

# Seurat\_filtering.Rmd

Anne Brussaard

2025-06-14

## 1 Filterings opties bekijken in Seurat

Na het uitvoeren van de tutorial met seurat wordt nu gekeken naar de effecten van filtering in Seurat van dataset E8.5 (GEO number GSE176588). Deze data is verkregen met VASA sequencing.

### 1.1 info

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 Deelvraag

Wat zijn de effecten van filtering van nFeature en percent.mt op de clustering van de data gevisualiseerd in een UMAP?

#### 1.2.1 Analyse

Stap 1.1: Eerst wordt gecontroleerd of alle packages zijn geinstalleerd, als dit niet zo is wordt dit gedaan.

Stap 1.2: Packages laden. De packages dplyr, ggplot2, pheatmap, tidyr, RColorBrewer, ggrepel, Seurat, Matrix, patchwork en here worden geladen. Deze packages worden met de library functie toegepast in analyse.

Stap 2: Data wordt geladen. De data wordt als csv bestand ingeladen en als object bewaard. De data is afkomstig van muis embryo's op dag E8.5.

Stap 3: MTX files worden geladen. Het object wordt samen met de matrix tabel ingeladen.

Stap 4: Seurat object maken. Met de packages seurat wordt een object gemaakt waardoor de data gebruikt kan worden voor preprocessing en data analyse.

Het percentage mitochondriale expressie wordt handmatig toegevoegd aan het seurat object omdat deze nodig is voor de kwaliteits check.

Stap 5: Kwaliteit check met visualisatie in violin plot van nFeature\_RNA (unieke genen per cel), nCount\_RNA (totaal aantal moleculen), en percent.mt (mitochondriale expressie). Dit wordt geplot in een violin plot zodat kan worden gekeken hoe de data is verdeeld. Een laag aantal unieke genen of moleculen kan namelijk wijzen op lege of slechte kwaliteit van de cellen. Een te hoog aantal kan juist wijzen op dubbele metingen. Veder is het mitochondriaal percentage belangrijk omdat een hoog percentage hiervan wijst op dode of slechte kwaliteit cellen. In deze violin plots wordt weergegeven hoe de data verdeeld is. Om te bepalen welke data er meegenomen wordt in verdere analyse worden er verschillende manieren van filteren

getest. Op de Y-as wordt het aantal genen, moleculen en percentage MT weergeven. Op de X-as staan de verschillende metingen die plaats hebben gevonden. Vanwege de grote hoeveelheid cellen zijn deze in verscheden metingen gesequenced.

