Seurat_tutorial

2025-11-23

De tutorial die is gevolgd is te vinden op de volgende site: https://satijalab.org/seurat/articles/pbmc3k_tutorial. De dataset die gebruikt wordt voor deze tutorial bestaat uit 2.700 single cells (PBMC's, peripheral blood mononuclear cells) die zijn gesequenced in de Illumina Nextseq 500. De dataset is van 10X Genomics, en de ruwe data kan gevonden worden op deze site: https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz.

QC en selecteren van de cellen

In Figure 1 worden de features, counts en percentage mitochondria weergegeven. Van de features blijven tussen de 200 en 2500 over. Van de mitochodria blijft onder de 5% over, daarboven is de cel waarschijnlijk van slechte kwaliteit of is deze stervende.

Normaliseren van de gegevens

De data wordt genormaliseerd door de totale expressie van elke cel om te zetten naar 10.000 en vervolgens een log-transformatie te nemen (het getal weer kleiner maken). Hierdoor kunnen de getallen per cel met elkaar vergeleken worden.

Kenmerkselectie

In Figure 3 zijn de 2000 cellen met de meest variërende kenmerken zijn rood gekleurd en de 10 meest variërende genen worden in onderstaand figuur aangegeven met hun naam.

De gegevens schalen

De gemiddelde expressie over de cellen wordt op 0 gesteld, ook wordt de variantie over de cellen op 1 gesteld.

Lineaire dimensionale reductie uitvoeren

Met behulp van principal component analyses (PCA) worden er van de oude variabelen, nieuwe variabelen gemaakt door lineaire combinaties te maken. Hieruit ontstaan enkele hoofdcomponenten (PC's). PC1 dekt het grootste deel van de variatie binnen de data. PC2 dekt daarna het grootste deel, enzovoort.

Dimensionaliteit

Om ruis te voorkomen in verdere analyse wordt er gekeken hoeveel PC's meegenomen moeten worden. Hiervoor kan er gebruik gemaakt worden van een elbowplot. In deze plot worden de PC's op de x-as geplaatst en de standaard deviatie op de y-as. Er ontstaat een soort van arm, waarbij de elleboog het keerpunt is. De PC's voor de elleboog worden meegenomen en de PC's na de elleboog worden niet meegenomen. Het verschil tussen de PC's na de elleboog is niet relevant genoeg om mee te nemen. In Figure 7 is de elleboog zichtbaar tussen 9 en 10.

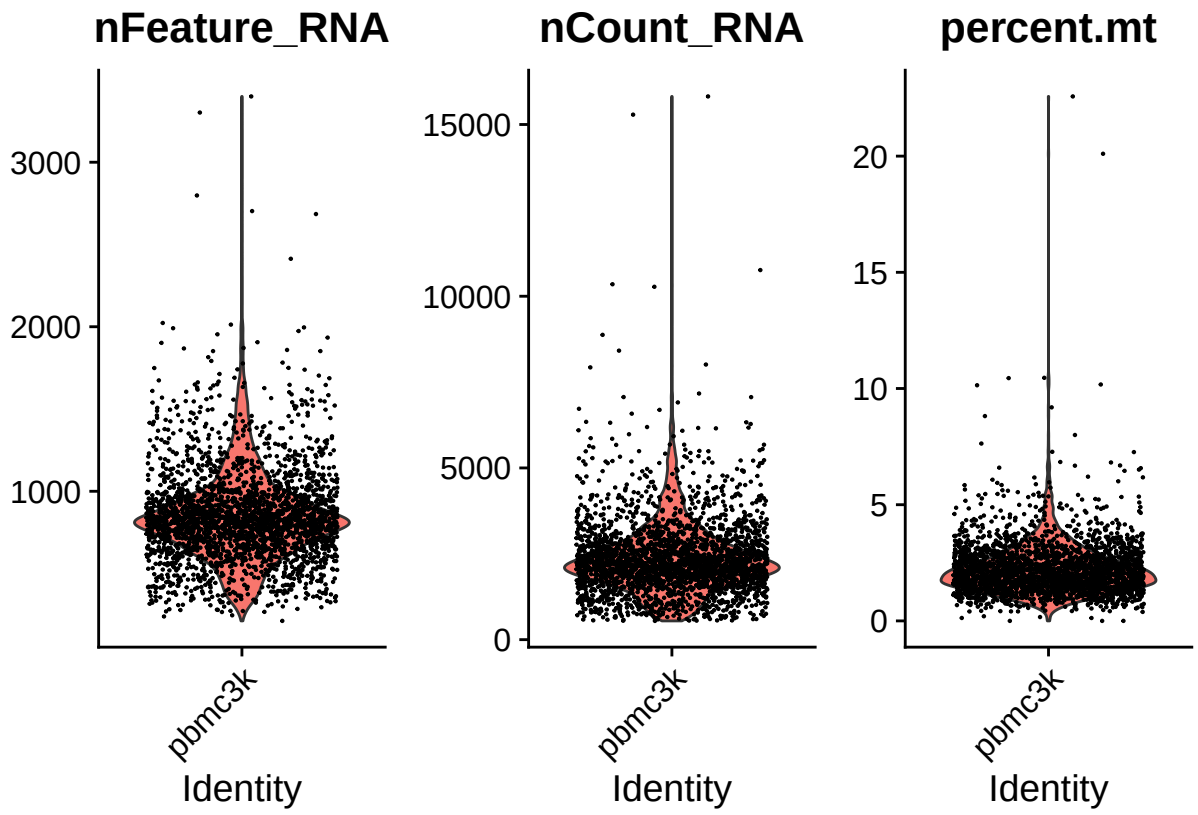

Figure 1: Vioolplots: Er zijn drie vioolplots naast elkaar weergegeven. In alle drie de vioolplots wordt op de x-as de dataset met pbmc's weergegeven. Elke stipje is een cel. Links wordt op de y-as het aantal unieke features weergegeven. In het midden wordt op de y-as het aantal RNA binnen de cel weergegeven. Rechts wordt op de y-as het percentage mitochondriale genen weergegeven.

