

Seurat pipeline

Anne Brussaard

2025-05-30

1 Seurat analyse op dataset E8.5

1.1 Deelvragen en flowschema

In dit bestand zullen de volgende deelvragen behandeld worden. 1. Wat is de kwaliteit van de data? 2. Hoeveel PCA's worden meegenomen in verdere analyse? 3. Welke clusters kunnen gebruikt worden voor de analyse met BRIE2.

Om deze vragen te kunnen beantwoorden zal de volgende workflow worden gevuld. 1. Data laden 2. Seurat object maken 3. Kwaliteits check data 4. Normalisatie data 5. PCA analyse 6. Clustering 7. UMAP 8. Selectie clusters

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()

1.2 Analyse

Stap 1.1: Eerst worden de packages geladen. De packages dplyr, ggplot2, pheatmap, tidyr, RColorBrewer, ggrepel, Seurat, Matrix, patchwork en here worden geladen.

Stap 1.2: Data wordt geladen. Het dataset E8.5 wordt gebruikt. Dit data set heeft de meeste data en in deze fase zit de meeste ontwikkeling van de embryo's waarbij veel alternatieve splicing betrokken is. De data is verkregen van de Hu server, de orginele data is te vinden via GEO number GSE176588.

Stap 1.3: MTX files worden geladen. MTX files zijn tekst bestanden waarin de data compact wordt opgeslagen.

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

Het percentage mitochondriale expressie wordt handmatig toegevoegd aan het seurat object. Deze data wordt los aangeleverd maar is wel essensieel voor het bepalen van de kwaliteit van de data.

1.2.1 Deelvraag: Wat is de kwaliteit van de data?

Stap 3.1: Kwaliteit check met visualisatie in 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 data ligt de mitochondriale expressie voor de meeste cellen bij 5%. De cellen die daar boven liggen worden dus als minder betrouwbaar gezien. Het hele dataset van tijdstip E8.5 wordt weergegeven.

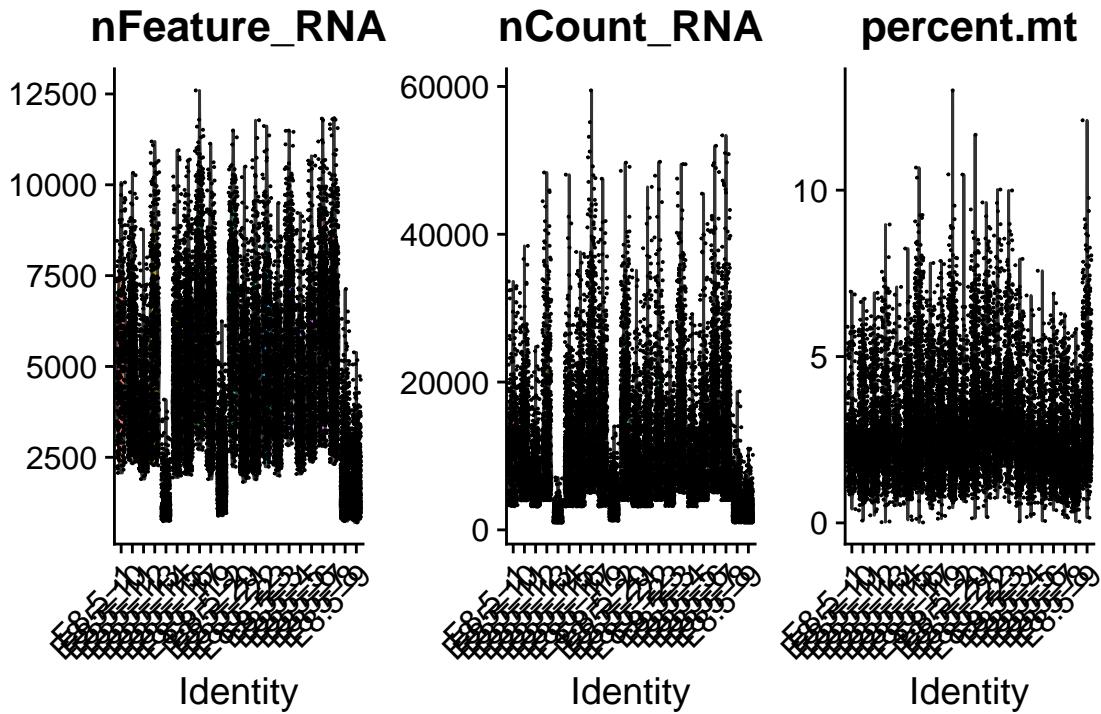


Figure 1: Dataset E8.5 in Violin Plots

In deze violin plot is het volgende te zien. Op de x as zijn de verschillende metingen weergeven die plaats moeten vinden om alle cellen van tijdspunt E8.5 te meten. op de Y-as zijn bij nfeature het aantal unieke genen te zien, bij nCount het aantal moleculen in de cel en bij percent mt het aantal mitochondriale expressie.

