seurat_tutorial

2025-09-22

seurat tutorial

Voor het project gaan we analyses uitvoeren met Seurat. Om kennis te maken met het progamma wordt er een tutorial uitgevoerd van Seurat. Seurat is een pakket in Rstudio voor de analyse van singel-cell RNA sequencing (scRNA-seq) data. Seurat beid veel functies aan voor het verwerken, visualiseren en het interpreteren van de genexpressie op celniveau. In de Seurat tutorial worden de volgende stappen doorlopen - inladen van de data - QC van de data en selecteren van de cellen voor vervolganalyse - normaliseren van de data en het schalen van de data - Dimensionaliteitsreductie met PCA - clusteren van de genen - Visualisatie van resultaten in overzichtelijke plots - Identificatie van markers die specifieke celtypen onderscheiden

doel

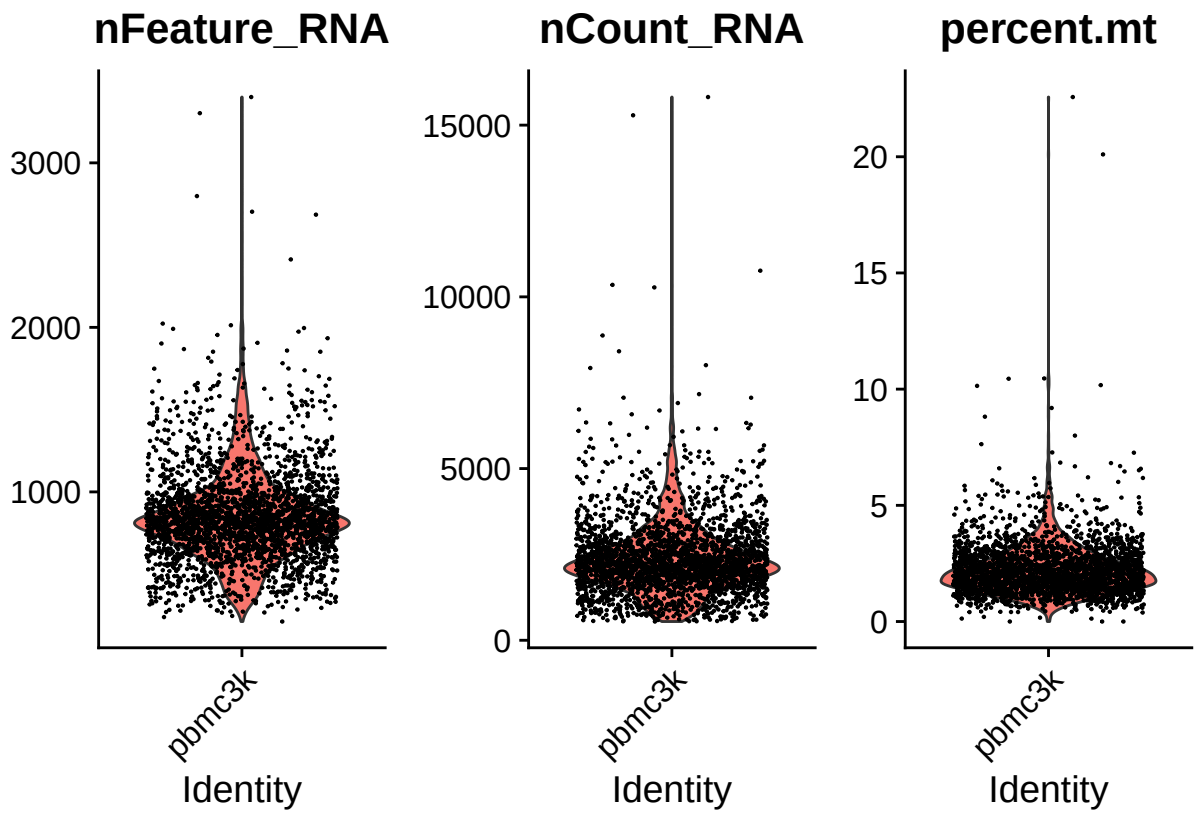
Het doel is om vertrouwd te raken met de verschillende functies van Seurat en te leren hoe ik deze functies kan gebruiken om de RNA-seq data te verwerken, visualiseren en analyseren.

