

Data_8_5_analyse_1_QC

2025-11-24

Inleiding

Om meer te weten te komen over de ontwikkeling van de mens, is onderzoek naar embryonale ontwikkeling essentieel. Dit is een complex proces waarbij inzicht in de vele moleculaire mechanismen cruciaal is. De muis is een geschikt modelorganisme vanwege de evolutionaire verwantschap met de mens en het korte ontwikkelingsstadium: binnen drie weken ontwikkelt een embryo zich tot een zelfstandig levende jonge muis. Hierdoor leent de muis zich uitstekend voor onderzoek naar embryonale ontwikkeling.

Een belangrijk onderdeel van de genregulatie wordt gevormd door long non-coding RNA (lncRNA). lncRNA's zijn RNA-moleculen langer dan 200 basenparen die niet coderen voor eiwitten, maar wel verschillende regulerende functies vervullen binnen de cel. Ze zijn vaak celtype-specifiek en kunnen daardoor dienen als betrouwbare markers voor verschillende celtypen.

In dit onderzoek worden twee pipelines ontwikkeld om nieuwe lncRNA-markers te identificeren uit single-cell RNA-seq (scRNA-seq) data van embryonale ontwikkeling:

- 1: Een volledige-genen pipeline, waarin alle genen samen worden geclusterd en vervolgens lncRNA-markers worden onderzocht.
- 2: Een lncRNA-only pipeline, waarbij eerst alleen de lncRNA's worden geselecteerd, zodat lage-expressie lncRNA's minder snel worden weggefilterd en beter onderzocht kunnen worden.

De analyses worden uitgevoerd met Seurat. In deze RMarkdown wordt de ruwe data van pipeline 1 geanalyseerd. Eerst wordt de kwaliteit van de cellen beoordeeld aan de hand van verschillende QC-metrics. Op basis van deze resultaten wordt bepaald welke cellen geschikt zijn voor verdere analyses, zoals normalisatie, feature selectie en clustering.

Doe

Hoofdvraag: Kan een lncRNA-gerichte single-cell RNA-seq pipeline nieuwe, cell-type-specifieke markers identificeren tijdens de embryogenese?

Deelvraag: Wat is de kwaliteit van de data en hoeveel data moet ik eruit filteren voor betrouwbare resultaten?

Hypothese

Er wordt verwacht dat wanneer de kwaliteit van de data bekijken wordt er ook gefilterd kan worden voor betrouwbare resultaten.

Data analyse met Seurat

In Seurat worden de volgende stappen doorlopen - inladen van de data - QC van de data en selecteren van de cellen voor vervolganalyse

Het inladen van de data

De data wordt ingeladen als een object. De data uit het onderzoek is al gedownload op de server. Dit is te vinden onder het volgende pad: ‘/home/data/projecticum/splicing/data/’. Omdat het om hele grote bestanden gaat is er voor gekozen om de data niet nog in het eigen project te zetten. Dit zou onnodig veel ruimte in beslag nemen.

De ruwe data van het onderzoek is te vinden bij de GEO website onder het volgende nummer: GSE176588.

Pre-procesing en QC van de data

Om de kwaliteit van de ruwe data te bekijken, wordt er o.a. gekeken naar het percentage van mitochondrial RNA. Dit is een maat voor de levensvatbaarheid van de cellen. Hoe hoger het percentage mitochondrial RNA, hoe slechter de levensvatbaarheid van de cel is. Bij de nFeature plot wordt er het totaal aantal features per cel weergeven. Bij de nCount_RNA plot wordt er gekeken naar het totaal aantal RNA moleculen per cel. Bij de percent.mt plot wordt er gekeken wat het percentage mitochondriaal RNA per cel is.