Stap 3.2: Visualisatie in Shatterplot. 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. De correlatie tussen nFeature en percentage.mt wordt hierin niet meegenomen omdat een hoog percentage MT kan samengaan met een zowel hoog of laag nFeature terwijl slechte kwaliteit vaak een hoge MT en een lage nCount heeft. Daarom is de correlatie van nFeature met percentage MT minder informatief over de kwaliteit.

1.2.2 Deel conclusie

uit de twee scatterplots is te zien dat de correlatie tussen nFeature_RNA en nCount_RNA 0,96 is. Dit betekend een zeer sterke positieve correlatie tussen het aantal genen en totaal aantal moleculen. Dit betekend goede kwaliteit van de cellen want de meeste cellen met veel RNA en inhoud brengen veel unieke genen tot expressie. Dit is wat de verwachting is bij goede kwaliteit single cell data. Voor de tweede scatterplot is een correlatie tussen percent.mt en nCount van -0,06 te zien. Dit betekend een licht negatieve correlatie tussen het percentage mitochondriale expressie en totaal aantal moleculen. Er is dus geen sterke relatie tussen de twee. Dit wijst op goede kwaliteit van de cellen omdat er niet veel cellen zijn die voornamelijk mitochondriale expressie hebben. Hieruit kunnen we concluderen dat de data die gebruikt wordt afkomstig is van goede kwaliteit cellen.

Stap 3.3: Filteren van de data. Op basis van de eerder uitgevoerde filter analyse is gekozen voor de volgende filtering nFeature_RNA <2000 >12000 zo worden de lege en dubbele metingen niet meegenomen. percent.mt <5 zo worden cellen die dood zijn of stress hebben niet meegenomen. nCount wordt hierin niet meegenomen