hypothese

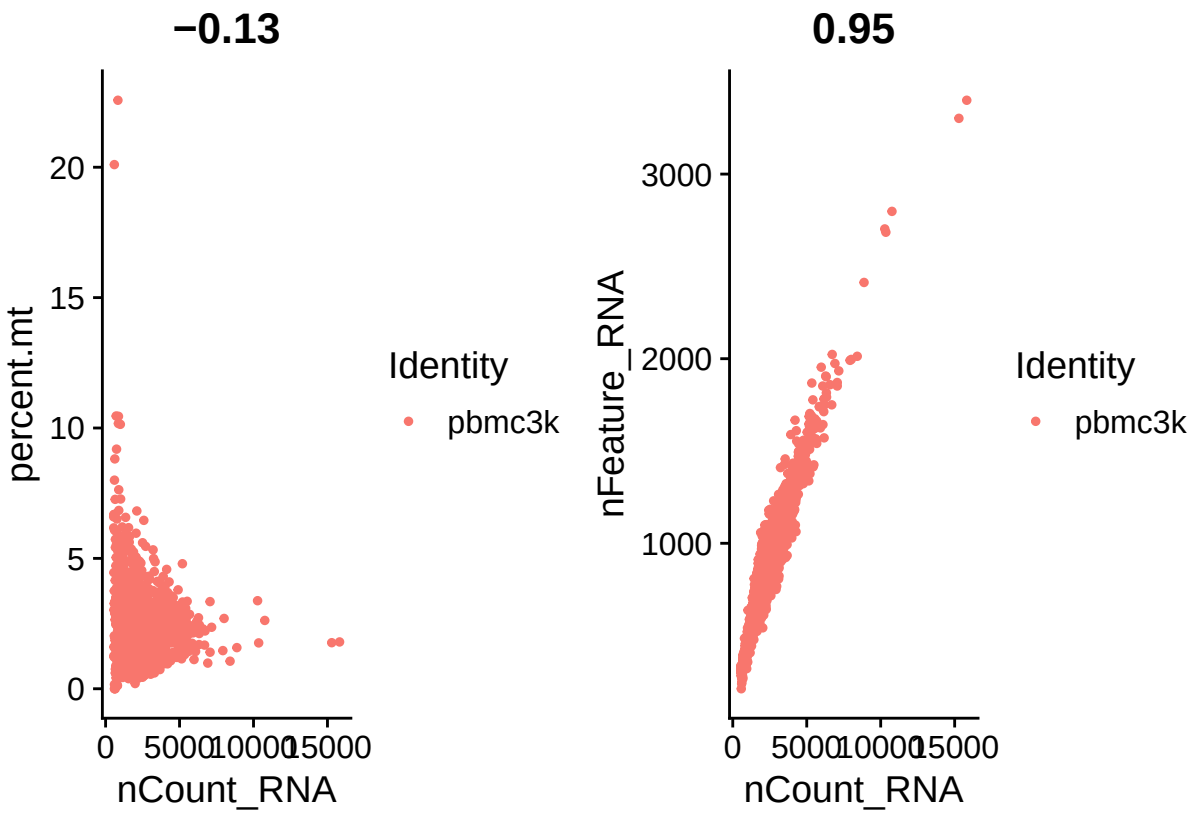
Er wordt verwacht dat de analyse met Seurat verschillende celtypen binnen de dataset kan clusteren en idetificeren en dat deze celtypen duidelijke verschillen in genexpressie laten zien.

het inladen van de data

De data wordt inladen als een object. Hierna wordt er een extra collom aangemaakt waarbij het pecentage mitochondriaal RNA berekend wordt. cellen die lage kwaliteit hebben vertonen vaak een hoog percentage mitochondriaal RNA. Om dit te visualiseren wordt er een vulcanoplot gemaakt. In de plot is nFeature_RNA: Hierbij wordt er gekeken hoe veel unieke genen er in de cel worden gedetecteerd, nCount_RNA: Hier wordt de som van het totaal aantal RNA-moleculen weergeven en percent.mt: hierbij wordt weergeven hoeveel mitochondriaal RNA een cel bezit.



met een featurescatter kan er gevisulaiseerd worden wat de relatie is tussen verschillende features. hierbij wordt weergeven de nCount_RNA tegen de percent.mt. Hier kan je zien of cellen met meer reads ook meer mitochondriale expressie hebben. wanneer hier een positieve corrolatie is, kan dit duiden op stress of dode cellen. In plot 2 wordt nCount_RNA vergeleken tegenover nFeature_RNA. Normaal geeft meer reads ook meer genen. Cellen die hier buiten vallen hebben mogelijk lage kwaliteit of zijn dubbelen cellen.