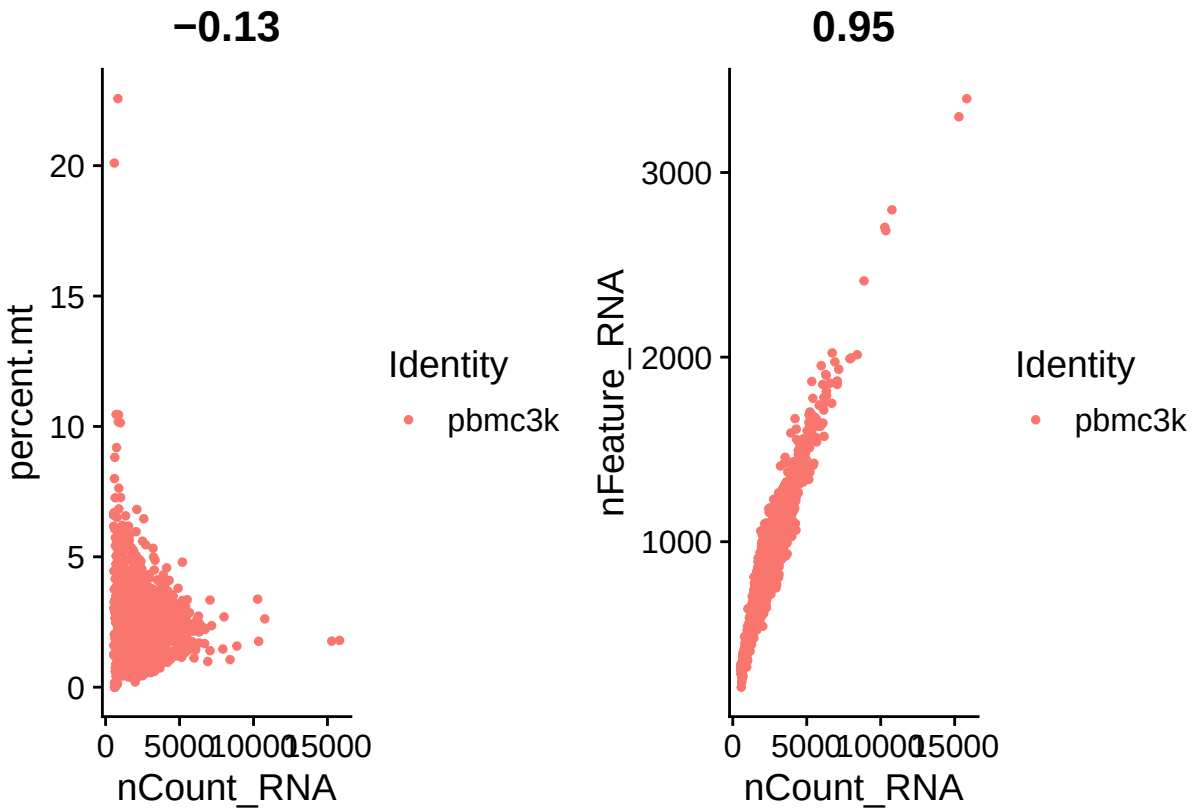

Figure 2: FeatureScatterplots: Er zijn twee scatterplots naast elkaar weergegeven, bij beide scatterplots staat op de x-as het aantal RNA binnen de cel. Elk stipje is een cel. Links is op de y-as het percentage mitochondriale genen weergegeven, waarbij alleen cellen meegenomen zijn die een percentage lager dan 5% mitochondriale genen hebben. Rechts is op de y-as het aantal unieke features per cel weergegeven, waarbij alleen de cellen meegenomen zijn die tussen de 200 en 2500 unieke features hebben.