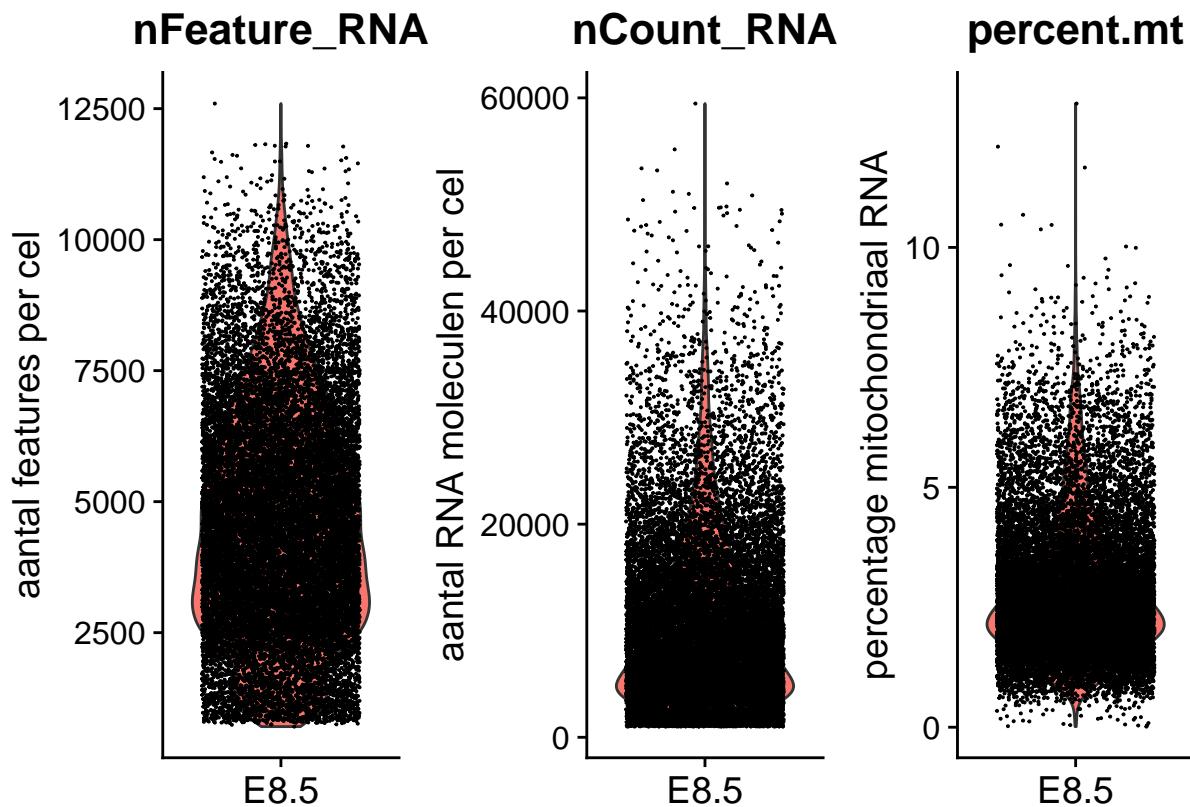


Figure 1: QC ruwe data visualisatie in vioplot van nFeature_RNA, nCount_RNA en percent.mt

In afbeelding 1 is te zien dat veel cellen veel features hebben. Dit kan te maken hebben met het celtype. cellen uit de embryomale ontwikkeling zijn actief aan het ontwikkelen en produceren veel verschillende genen. Het percentage mitochondriaal RNA is niet heel hoog bij de meeste cellen. Een hoog percentage kan duiden op slechte kwaliteit van de cellen. Het grootste gedeelte van de data heeft een goede kwaliteit. Cellen met een hoger percentage mitochondriaal RNA dan 5 % worden eruit gefilterd. Om wat meer te kunnen zeggen over de kwaliteit van de cellen, kunnen de verschillende waarden ook tegenover elkaar gezet worden in een

FeaturesScatter plot. Hieronder zijn twee van deze plotten zichtbaar waarbij de nCount_RNA tegenover de percent.mt is gezet. In de tweede grafiek zijn de nCount_RNA gezet tegenover de nFeature_RNA.