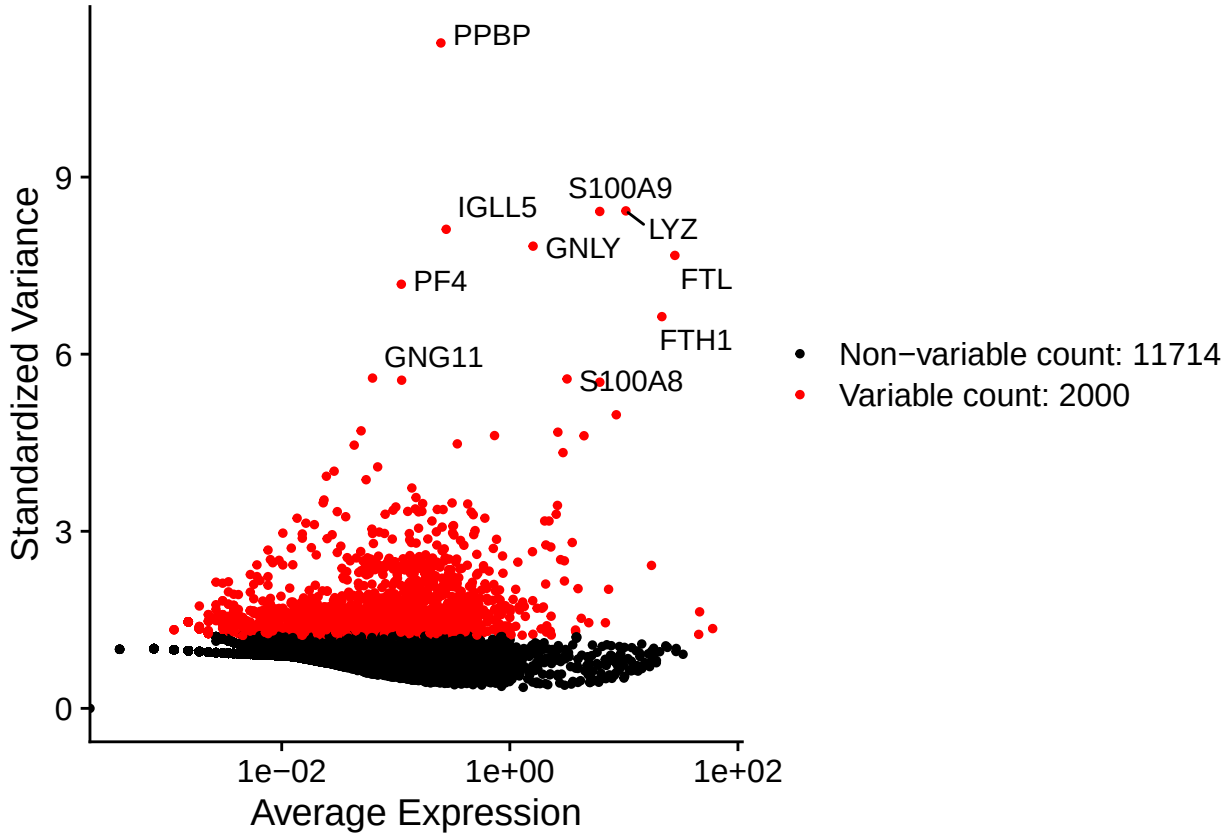
De cellen worden gefilterd op feature count van meer dan 2500 of minder dan 200 features hebben. Er word ook gefilterd op minder dan 5% mitochondriaal pecentage.

normaliseren van de data

het normaliseren van de data wordt gedaan door middel van de log methode. deze functie staat standaard op 10.000 ingesteld.

feature selectie van de data

De data moet geselecteerd worden op genen die de meeste variabiliteit tonen. Dit zorgt ervoor dat bij verdere analyses er gefocused wordt op genen die variabel van elkaar zijn. Deze functie staat standaard ingesteld op 2000 features per dataset. De data wordt weergeven in een VariableFeaturePlot. In de plot staan de top 10 meest variabele genen weergeven. De labels geven de namen van de genen aan in het plot.



