

# seurat

Jalisa van der Zeeuw

2025-06-01

## 0.1 Inleiding

Voordat alternatieve splicing-analyse met SUPPA2 kan worden uitgevoerd, moet de ruwe VASA-seq data eerst zorgvuldig worden voorbewerkt. In dit project analyseren we VASA-seq data afkomstig van muizenembryo's. VASA-seq is een techniek die het mogelijk maakt om het transcriptoom van individuele cellen te profileren, waardoor we gedetailleerde genexpressie-informatie krijgen per cel. Omdat deze data zowel biologische variatie als ruis bevat, is een grondige kwaliteitscontrole en opschoning noodzakelijk voordat verdere analyses kunnen plaatsvinden. De preprocessing wordt uitgevoerd met behulp van de Seurat-toolkit. Seurat biedt een uitgebreide workflow voor kwaliteitscontrole, normalisatie, identificatie van variabele genen, dimensionele reductie, en clustering van cellen.

## 0.2 Doelstelling

Het doel van deze analyse is om uit de VASA-seq data betrouwbare biologische inzichten te verkrijgen door middel van zorgvuldige verwerking en analyse in Seurat. Hierbij ligt de focus op het identificeren van celtypen en celtoestanden, die vervolgens gebruikt kunnen worden voor verdere analyse van alternatieve splicing met SUPPA2.

Het doel van deze analyse is om de volgende deelvraag te beantwoorden: *Hoe wordt VASA-seq data verwerkt met Seurat om cellen te filteren, clusteren en geschikt te maken voor analyse met SUPPA2?*

### 0.2.1 Packages laden

We beginnen met het laden van de benodigde R packages voor de analyse. Deze pakketten zijn nodig voor data-manipulatie (`dplyr`, `tidyverse`), visualisatie (`ggplot2`), en het werken met single-cell data (`Seurat`, `Matrix`).

### 0.2.2 Inladen van metadata

In deze stap laden we de metadata in voor zowel de features (genen) als de samples (cellen). De featuremetadata bevat bijvoorbeeld de namen en eigenschappen van de genen, terwijl de samplemetadata informatie bevat over de individuele cellen, zoals een cel-ID of celtype. Deze informatie is essentieel om de ruwe readcounts correct te koppelen aan genen en cellen, en vormt dus een onmisbaar onderdeel van de dataset.

### 0.2.3 Inladen van de count matrix

De ruwe expressiedata van de cellen wordt ingelezen in de vorm van een sparse matrix, waarbij rijen genen voorstellen en kolommen individuele cellen. Deze count matrix bevat het aantal reads per gen per cel, en vormt de basis voor de downstream-analyse. Met de functie `ReadMtx()` worden de matrix zelf, de bijbehorende geninformatie (features) en celinformatie (barcodes) gecombineerd tot één structuur.

## 0.2.4 Aanmaken van het Seurat-object

De ingelezen count matrix wordt nu omgezet naar een Seurat-object met de functie `CreateSeuratObject()`. Dit object, dat hier wordt opgeslagen onder de naam `seurat`, bevat de expressiegegevens van de cellen en wordt gebruikt als uitgangspunt voor alle volgende analyses. Na het aanmaken wordt het object kort bekeken met `seurat` en `heads(seurat)` om te controleren of het correct is opgebouwd en de data er logisch uitziet.

```
## An object of class Seurat
## 45033 features across 18135 samples within 1 assay
## Active assay: RNA (45033 features, 0 variable features)
## 1 layer present: counts

##                                     orig.ident nCount_RNA nFeature_RNA
## E8.5-10_i22-AACTCAGTTTATCTGT    E8.5-10      5475.30     3046
## E8.5-10_i22-AAGAACAGATCTTGTT   E8.5-10      10520.87     4497
## E8.5-10_i22-AAGCCTTCAGCTACGG   E8.5-10      9080.69      4119
## E8.5-10_i22-AAGCCTTCCCCGAATG   E8.5-10      12031.87     4873
## E8.5-10_i22-AAGCCTTCCTCTGGA    E8.5-10      17779.75     6563
## E8.5-10_i22-AAGGATGACGTACCTA   E8.5-10      7944.99      3779
## E8.5-10_i22-AAGCGCCTCTACGAGC   E8.5-10      16567.70     5800
## E8.5-10_i22-AAGTATTGCTCGCGTA   E8.5-10      7269.05      3515
## E8.5-10_i22-AAGTTGTCACGGTAGC   E8.5-10      12097.91     4835
## E8.5-10_i22-AAGTTGTCCGTATTTC  E8.5-10      8067.23      3859
```