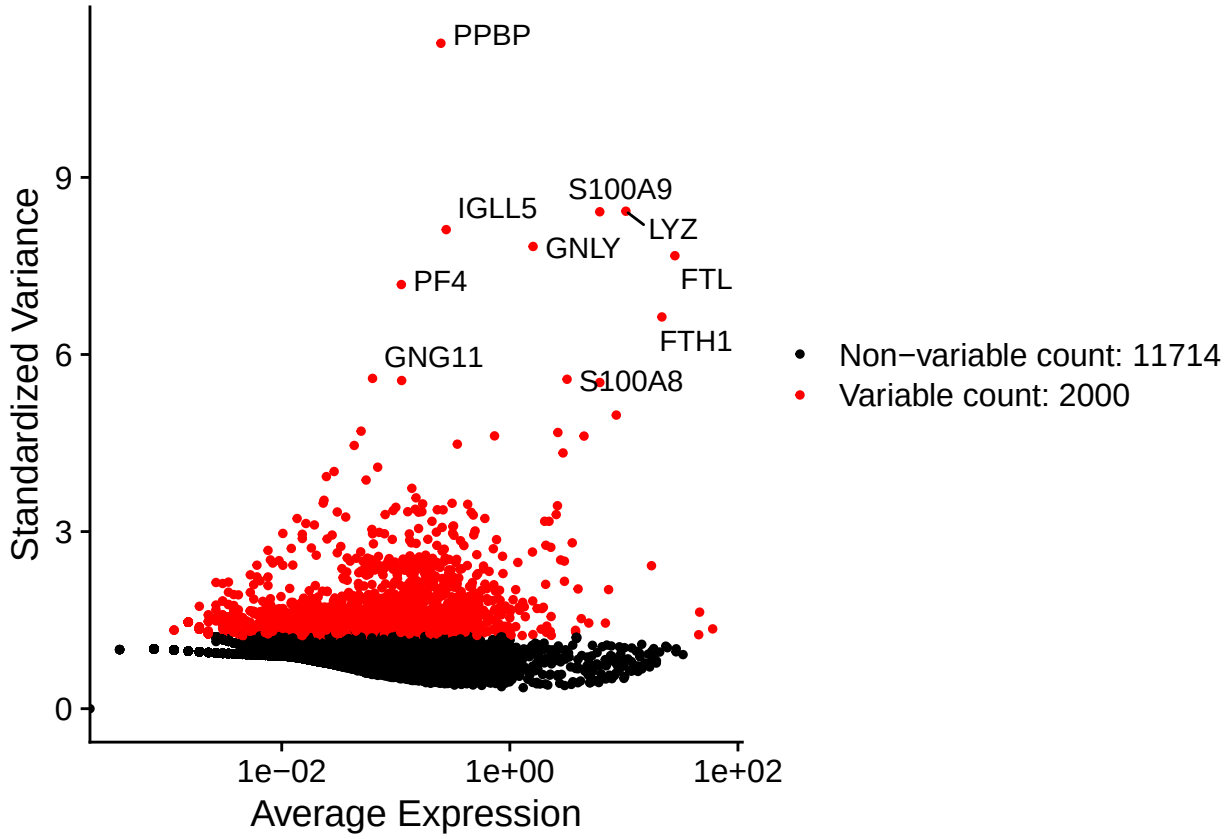

Figure 3: VariableFeaturePlot: Op de x-as wordt de gemiddelde expressie weergegeven. Op de y-as wordt de standaard variantie weergegeven. De 2000 cellen met de meest variërende kenmerken zijn rood gekleurd en de 10 cellen met de meest variërende genen zijn aangegeven met hun naam.

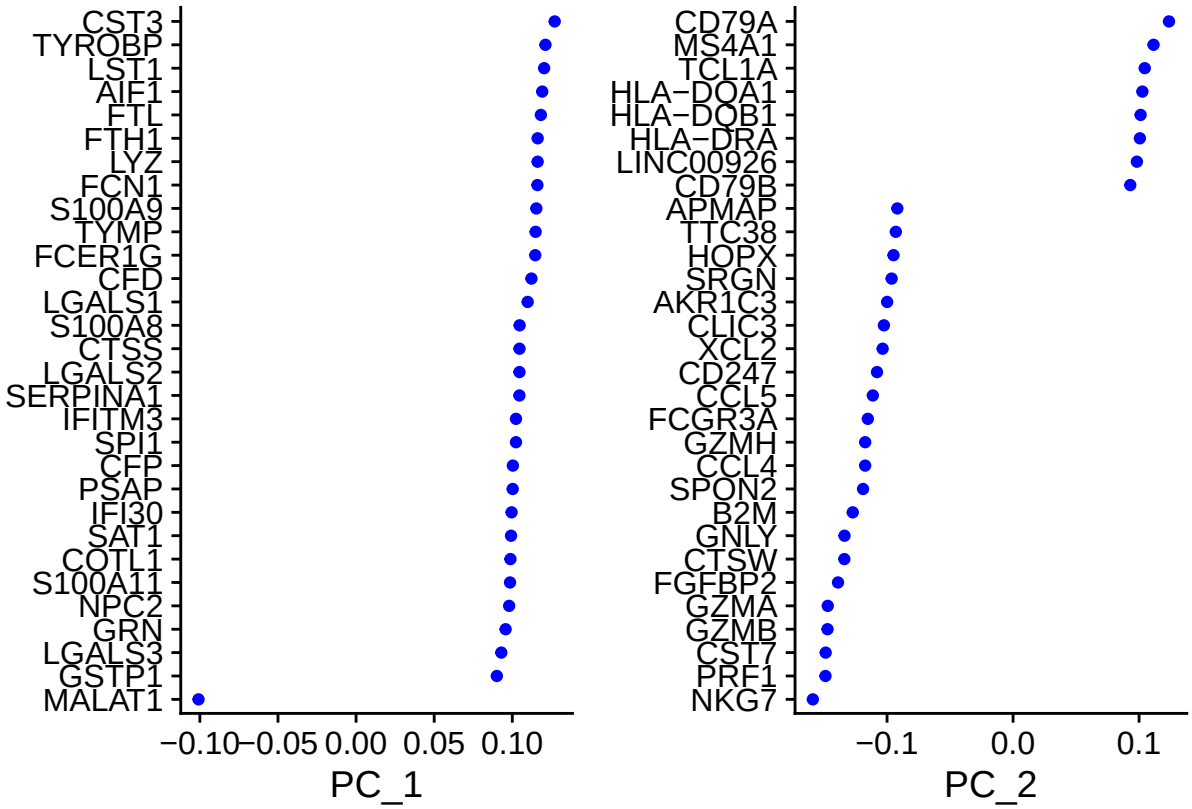

Figure 4: VizDimLoadings. De topgenen worden weergegeven die geassiocieerd worden met PC1 of PC2.