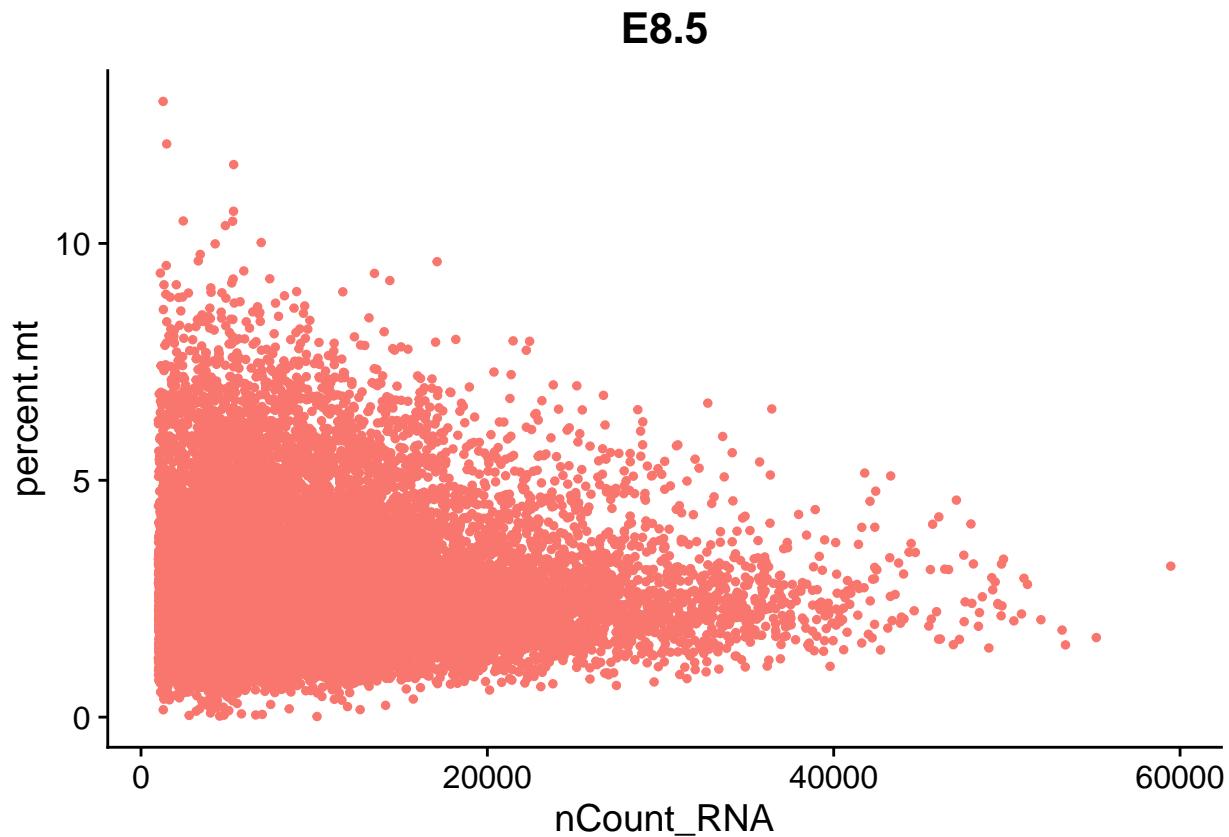


Figure 2: Scatterplot relatie tussen nCount_RNA en percent.mt

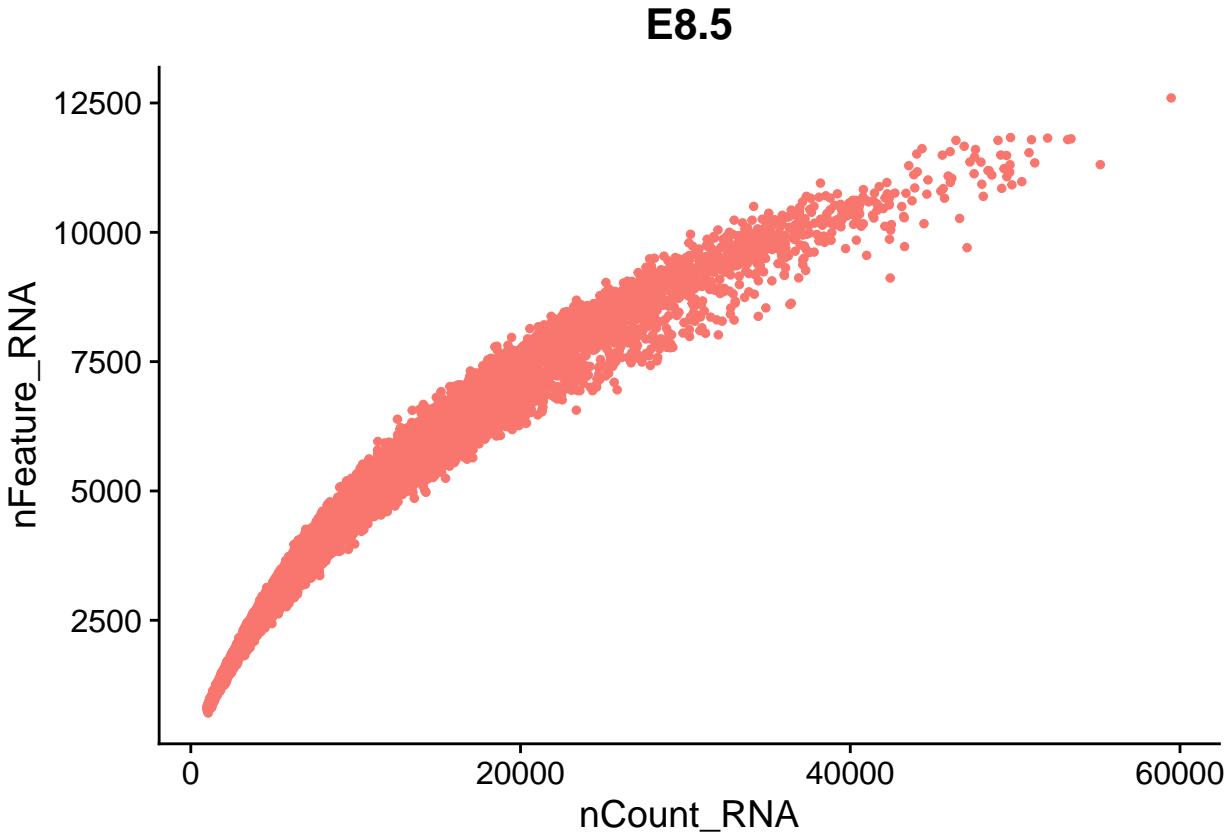


Figure 3: Scatterplot relatie tussen nCount_RNA en nFeature_RNA

Bij figuur 2 waarbij nCount_RNA tegenover de percent.mt is gezet, wordt er gekeken of er een correlatie is tussen het aantal RNA moleculen in een cel en de hoeveelheid mitochondriaal RNA. Wanneer er cellen zijn met veel RNA moleculen en een hoog percentage mitochondriaal RNA, kan dit betekenen dat er meercellige druppels zijn geweest bij de isolatie. Dit kan dus zorgen voor verhoogde waardes.

Bij figuur 3 waarbij nCount_RNA tegenover de nFeature_RNA wordt er verwacht dat wanneer de count toeneemt de nFeature ook toeneemt. Dit moet lijken op een rechte lijn. Wanneer de count hoog is maar de features laag, dan kan dit betekenen dat weinig cellen veel gesequenced zijn. Wanneer de count laag is maar de features hoog kan het betekenen dat er veel features zijn, maar er niet diep genoeg gesequenced is.

Conclusie

Bij de data in figuur 2 is er gekeken of er een correlatie is tussen nCount_RNA tegenover de percent.mt. Wanneer er cellen zijn met veel RNA moleculen en een hoog percentage mitochondriaal RNA, kan dit betekenen dat er meercellige druppels zijn geweest bij de isolatie. Dit is in figuur 2 niet te zien. Er zijn wel cellen te zien met een lage count met wel hogere mitochondriale waarde. Dit kan wijzen op slechte kwaliteit van de cellen. Dit is ook te verwachten en wordt er in de volgende stappen uitgefilterd.