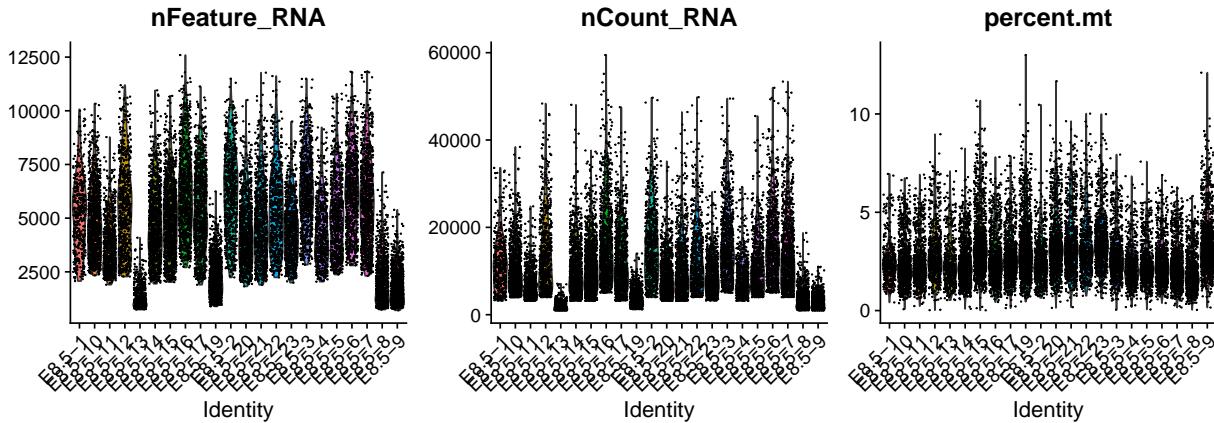
## 0.2.5 Berekenen van het mitochondriaal percentage

Om de kwaliteit van de cellen te beoordelen, wordt het percentage mitochondriale genexpressie per cel berekend. Een verhoogd percentage mitochondriale RNA's kan wijzen op beschadigde of afstervende cellen. Mitochondriale genen worden herkend aan het voorvoegsel “mt-” in hun naam. Deze genen worden geselecteerd op basis van hun naam, waarna het percentage mitochondriale expressie per cel wordt toegevoegd aan het Seurat object, `seurat`, als nieuw metadata-kolom `percent.mt`.

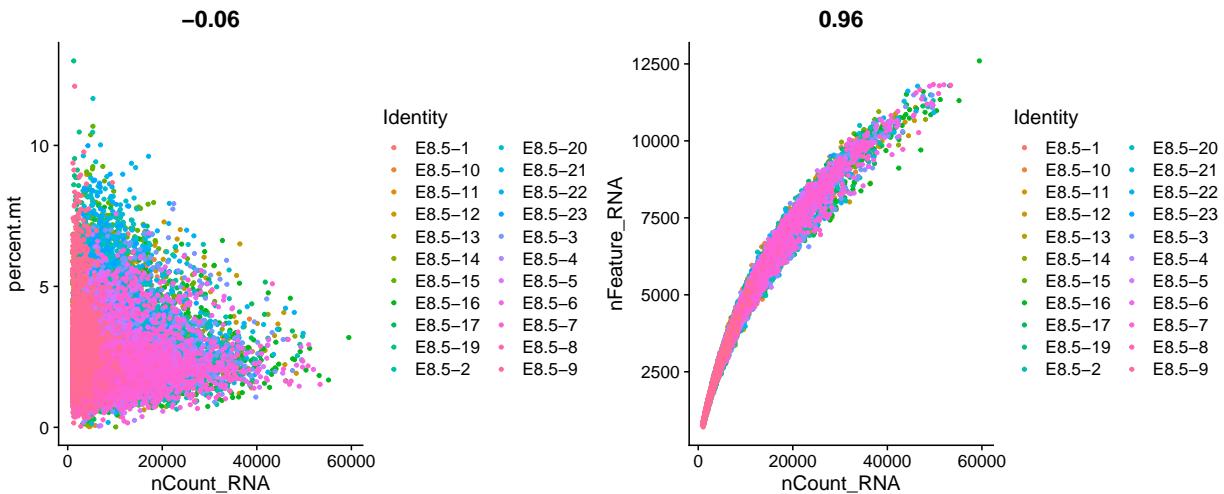
## 0.2.6 Quality Control visualisaties

In deze stap maken we verschillende grafieke om de kwaliteit van onze data te beoordelen. Met een violin plot (VlnPlot) visualiseren we per cel: - het aantal gedetecteerde genen (`nFeature_RNA`) - het totaal aantal getelde transcripts (`nCount_RNA`) - het percentage mitochondriale genen (`percent.mt`).