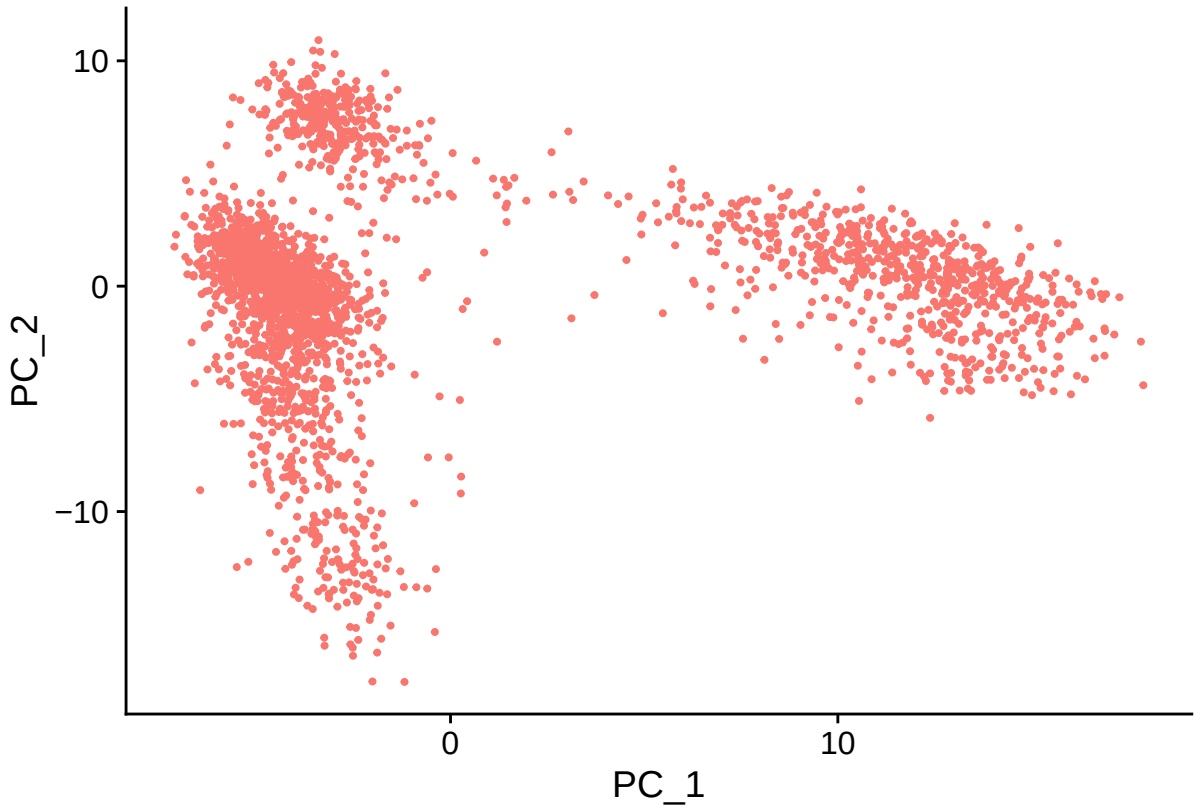

Figure 5: DimPlot, Op de x-as wordt PC1 geplot en op de y-as wordt PC2 geplot. Elke stip representeert een cel.

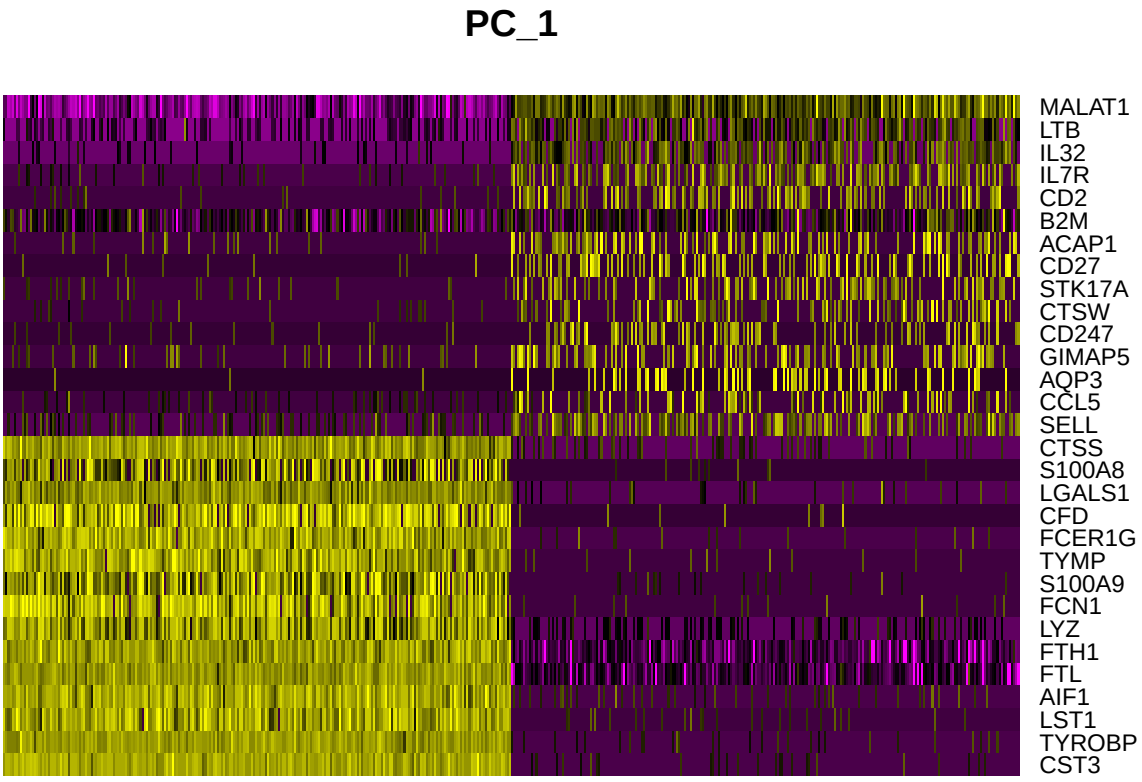

Figure 6: DimHeatmap, PC1 wordt weergegeven met als de verschillende genen in de rijen, en verschillende cellen in de kolommen. Paars is een lage PCA-score, geel is een hoge PCA-score. Clusters worden hiermee gevisualiseerd