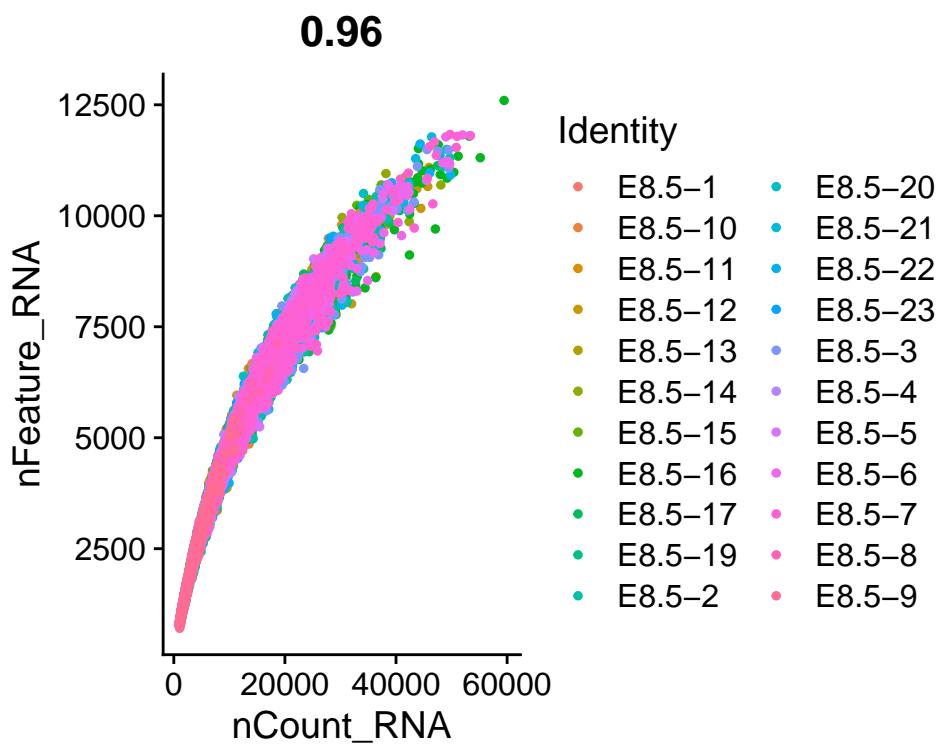


Figure 2: Correlatie analyse

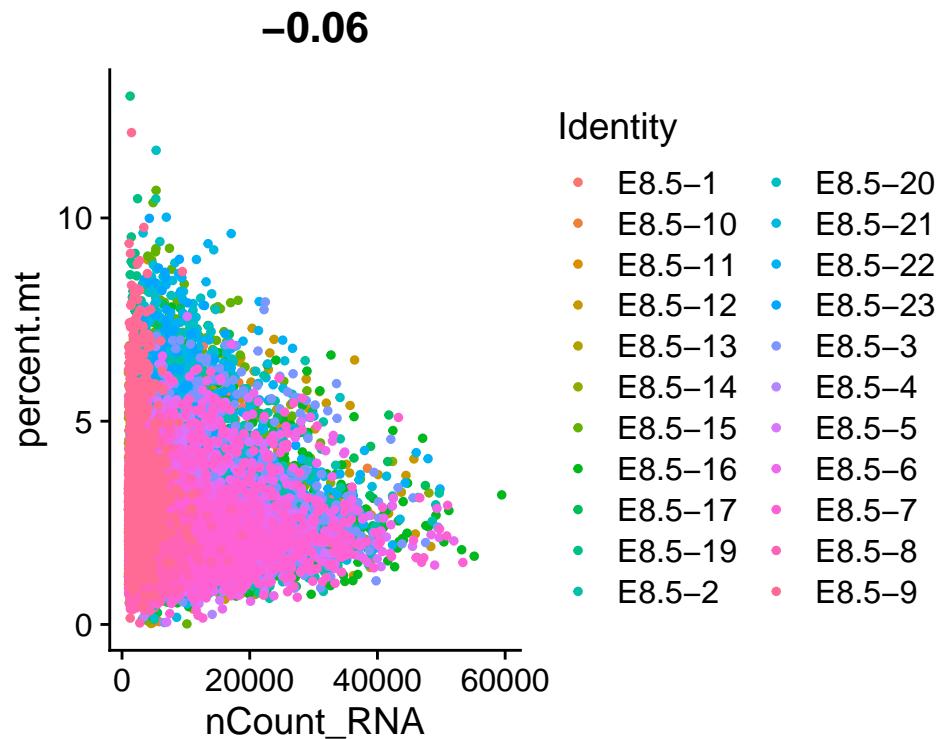


Figure 3: Correlatie analyse

omdat deze alleen een indicatie geeft over de kwaliteit maar niet meegenomen in verder analyse omdat het geen informatie bevat over genexpressie.

Stap 4.1: Normalisatie van de data. 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 hetzelfde 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 zijn met genen met lage counts. Ook helpt het bij het normaal verdelen van de data voor PCA analyse.

Stap 4.2: Variabele genen vinden. Deze stap vindt plaats om allen de genen te gebruiken die in expressie sterk van elkaar verschillen. Ook wordt op deze manier ruis van genen die constant een lage expressie hebben verwijderd. Dit zorgt voor een efficiëntere analyse en kan er een beter beeld gevormd worden van welke genexpressie patronen er zijn.

Stap 4.3: Variable genen plotten. Deze variabele genen worden vervolgens weergegeven in een plot om een beeld te krijgen van hoeveel genen in de data variabel zijn en wat de top 10 genen zijn. Hierin worden zowel 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.