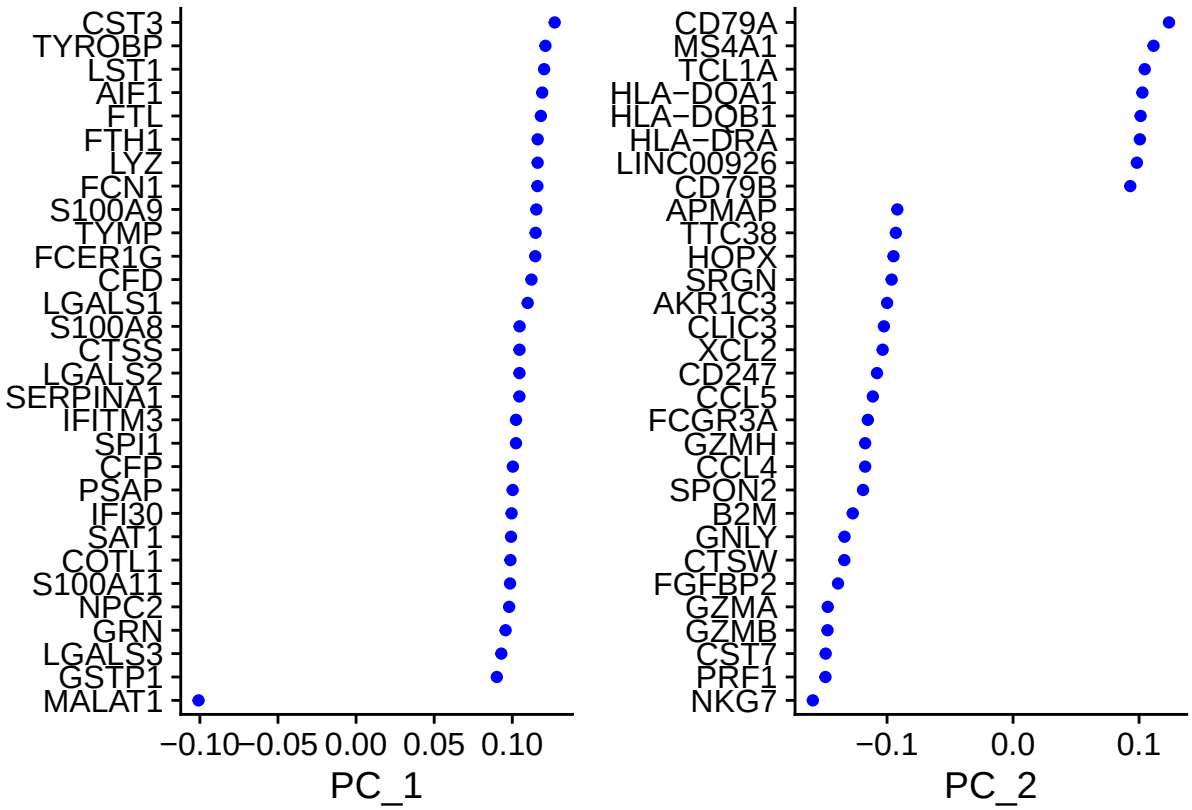
schalen van de data

De functie zorgt ervoor dat de data vergelijkbaar wordt op de schaal en het gemiddelde. Dit zorgt ervoor dat de data beter verwerkt wordt bij verdere analyse zoals bij de PCA. Hierdoor wordt voorkomen dat genen die veel tot expressie komen niet overheersen.

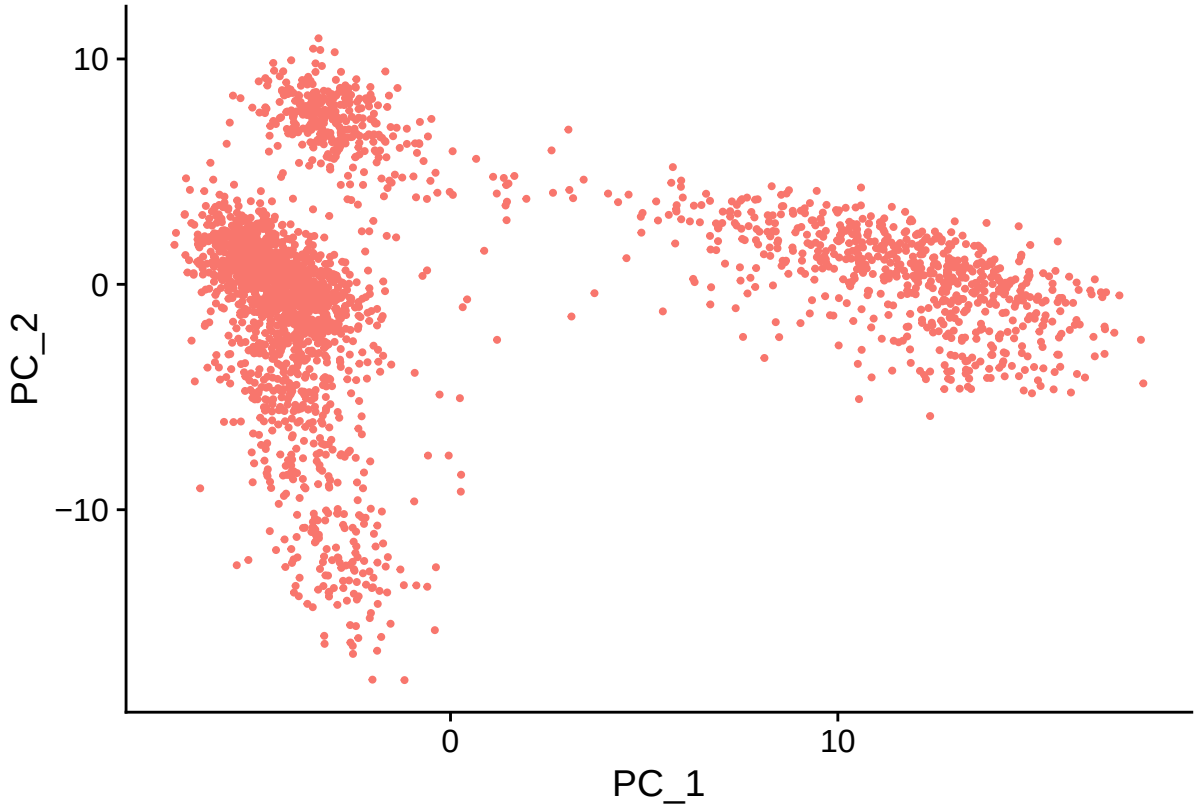
Bij een PCA analyse wordt de grote hoeveelheid data vereenvoudigt. Dit wordt gedaan door te kijken welke De data kan ook in grafieken worden weergeven voor een vergemakkelijkeren weergaven. hieronder worden de eerse 5 PC's weergeven met daarin 5 features.

```
## 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
```