In figuur 3 wordt er verwacht nCount_RNA tegenover de nFeature_RNA wordt er verwacht dat wanneer de count toeneemt de nFeature ook toeneemt en dit een rechte lijn vormt. Mijn eigen data ziet er goed uit. Het lijkt op een rechte lijn en er is te zien dat de count toeneemt de nFeature ook toeneemt.

Voor vervolg analyse is er voor gekozen om te filteren op de volgende settings: nFeature_RNA > 200 & nFeature_RNA < 7500 & percent.mt < 5. Hiermee worden de mogelijk slechte cellen eruit gefilterd, maar

wel de meeste data meegenomen.

Discussie

Bij het filteren van de features is er gefilterd op: nFeature_RNA > 200 & nFeature_RNA < 7500. Hiervoor is gekozen aan de hand van figuur 1 waar de features weergegeven zijn. Hierin is te zien dat er veel verschillende features aanwezig zijn. Omdat er zo veel features zijn, is de waarde best hoog gezet zodat er niet te veel data uit gefilterd wordt. Wanneer het hoger gezet wordt, kan het mogelijk zijn dat er data meegenomen wordt van meercellige druppels bij de isolatie. De cellen die worden meegenomen hebben een mitochondriaal percentage onder de 5%. Hier valt de meeste data onder zoals te zien is in Figuur 1. Wanneer dit percentage hoger gezet wordt, kan het zijn dat er cellen met een slechte kwaliteit worden meegenomen.

Er bestaat een risico dat er te veel cellen of features worden weggefilterd. Dit kan de verdere analyse beïnvloeden, omdat potentieel belangrijke data, zoals genen die als markers kunnen dienen voor nieuwe clusters, verloren gaan. Daarom is het belangrijk om de resultaten kritisch te evalueren en de invloed van de filtering op de downstream-analyses zorgvuldig te beoordelen.