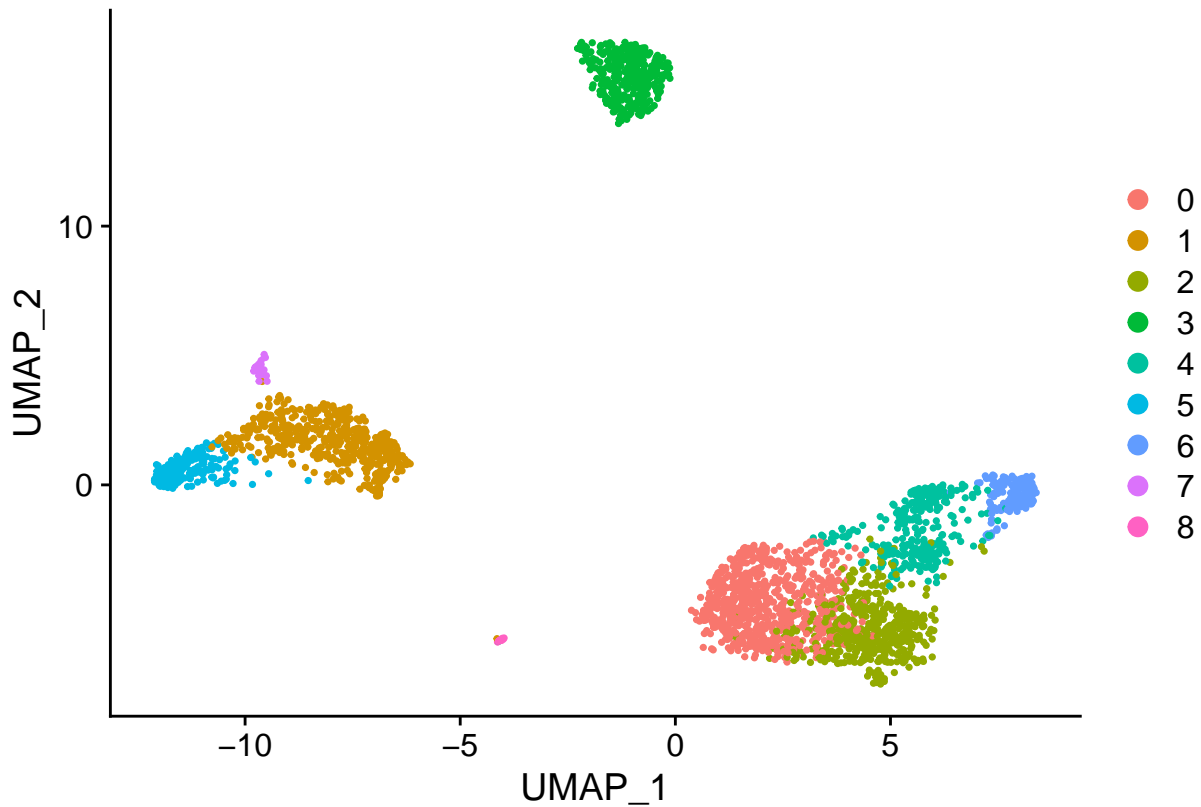


Op basis van deze ElbowPlot zijn de eerste 10 PCs geselecteerd omdat de “elbow” stopt rond PC9-10, wat wijst op een signaal in de eerste 10 PCs.

Stap 3.11. De cellen worden geclusterd op basis van de eerste 10 PCs.

Stap 4.1 De UMAP maken. De gevormde clusters worden weergegeven aan de hand van 10 PCs. In deze UMAP wordt weergegeven welke clusters er zijn en hoe deze zijn verdeeld.