De data kan ook gevisualiseerd worden in een 2D plot. Hierin worden de eerste twee dimenties weergeven. De top genen worden weergeven.

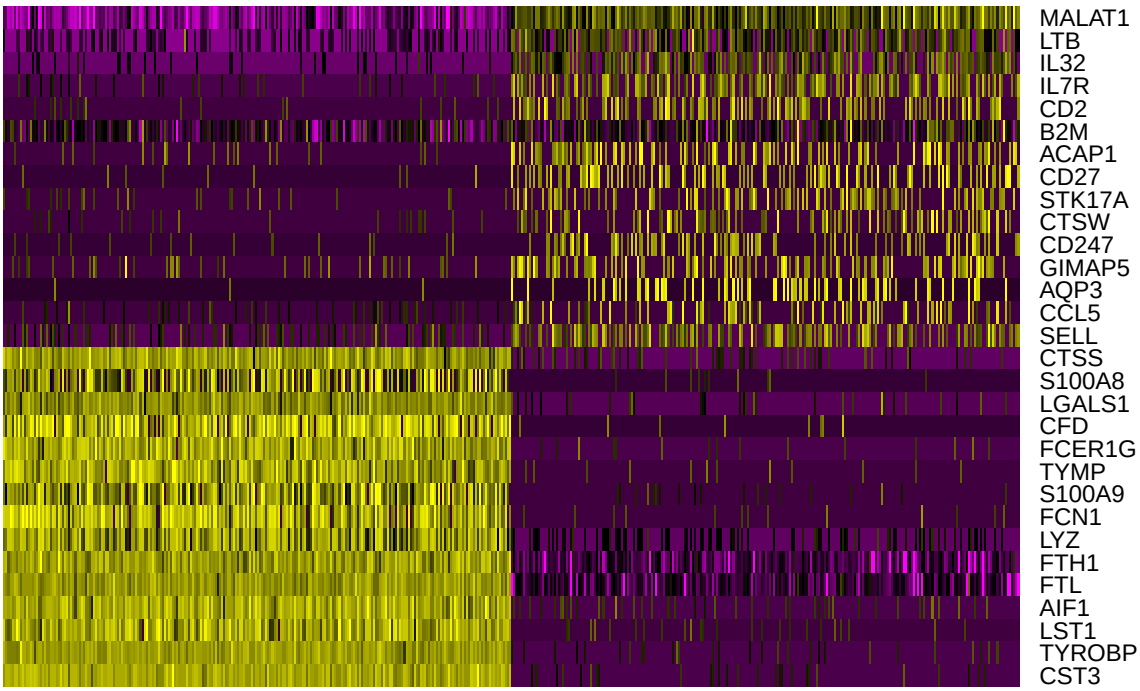


De data kan ook worden weergeven in een dimplot. Een dimplot is een plot waar alle cellen zijn eigen puntje hebben. PC1 en PC2 zijn tegenover elkaar uitgezet.