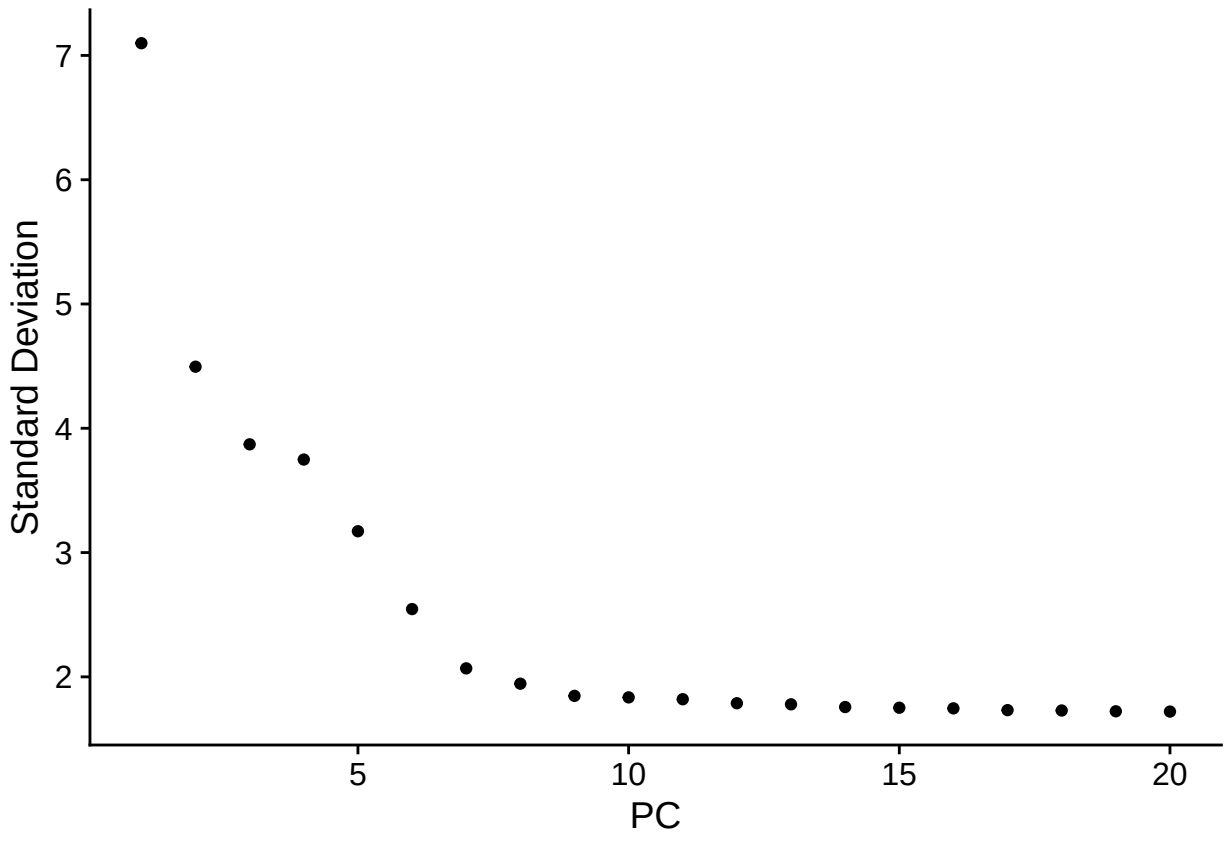

Figure 7: ElbowPlot: Op de x-as staan de PC's en op de y-as staan de standaard deviatie.

Clusteren

Nadat de juiste PC's zijn gekozen (1-10), verdeelt Seurat de cellen in groepen (clusters). Het programma kijkt hiervoor welke cellen het meest op elkaar lijken en verbindt die in een netwerk. Vervolgens zoekt het naar dicht bij elkaar liggende groepjes cellen binnen dat netwerk. Met de resolutie-instelling kun je bepalen of je grotere, bredere groepen of juist kleinere, gedetailleerdere groepen krijgt. In dit geval is er gekozen voor een resolutie van 0.5.

Niet-lineaire dimensionale reductie uitvoeren (UMAP/tSNE)

Seurat kan technieken zoals UMAP en tSNE gebruiken om de data overzichtelijk weer te geven in twee dimensies. Deze methoden plaatsen cellen die op elkaar lijken dicht bij elkaar, zodat de eerder gevonden clusters ook zichtbaar worden in de plot. Het is een handig hulpmiddel om patronen in de data te verkennen. Het blijft echter een visualisatie: lokale relaties (cellen die erg op elkaar lijken) worden goed weergegeven, maar de grotere, globale structuur is minder betrouwbaar. Gebruik UMAP of tSNE daarom vooral om een indruk van de data te krijgen, maar niet om er alleen op basis van de plot biologische conclusies te trekken.