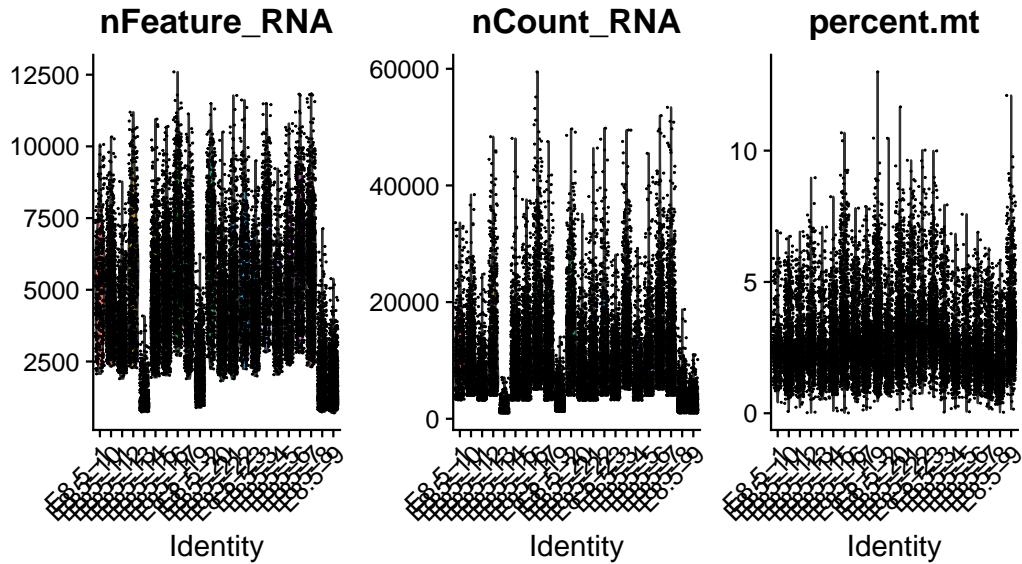


Figure 1: Visualisatie kwaliteit dataset E8.5

Stap 6. De data kan vervolgens op verschillende manieren worden gefilterd nu is gekozen voor de de opties strict, mild en geen filter. Dit is gedaan op basis van de violin plot uit stap 5, er zijn namelijk niet hele grote verschillen in de data. Er is gekozen voor strikt filtering, milde filtering en geen filtering. Dit wordt gedaan te kijken welke filtering het beste is voor de data. De verschil in filtering zit op de nFeature en percent.mt. Er is voor gekozen om de percent.mt niet te testen, dit omdat 5% vanuit de literatuur wordt aanreikt als cutoff waarden voor muizen cellen. De nCount wordt ook niet meegenomen in deze filter stap omdat deze data geen genexpressie informatie bevat en niet verder wordt gebruikt in de berekening. De filters zijn: Seurat strict = nFeature >2500 <10000 & percent.mt <5% Seurat mild = nFeature >2000 <12000 & percent.mt <5% Seurat no filter = geen filter

Stap 7. Opnieuw worden de violin plots gemaakt om het effect van de filter stappen te zien op de data. zie figuur 2. We zien nu het effect van de filtering op de data. De bovenste rij is het strikte filter en we zien daar een duidelijk lijn welke cellen wel of niet wordt meegenomen. Bij de tweede rij (het milde filter) is deze lijn een stuk minder strak voor nFeature en nCount. Voor de laatste rij is geen filter gebruikt en bevat dus alle cellen. Zie figuur 2 Verschillende filterings opties.

Stap 8. Vervolgens wordt een functie geschreven om de de verder seurat analyse uit te voeren op de verschillende filter mogelijkheden. Hierin wordt de data genormaliseerd, er variabele genen geselecteerd, de data geschaald, een pca analyse uitgevoerd er clusters gemaakt en deze clusters worden uitgezet in UMAP's. De stappen van de functie zullen niet verder worden gevisualiseerd dit omdat de focus van dit script ligt op het vergelijken van de filter stappen, zo zijn voor alle 3 de datasets 10 PC's meegenomen.

Stap 9. Deze functie wordt uitgevoerd op de 3 filter opties. Dit houdt in dat de data wordt genormaliseerd voor nFeature, er wordt een berekening uitgevoerd zodat alle cellen "dezelfde" hoeveelheid expressie hebben zodat ze vergelijkbaar zijn. Vervolgens worden de variabele genen geselecteerd, dat zijn genen die in expressie sterk van elkaar verschillen. Hierna volgt de schaling om te voorkomen dat hoge expressie niveaus invloed hebben op de PC analyse. Daarna volgt de PC analyse waarbij wordt gekeken welke soort expressie patronen er zijn. Hiervan worden er 10 geselcteerd. Daarop volgend is de clustering en wordt de UMAP gemaakt, er wordt gegroepeerd op basis van expressie patroon.