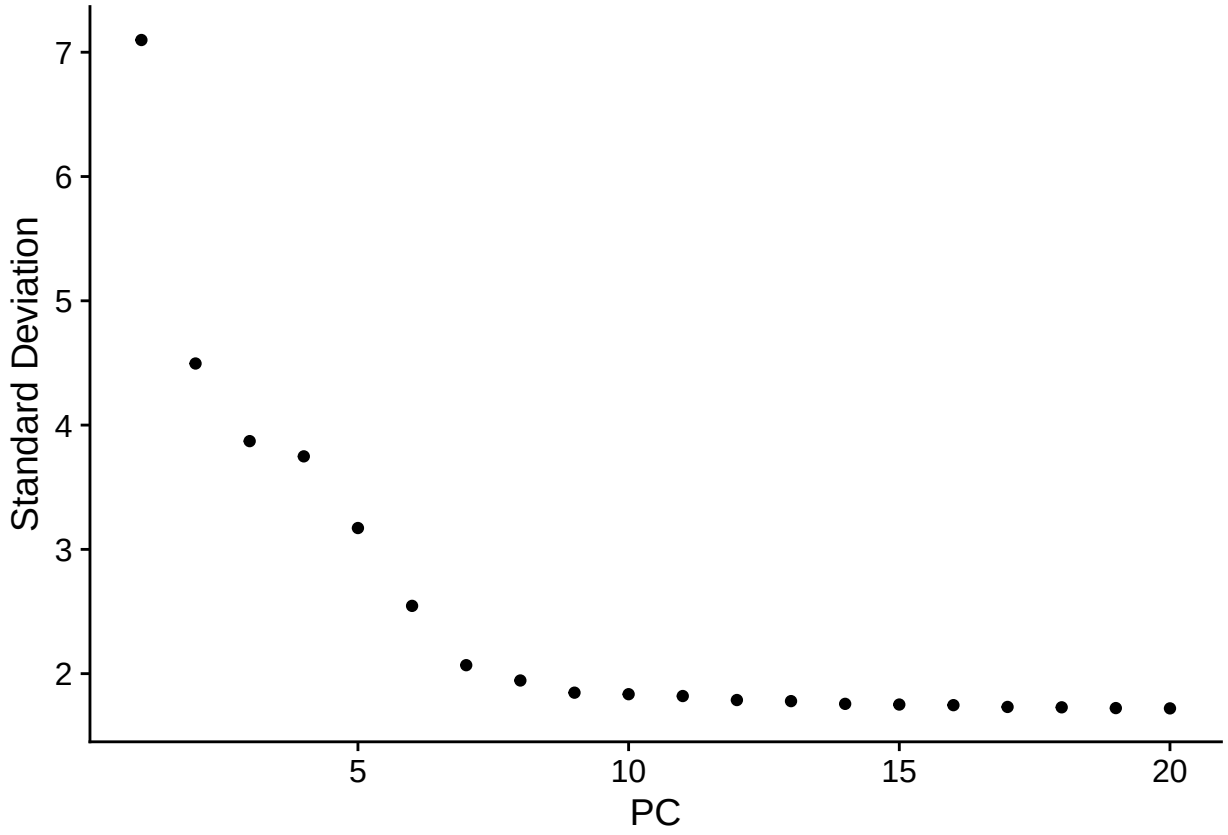


Een heatmap wordt gebruikt om te kijken welke onderdelen de grootste variatie van elkaar hebben. Dit kan ook gebruikt worden om te bepalen welke PC's er meegenomen worden voor verdere analyse downstream. Door cellen op een getal in te stellen, worden de meest extreme cellen aan beide uiteinden van het spectrum weergegeven, wat het plotten van grote datasets aanzienlijk versnelt.