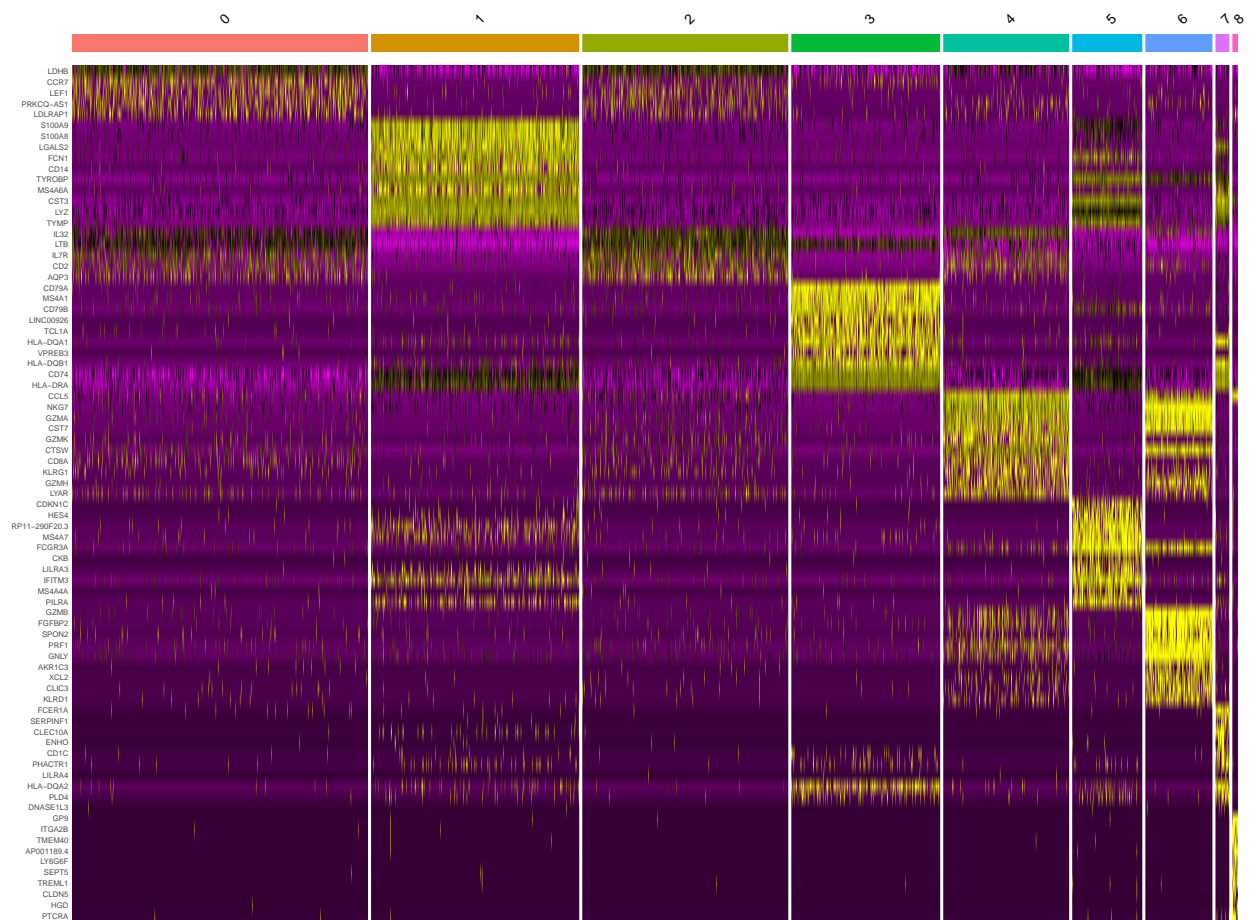




Stap 5.1 Cluster biomarkers vinden Alle markers van alle clusters worden gevonden en alleen de positieve worden gerapporteerd. Dit is belangrijk om te bepalen wat voor cellen er in de clusters zitten, en kan gebruikt worden om te bepalen welke clusters biologisch relevant zijn voor verder onderzoek.