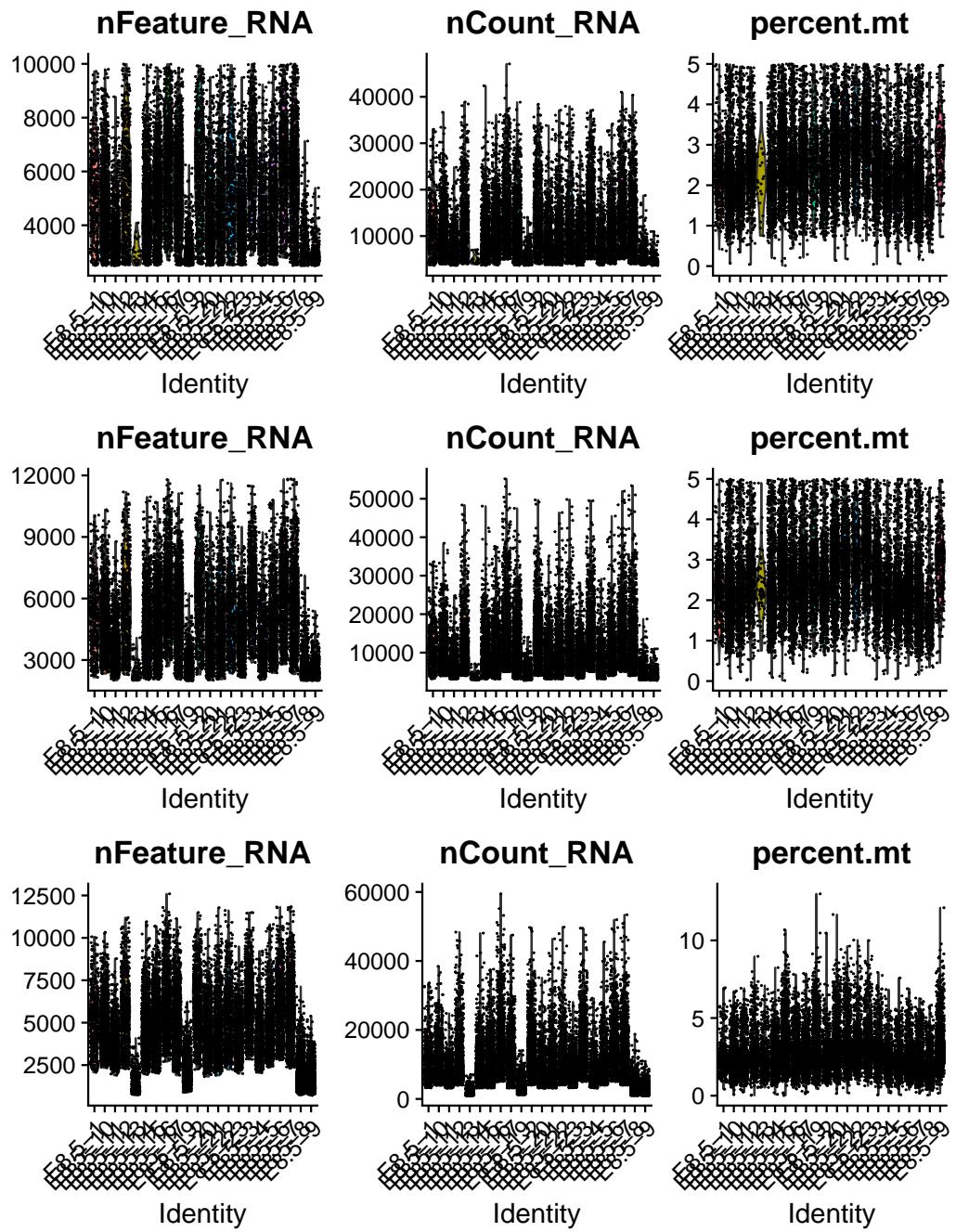


Figure 2: Verschillende filterings opties (strikt (rij 1), mild (rij 2) en geen filter(rij 3))

Stap 10. Van de 3 verschillende filter opties worden UMAPs gemaakt, om het verschil te laten zien. Zie figuren 3, 4 en 5.