PC_1

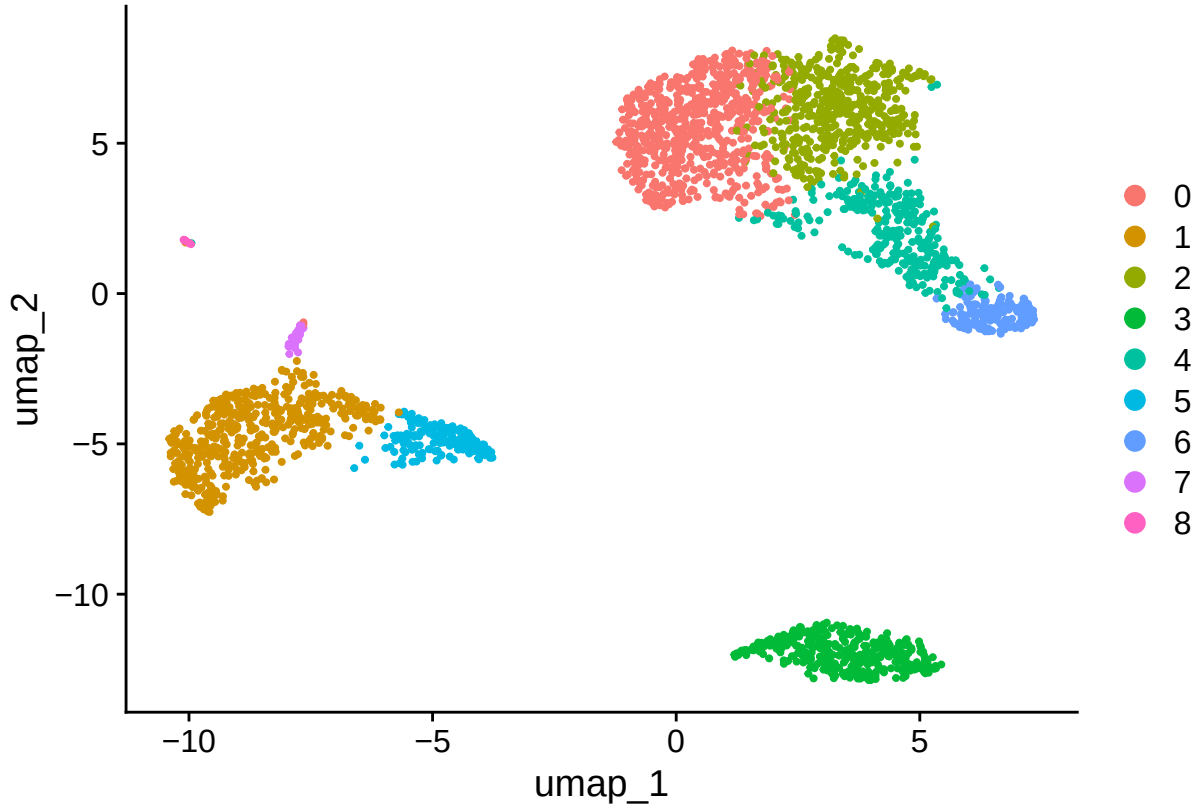


Clusters worden gebaseerd op de PCA waardes. Dit zijn correlaties tussen cellen die met de genexpressie op elkaar lijken. om te bepalen hoeveel van de PC's we mee nemen wordt er een ElbowPlot van de PC's gemaakt. In de plot wordt weergeven hoeveel variatie de PC's bezitten.



Aan de hand van de grafiek is er voor gekozen om de eerste 10 PC's mee te nemen met de verdere dataverwerking. De data moet worden geclusterd. Voor het maken van de clusters moet er eerst bekeken worden wat de ruimte moet zijn tussen de verschillende cellen

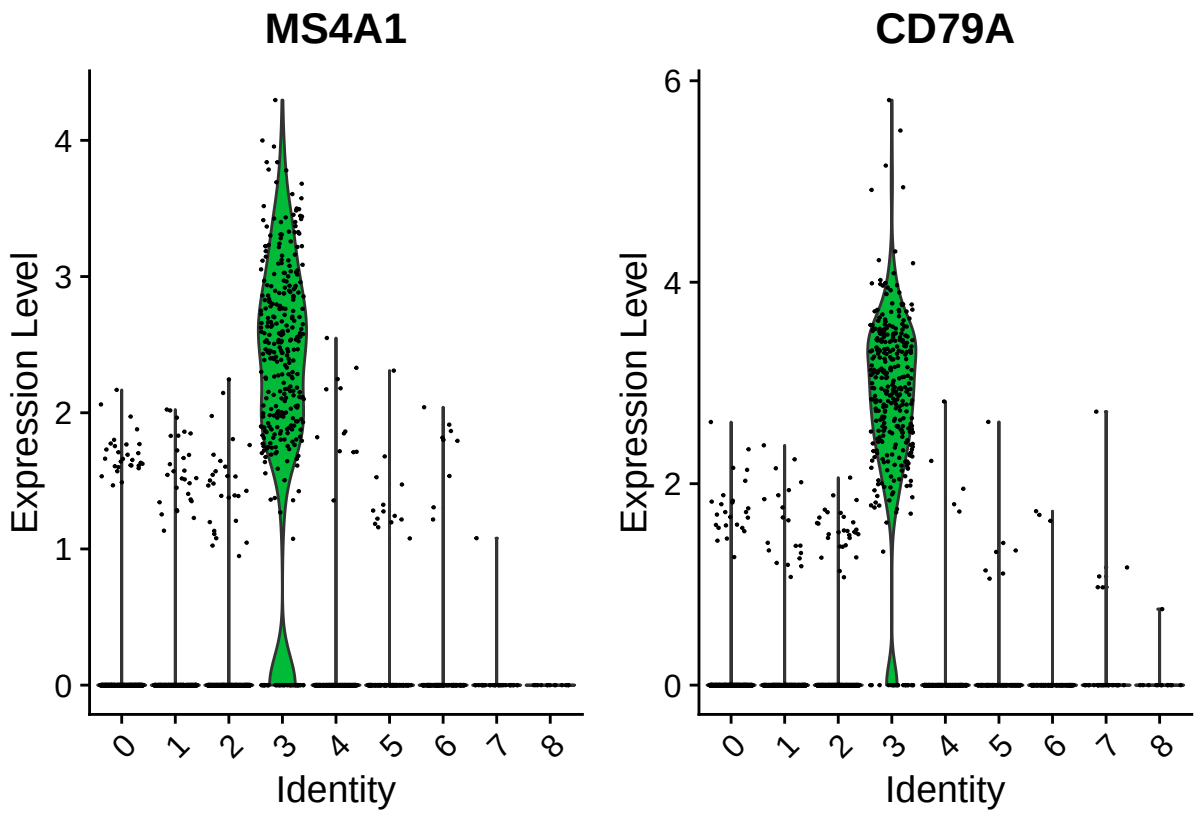
```
## Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck
##
## Number of nodes: 2638
## Number of edges: 95927
##
## Running Louvain algorithm...
## Maximum modularity in 10 random starts: 0.8728
## Number of communities: 9
## Elapsed time: 0 seconds
```