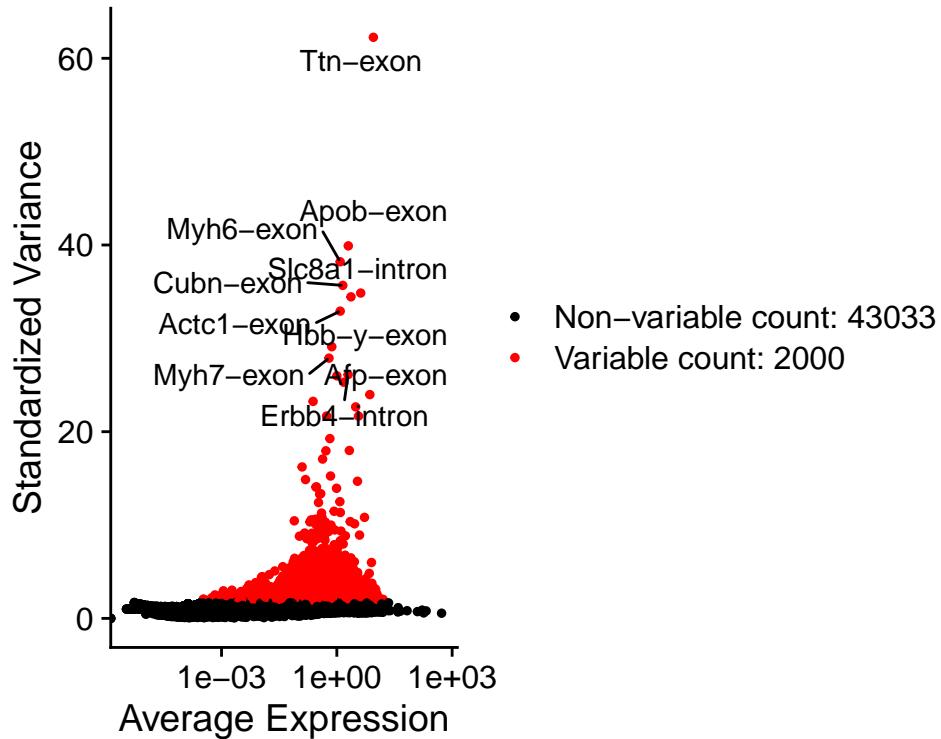


Figure 4: Variabele genen

Stap 4.4: Schalen van genen. Dit is nodig om te zorgen dat alle genen een vergelijkbaar expressie niveau. Zo wordt voorkomen dat de PCA beïnvloed wordt door genen die hoge expressie hebben. Er wordt een lijst gemaakt van alle gebruikte genen, deze krijgen allemaal een gemiddelde en standaard deviatie van de expressie

1.2.3 Deelvraag: Hoeveel PCAs worden meegenomen in verdere analyse?

Stap 5.1: PCA analyse uitvoeren Single cell data bevat namelijk duizende genen per cel (dimensies) PCA analyse reduceert dit naar kleinere hoofd dimenties. De belangrijkste verschillen tussen cellen worden samengevat waardoor verder analyse efficenter kan verlopen.

Stap 5.2: PCA analyse plotten. Deze PCA analyse wordt gevisualiseerd in een Elbowplot. Op de X-as worden de verschillende PC weergeven en op de Y-as de standaard deviatie. Hoe hoger de standaard deviatie hoe groot het verschil is in PC, dit zijn ook de belangrijkste PC om mee te nemen in verdere analyse.