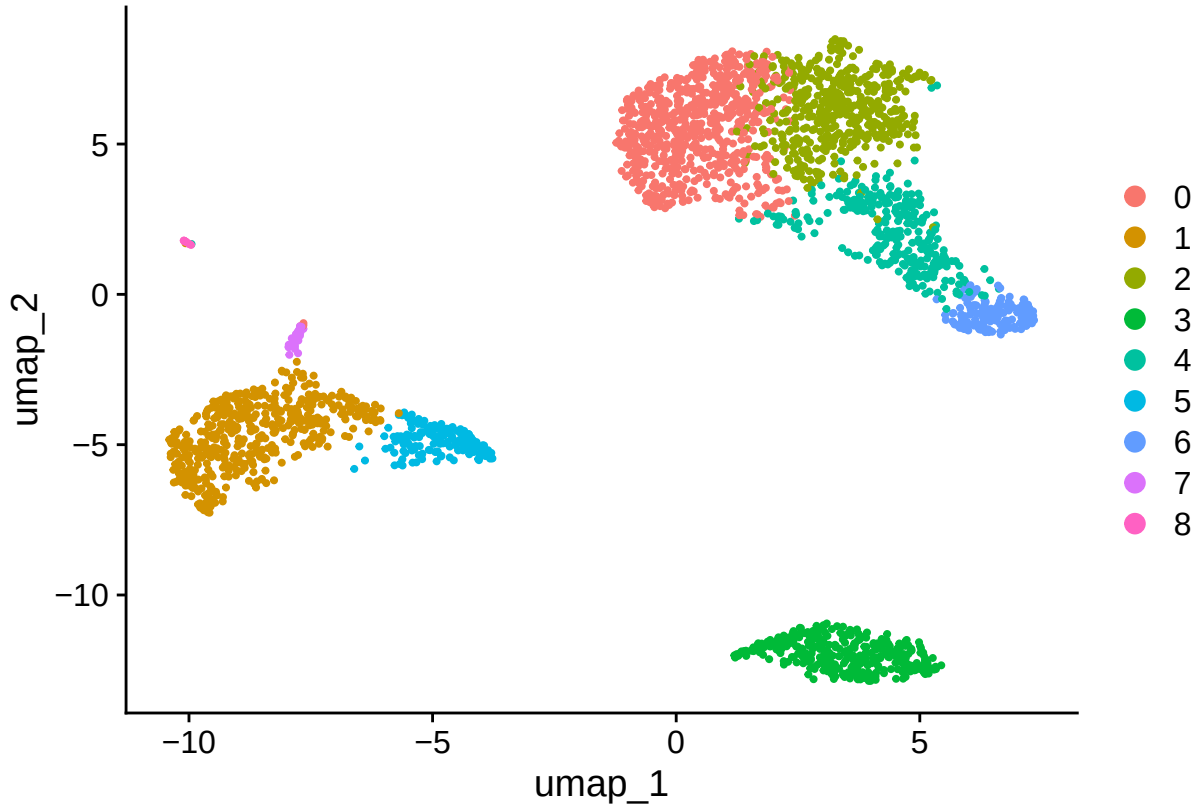

Figure 8: DimPlot: Op de x-as wordt umap_1 geplot en op de y-as wordt umap_2 geplot. Elke stip representeert een cel. De kleur geeft aan bij welk cluster deze cel hoort.

Het vinden van differentieel tot expressie gebrachte kenmerken (clusterbiomarkers)

Vervolgens kunnen er voor elke cluster genen gezocht worden die duidelijk meer (of minder) tot expressie komen dan in andere cellen. Dit helpt om te bepalen welke genen een cluster kenmerken.