Stap 11. De UMAPs weergeven. zie figuur 4 (strikt), figuur 5 (mild) en figuur 6 (geen filter)

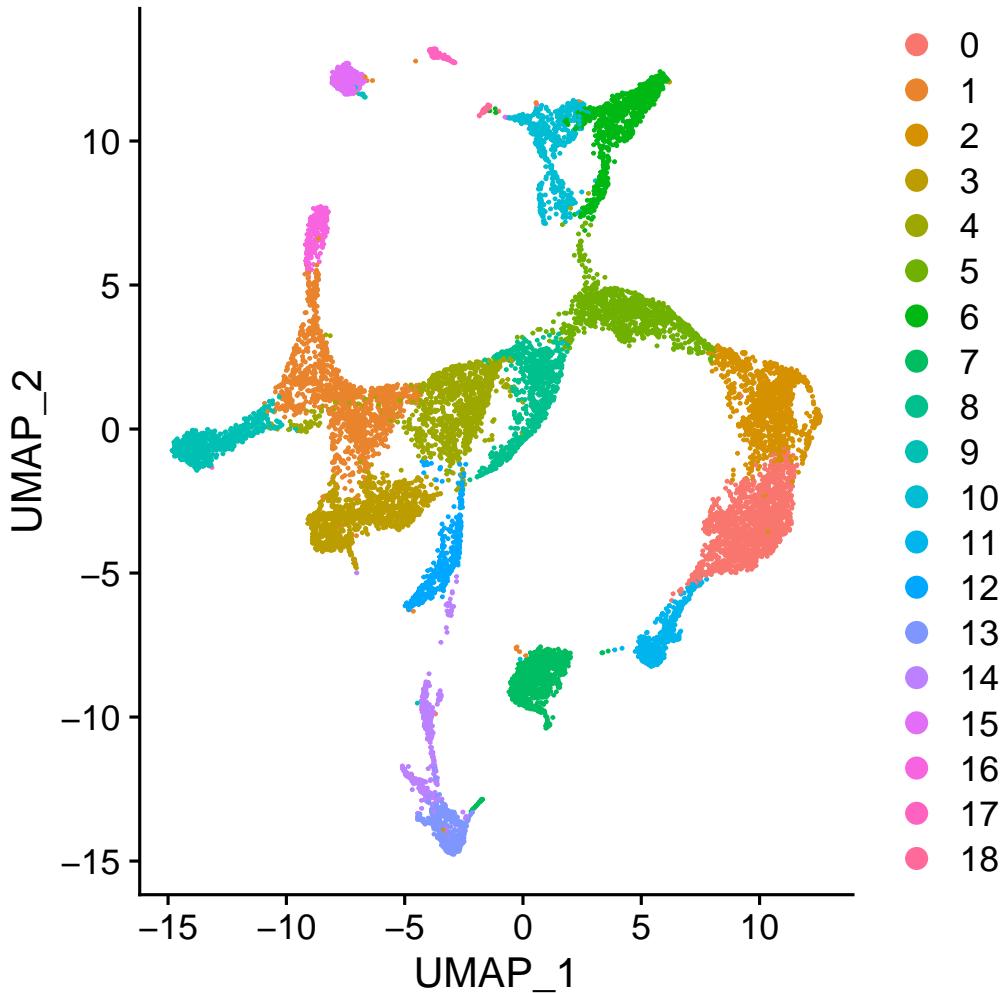


Figure 3: UMAP filter strikt

### 1.2.2 Conclusie

In de umaps is te zien dat de milde en strikte filter ongeveer dezelfde clusters bevatten. Toch is op basis van deze UMAPS is gekozen voor filter optie mild. Dit omdat met deze manier van filteren zo veel mogelijk data wordt meegenomen wat bijdraagt aan de betrouwbaarheid van de verdere analyse. Bij deze optie zijn er ook duidelijke clusters te zien in vergelijking met de geen filter optie. Ook kunnen door de strenge filtering misschien cellen worden weg gefilterd die zorgen voor interessante variatie. Verder is het doel van het onderzoek nu om een pipeline op te zetten voor BRIE2, daarom is de biologische relevantie minder prioriteit.