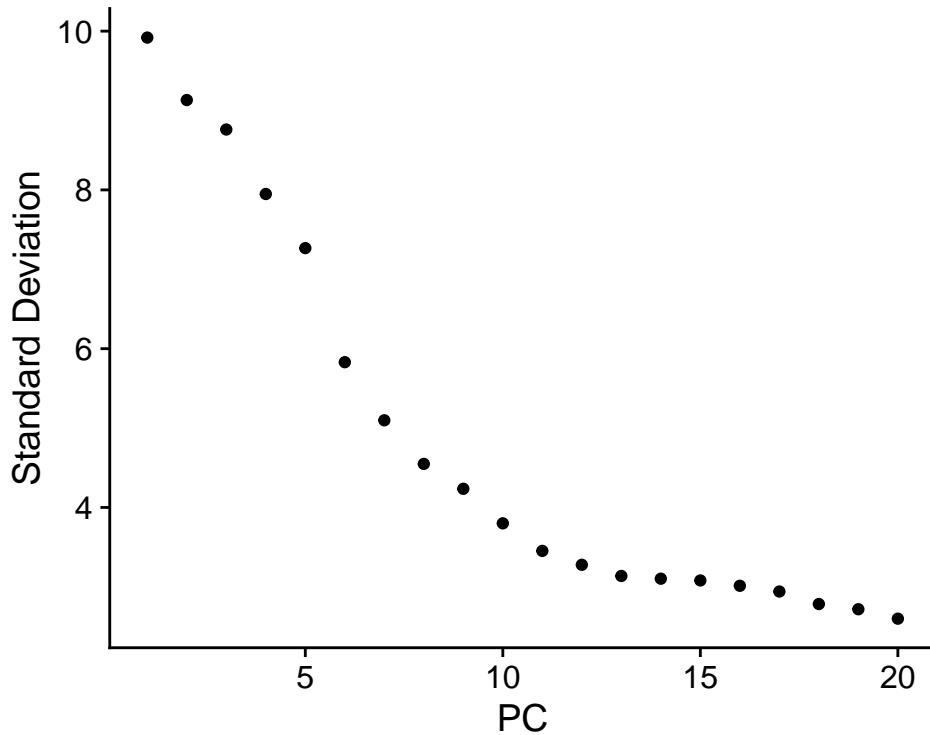


Figure 5: PC analyse

1.2.4 Deel conclusie:

Vanuit de eerder gevolgde tutorial is geadviseerd om 10 PC's mee te nemen in cluster analyse voor betrouwbaar resultaat. Ook is te zien in de elbowplot dat de "elbow" afneemt na 10 PC's. Daarom zullen de eerste 10 PC's worden meegenomen in de verder cluster analyse.

Stap 6.1: Clusteren Er zal cluster analyse worden uitgevoerd op de eerste 10 PC's. Dit houdt in dat er op basis van gen expressie patronen groepen gemaakt worden.

Stap 7.1: UMAP clusters De manier waarop deze clusters worden weergegeven is een UMAP.

In de UMAP is iedere punt 1 cel, de kleur is het cluster en de afstand tussen de cellen is hoeveel verschil er is in genexpressie patroon. Dit betekend dat hoe dichter de cellen bij elkaar liggen hoe meer de expressie op elkaar lijkt. Toch is er dan wel verschil waardoor ze behoren tot een ander cluster. Daarom worden vaak clusters gekozen voor verdere analyse die dicht bij elkaar liggen. In deze UMAP worden 17 verschillende clusters weergeven.

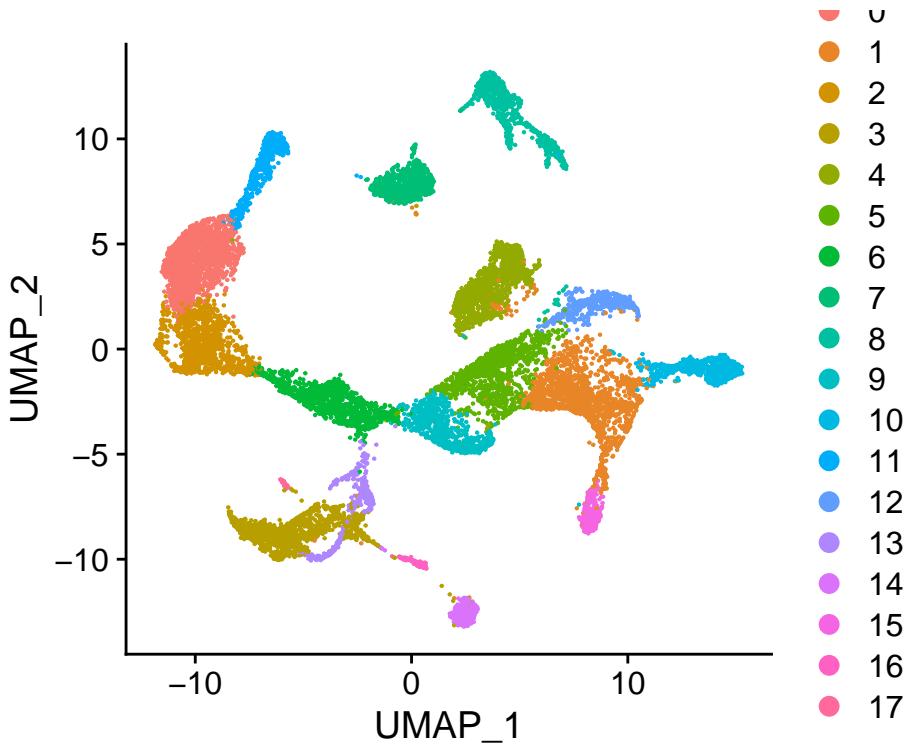


Figure 6: Clustering E8.5 data

1.2.5 Deelvraag: Welke clusters worden geselecteerd voor analyse in BRIE2?

In eerste instantie is gekozen voor cluster 5 en 9 omdat deze clusters dicht bij elkaar liggen, er kan dan gekeken worden of er verschil is in alternatieve splicing. Maar er zal dan ook eerst gekeken worden of er in deze clusters dezelfde genen voorkomen die in alternatieve splicing van elkaar kunnen verschillen. Om te kunnen bepalen welke clusters meegenomen worden in analyse met BRIE2 zal nu verder gekeken worden naar de biomarkers.