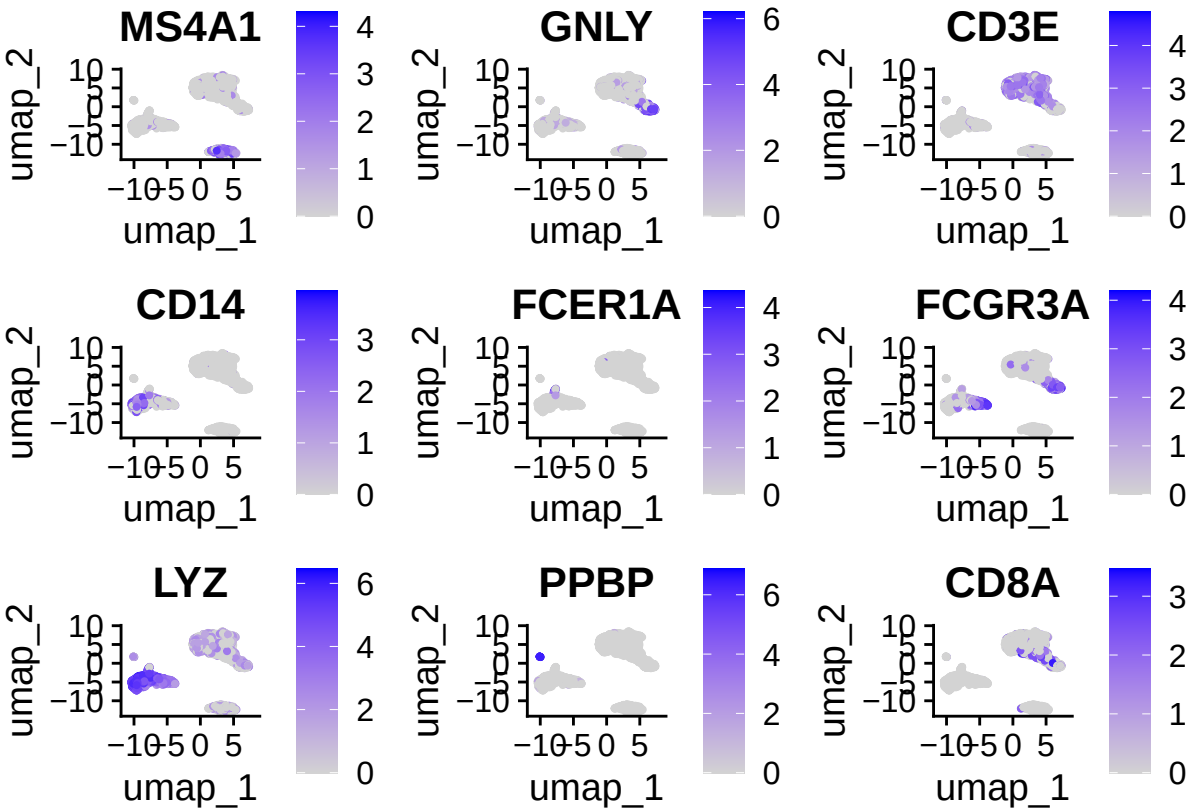
Nu de data uitgezet is in clusters kan er ook gekeken worden of de verschillende clusters ook unieke biomarkers hebben.

De unike biomarkers kunnen op verschillende manieren weergegeven worden. Een manier is met een VlnPlot. Hierin kan je zelf aangeven welke genen je wilt bekijken en wordt er weergeven in welke clusters het gen tot expressie komt.

```
##       myAUC  avg_diff power avg_log2FC pct.1 pct.2
## RPS12 0.827 0.5059247 0.654  0.7387061 1.000 0.991
## RPS6  0.826 0.4762402 0.652  0.6934523 1.000 0.995
## RPS27 0.824 0.5047203 0.648  0.7372604 0.999 0.992
## RPL32 0.821 0.4294911 0.642  0.6266075 0.999 0.995
## RPS14 0.811 0.4334133 0.622  0.6336957 1.000 0.994
```



Een andere manier is met een FeaturePlot. Dit is een plot om te weergeven welke features waar in de clusters zitten. Dit wordt gevisualiseerd door middenl van een FeaturePlot.


MS4A1
GNLY
CD3E
CD14
FCER1A
FCGR3A
LYZ
PPBP
CD8A
umap_1
umap_2