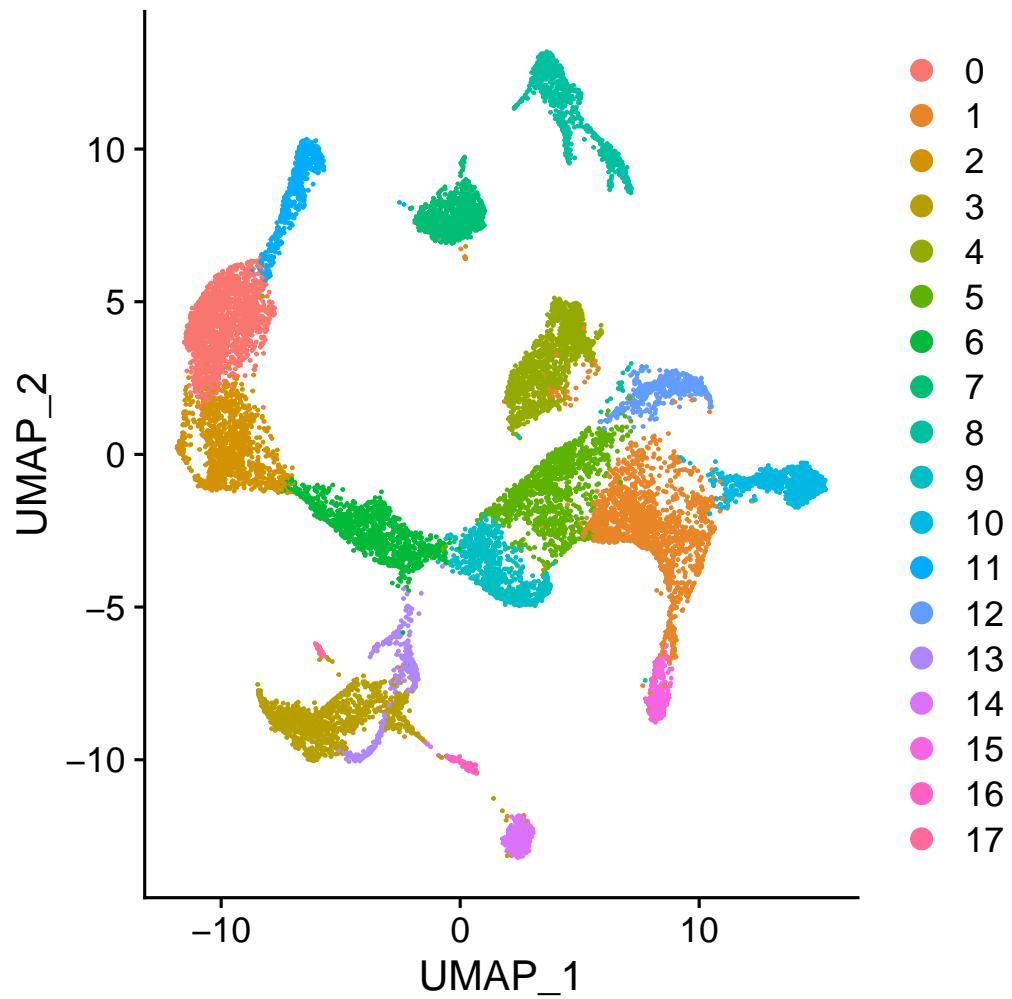


Figure 4: UMAP filter mild

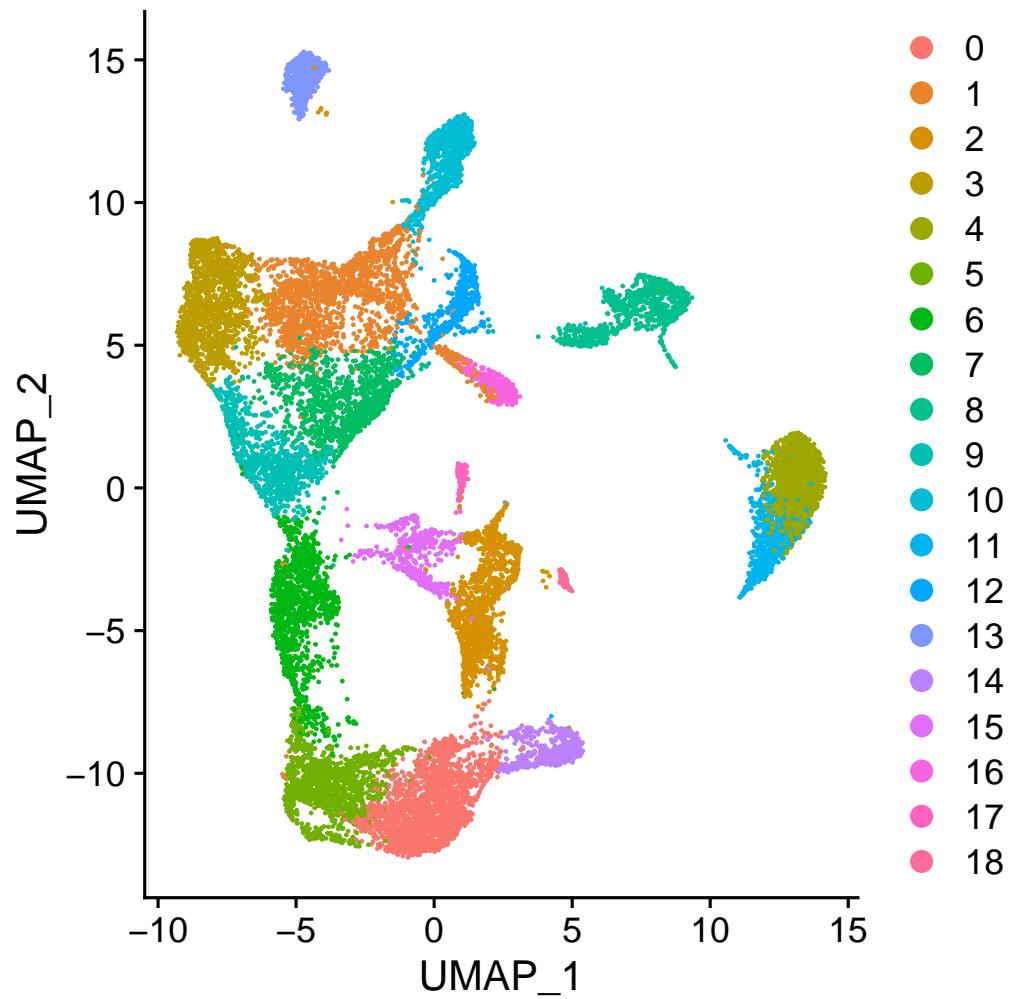


Figure 5: UMAP no filter

### 1.2.3 Discussie

Strengere filtering kan wel zorgen voor minder ruis wat kan zorgen voor een duidelijker biologisch verschil. Daarom zou na het ontwikkelen van de analyse pipeline van brie2 is ontwikkeld om zo andere clusters te kunnen vergelijken. Ook zou het nog een mogelijkheid zijn om naar een andere methode van visualisatie. Zo zou er voor specifieke genen het verschil in expressie worden weergeven in een violin plot. Om het hele dataset zichtbaar te maken zou ook nog een heatmap gebruikt kunnen worden, al kunnen onderlinge relaties tussen groepen hier wel minder goed in worden weergeven. Omdat nu gekozen is voor filter mild zou ik wel aanraden om na het opzetten van de pipeline deze opnieuw door te lopen met de strenger gefilterde data om te zien of er verschil is in uitkomst. Ook kan op deze manier de pipeline gecontroleerd worden.