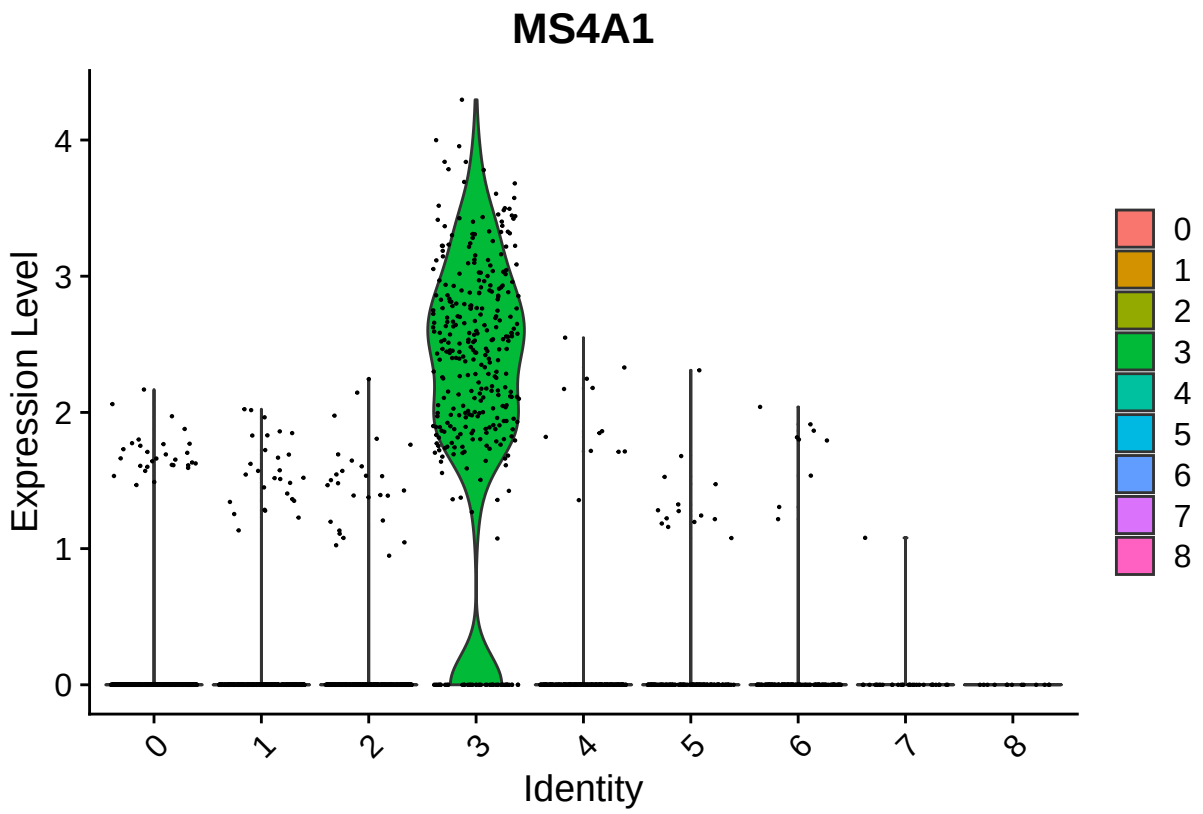

Figure 9: ViolinPlot: Het expressielevel van het gen MS4A1 wordt per cluster weergegeven. In cluster 3 komt dit gen het meest tot expressie.

Toewijzen van celtype-identiteit aan clusters

Als laatste stap worden er bekende celtypen gekoppeld aan de clusters.

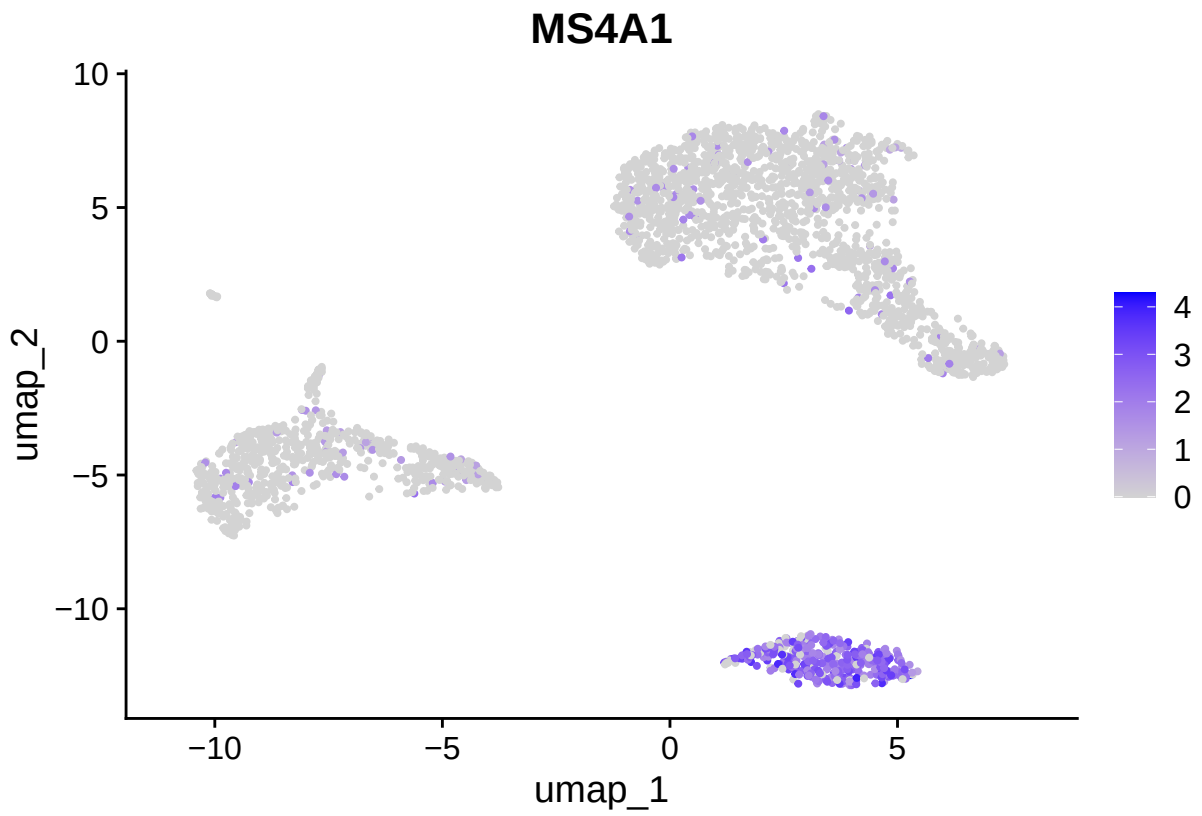

Figure 10: FeaturePlot:Op de x-as staat umap_1 en op de y-as staat umap_2. De expressie van MS4A1 wordt aangegeven in het paars.