Stap 8.1. Biomarkers markers vinden in alle clusters en alleen de positieve worden gerapporteerd. Deze biomarkers kunnen helpen met het vinden van verschillen in de clusters.

Stap 8.2 Top 15 genen van cluster 5 en 9 selecteren. De top 10 wordt geselecteerd om te kijken of er overlap is tussen de cluster genen.

Stap 8.3: Top 15 genen van cluster 5 en 9 plotten

In de Dotplot afbeelding zijn op de X-as de genen weergegeven, op de Y-as zijn de clusters weergegeven. De grote van het bolletje staat voor in hoeveel procent van de cellen in dat cluster dat gen tot expressie komt en de kleur intensiteit voor de hoeveelheid expressie van dit gen van het totaal.

1.2.6 Deelconclusie:

Op basis van deze resultaten kan geconcludeerd worden dat cluster 5 en 9 geschikt zijn voor verdere analyse. Te zien is dat een aantal genen in beide clusters voorkomen. Dit geeft de mogelijkheid om te kijken naar verschillende isovormen of splicing events van deze genen in beide clusters.

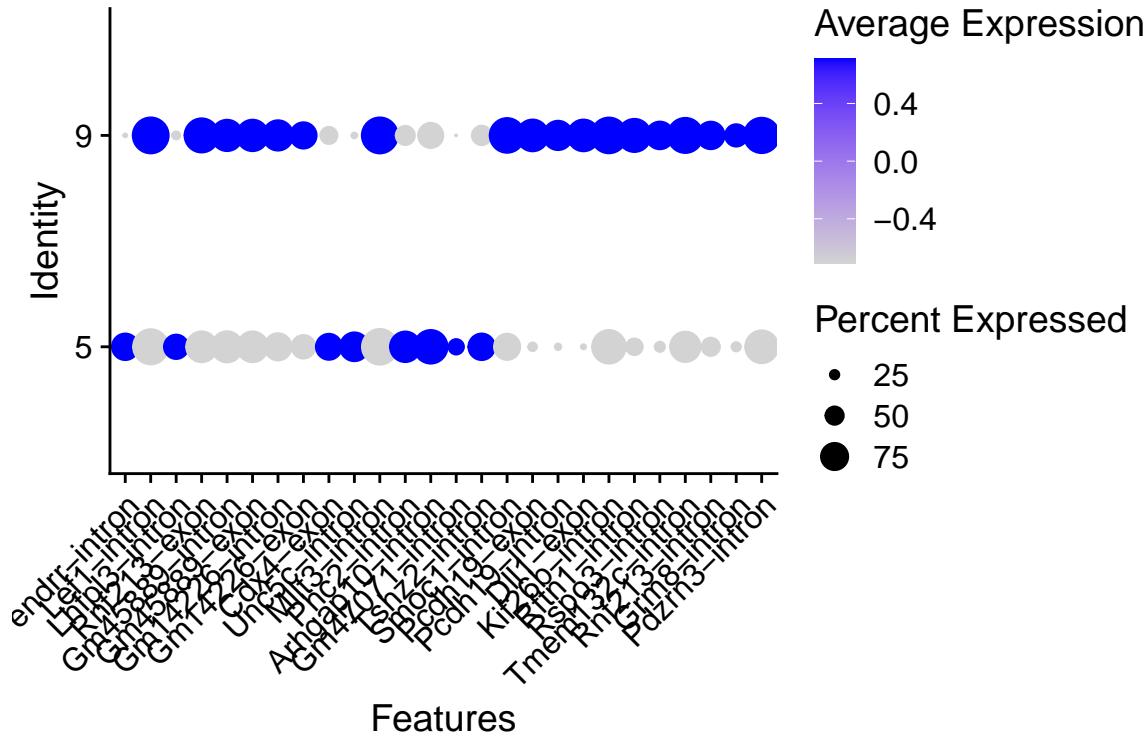


Figure 7: Top 15 genen cluster 5 en 9

1.3 Conclusie

Op basis van deze experimenten is geconstateerd dat de data van goede kwaliteit is en gebruikt kan worden voor analyse. Verder is bepaald dat er 10 PC patronen worden meegenomen voor het vormen van de clusters. Ook is bepaald dat cluster 5 en 9 gebruikt zullen worden voor verdere analyse in BRIE2,