Deze plots helpen om cellen met slechte kwaliteit, zoals cellen met weinig genexpressie of veel mitochondriale RNA te identificeren. Het percentage mitochondriale genexpressie is een indicatie van de kwaliteit van de cellen; een hoog percentage kan duiden op beschadigde of gestresste cellen. Door deze plots te bekijken, kunnen we bepalen welke cellen geschikt zijn voor verdere analyse.



Vervolgens maken we scatterplots waarbij deze verschillende kwaliteitskenmerken met elkaar vergeleken worden. Door deze kenmerken in relatie tot elkaar te bekijken, kunnen we mogelijke verbanden of afwijkingen identificeren die met afzonderlijke visualisaties niet direct zichtbaar zijn. Bijvoorbeeld, een hoge mitochondriale expressie samen met een laag aantal gedetecteerde genen kan duiden op beschadigde cellen. Maar een hoog mitochondriale expressie samen met een hoog aantal gedetecteerde genen, kan weer duiden op doublets. Dit geeft aanwijzingen over eventuele slechte kwaliteit cellen of technische artefacten.



### 0.2.7 Filtering

Om de kwaliteit van de dataset te waarborgen, worden de cellen gefilterd. De cellen die tussen de 200 en 2500 genen hebben en minder dan 6% mitochondriale expressie worden behouden. Deze gefilterde dataset, hierna genoemd `seurat.filtered`, wordt gebruikt voor downstream analyse. De gekozen grenzen voor filtering zijn gebaseerd op basis van de zojuist gemaakte plots, rekening houdend met de biologische aard van de data (muizen embryo's).

### 0.2.8 Normalisatie, selectie van variabele genen en dataschaling met SCTransform

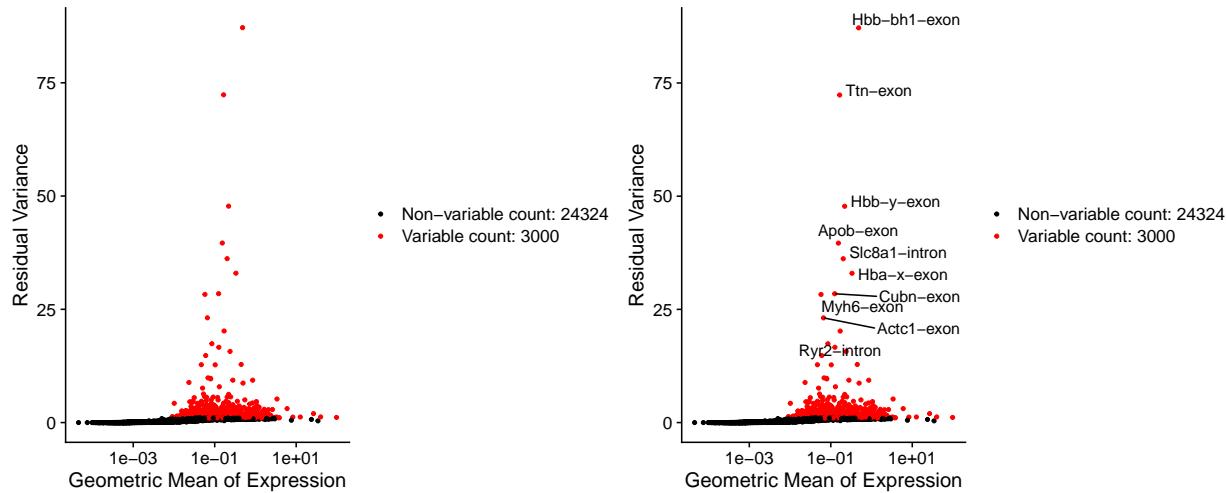
In deze stap passen we normalisatie en variable geneselectie toe met de functie `SCTransform`. Deze methode vervangt de klassieke log-normalisatie en integreert tegelijk dataschaling en regressie. Hierbij wordt het effect van mitochondriale genexpressie (`percent.mt`) gecorrigeerd, omdat dit de downstream analyses kan verstören. Door regressie op `percent.mt` verminderen we deze technische variatie, zodat de biologische

verschillen beter naar voren komen. Normalisatie zorgt ervoor dat verschillen in sequencing-diepte tussen cellen worden gecorrigerd, zodat genexpressiewaarden beter vergelijkbaar zijn. Dataschaling brengt de genexpressiewaarden op een vergelijkbaar schaalniveau, wat belangrijk is voor de downstream methoden (PCA en clustering). Tegelijk selecteert **SCTransform** automatisch de meest variabele genen (HVG's), wat belangrijk is voor de downstream analyses. De hoog variabele genen zijn genen die tussen cellen sterk verschillen in expressie, en geven ons de meeste informatie. Op basis van deze informatie kunnen we verschillen in celtypen of toestanden onderscheiden.

