Data_8_5_analyse_2_PC_resolutie

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 celtypes.

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.

Doel

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

Deelvraag:

Hypothese

Data analyse met Seurat

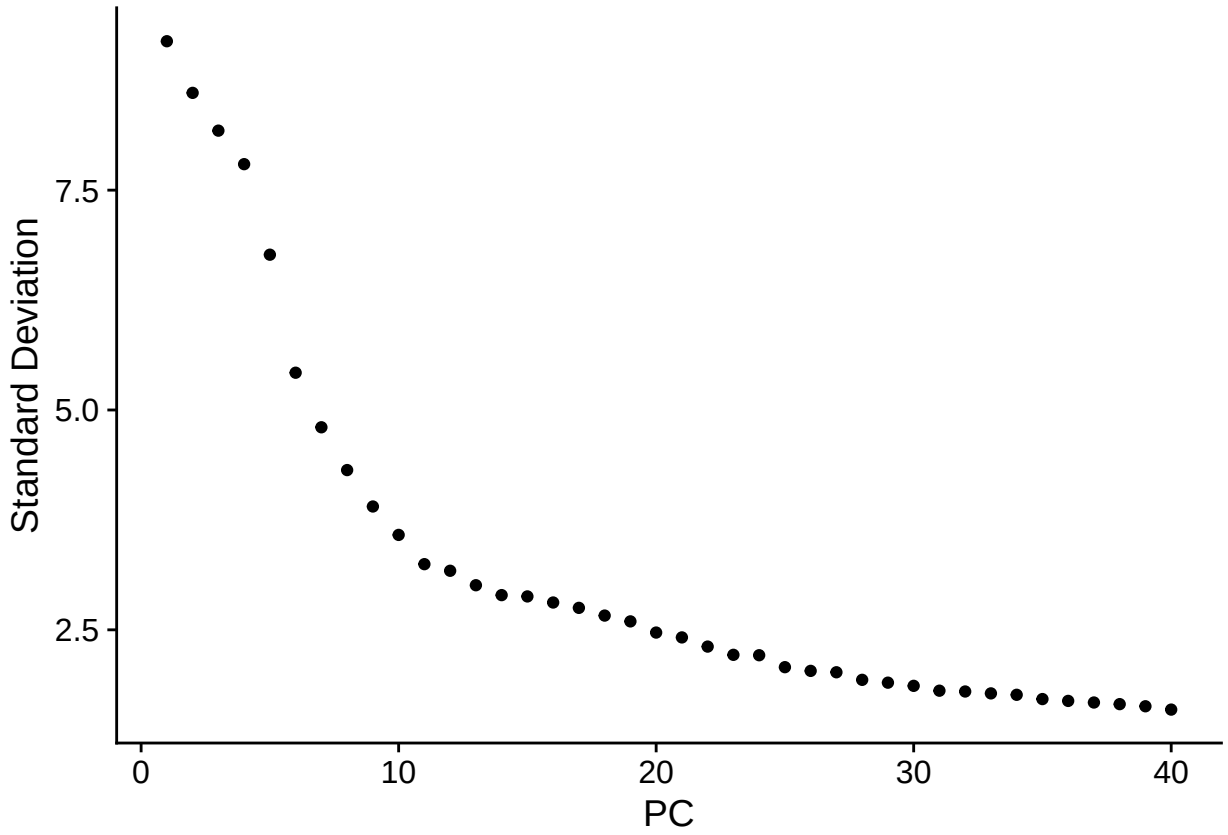
In Seurat worden de volgende stappen doorlopen

Het inladen van de data

De data wordt inladen 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.

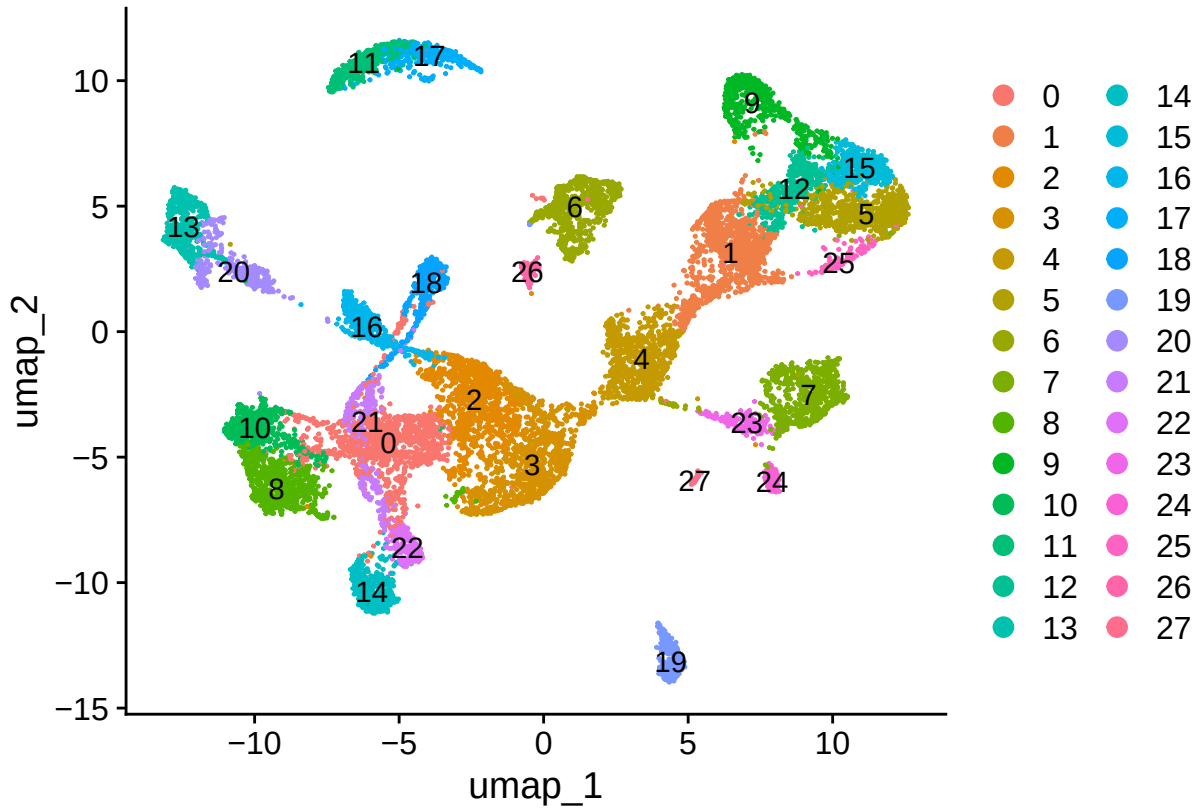
De ruwe data van het onderzoek is te vinden bij de GEO website onder het volgende nummer: GSE176588.

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: 14817
## Number of edges: 507269
##
## Running Louvain algorithm...
## Maximum modularity in 10 random starts: 0.9115
## Number of communities: 28
## Elapsed time: 1 seconds
```


umap_2
umap_1
10
5
0
−5
−10
−15
−10
−5
0
5
10
11
17
9
15
12
5
6
13
1
25
26
20
18
16
4
2
7
10
21
0
23
3
27
24
8
22
14
19
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27