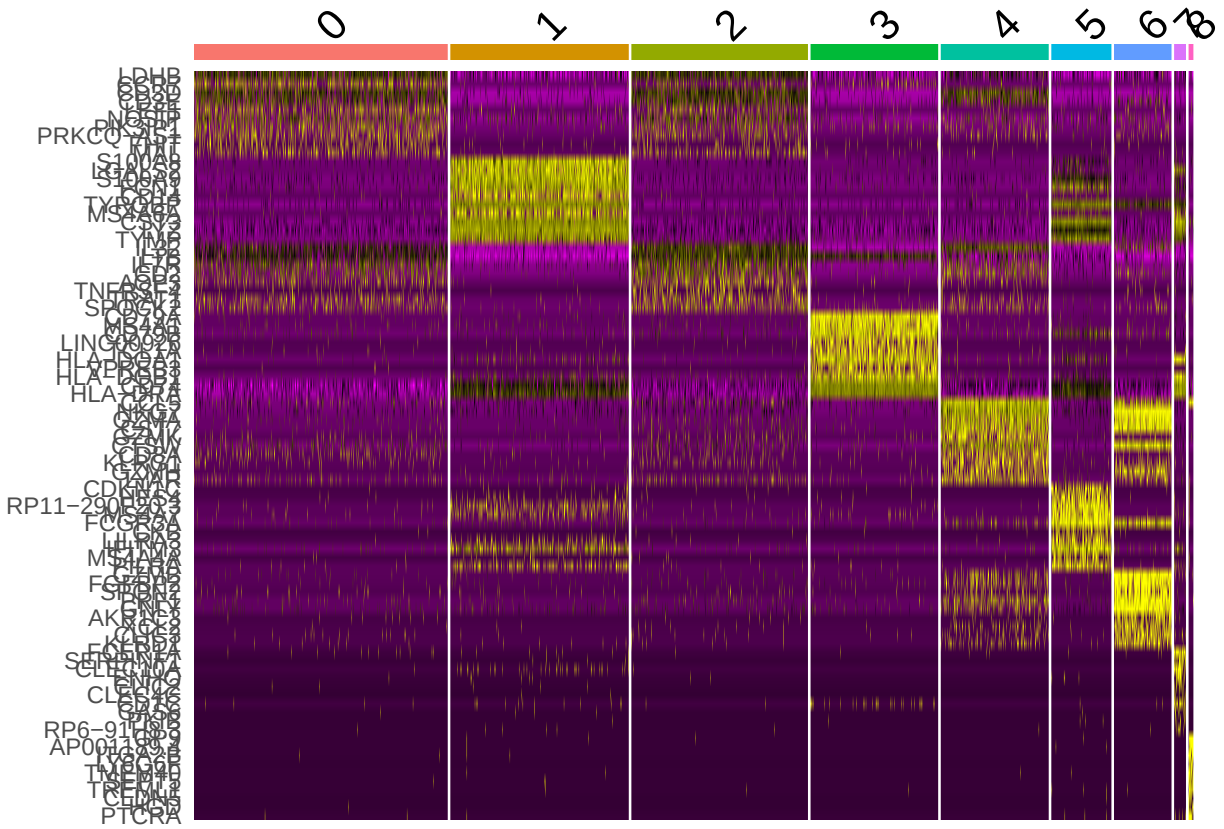

Figure 11: DoHeatMap: Op de x-as worden de clusters weergegeven en op de y-as worden de verschillende markers weergegeven. De 10 belangrijkste markers worden per cluster weergegeven. De gele kleur geeft een hoge expressie van het gen aan en de paarse kleur een lage expressie.

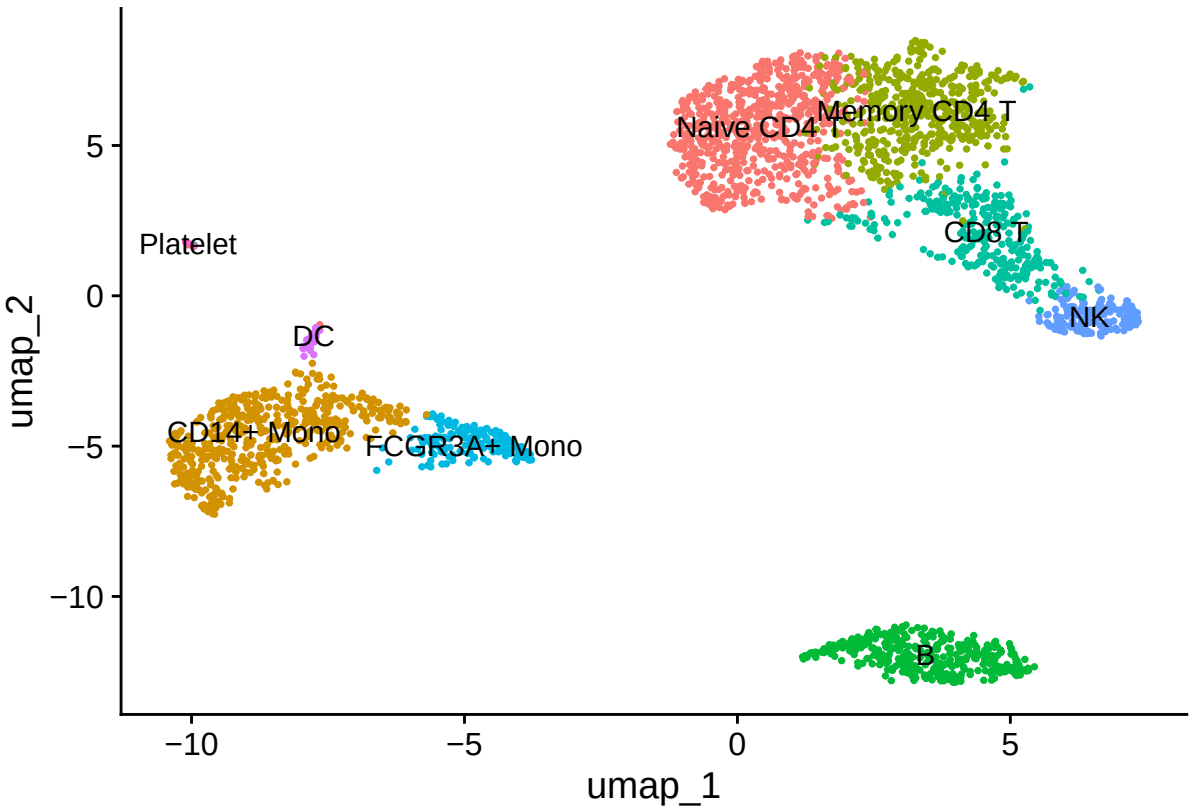

Figure 12: DimPlot: Op de x-as staat umap_1 en op de y-as staat umap_2. De clusters zijn aangegeven met verschillende kleuren en in de clusters staan de celtype namen van de desbetreffende cluster.