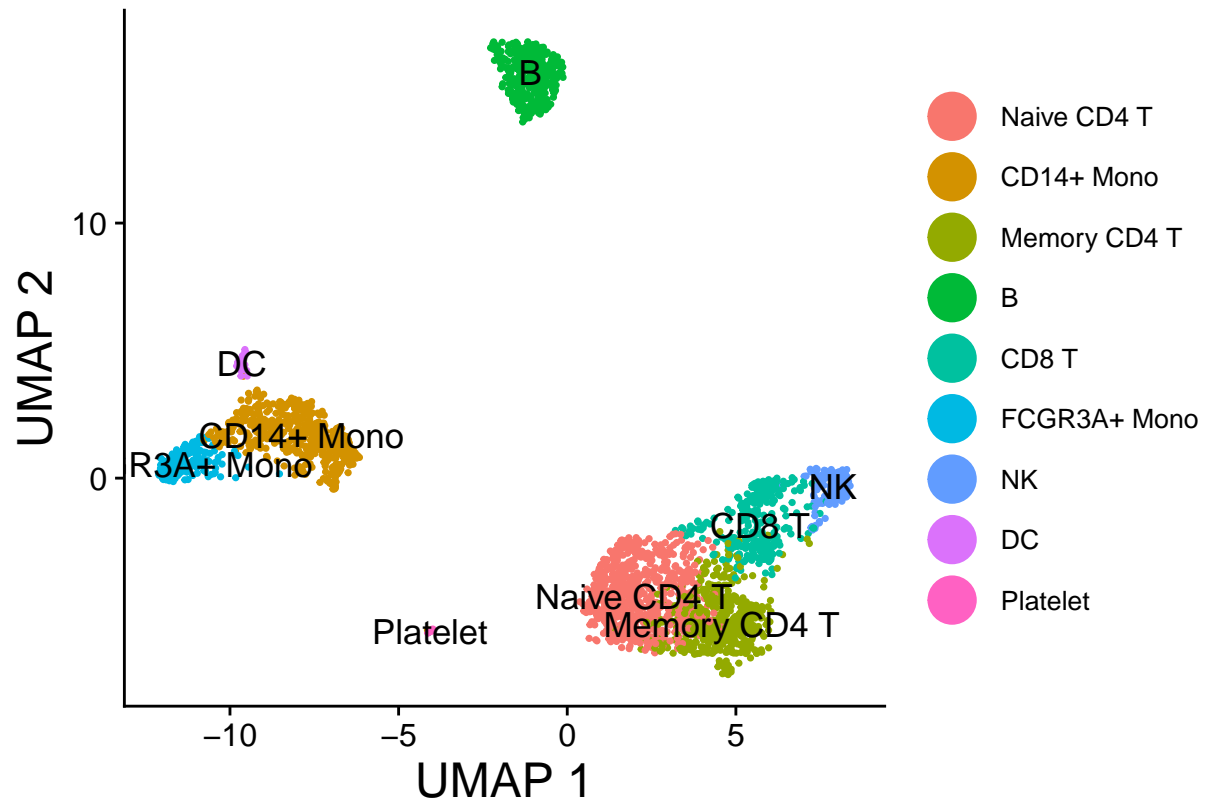
```
## # A tibble: 939 x 7
## # Groups:   cluster [9]
##       p_val avg_log2FC pct.1 pct.2 p_val_adj cluster gene
##       <dbl>      <dbl> <dbl> <dbl>      <dbl> <fct>   <chr>
## 1 3.75e-112      1.09 0.912 0.592 5.14e-108 0      LDHB
## 2 9.57e- 88      1.36 0.447 0.108 1.31e- 83 0      CCR7
## 3 1.35e- 51      1.08 0.342 0.103 1.86e- 47 0      LEF1
## 4 6.27e- 43      1.02 0.33  0.112 8.60e- 39 0      PRKCQ-AS1
## 5 6.26e- 30      1.10 0.247 0.085 8.59e- 26 0      LDLRAP1
## 6 0              5.57 0.996 0.215 0          1      S100A9
## 7 0              5.48 0.975 0.121 0          1      S100A8
## 8 0              3.81 0.909 0.059 0          1      LGALS2
## 9 0              3.40 0.952 0.15  0          1      FCN1
## 10 1.03e-295     2.82 0.667 0.027 1.42e-291 1      CD14
## # i 929 more rows
```

Stap 5.2 Een heatmap wordt gemaakt voor de top 20 markers. Er wordt aangegeven in welke clusters de genen voorkomen.



Stap 5.3. De cell type worden aan de clusters gekoppeld.

Stap 5.4. Er wordt een UMAP gemaakt waarin wordt aangegeven welk cluster welk celtype is.



Conclusie: Het is gelukt om door het volgen van de tutorial de preprocessing en visualisatie van de aangereikte data uit te voeren.