Hieronder zien we het aantal HVG's dat **SCTransform** heeft geselecteerd en de top10. Vervolgens zie je plots met alle genen, waarbij de hoog variable genen rood zijn gemarkeerd. Deze HVG's worden meegenomen voor derere analyse. Daarnaast wordt dezelfde plot getoond, maar dan met de top10 HVG's gelabeld, zodat je direct de belangrijkste variabele genen kunt herkennen.

```
## [1] 3000

## [1] "Hbb-bh1-exon"   "Ttn-exon"        "Hbb-y-exon"      "Apob-exon"
## [5] "Slc8a1-intron"  "Hba-x-exon"      "Cubn-exon"       "Myh6-exon"
## [9] "Actc1-exon"     "Ryr2-intron"
```



## 0.2.9 Principal Component Analysis (PCA)

We voeren een PCA uit op de hoogst variabele genen om de belangrijkste variaties in de data te identificeren. PCA reduceert de hoge dimensies van de dataset naar een kleiner aantal samengestelde variabelen, genaamd principal components (PC's). Met dimensies wordt bedoeld het aantal variabelen waarop de cellen worden geanalyseerd. Elke dimensie komt in dit geval overeen met de expressiewaarde van een gen in een cel. Hierdoor wordt het de data eenvoudiger te visualiseren en interpreteren. De eerste paar principal components bevatten de meeste informatie over de variatie tussen cellen.

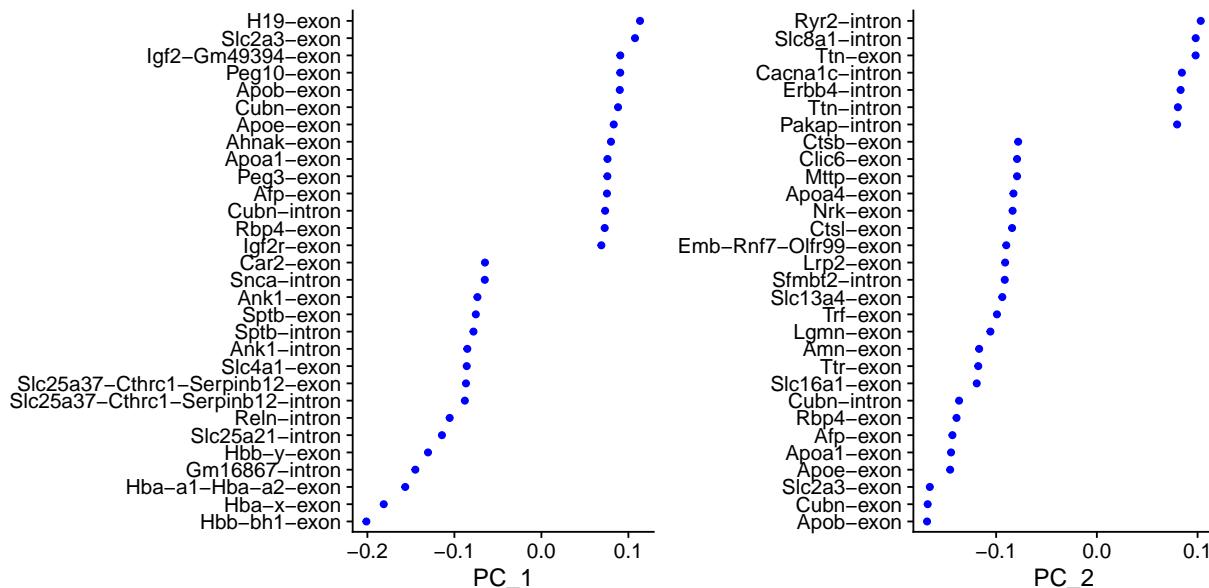
De volgende plot toont de genen die het meest bijdragen aan de geselecteerde principal components (PC's), in dit geval PC1 en PC2. Elke stip vertegenwoordigt een gen en de positie op de x-as geeft de bijdrage (loading) van dat gen aan de PC. Genen aan de uiteinden van de assen (van de 0 af) dragen het sterkst bij aan de variatie in die PC. De genen links van 0 dragen daarbij negatief bij, rechts van de 0 positief - dat betekent dat deze genen in verschillende richtingen bijdragen aan de scheiding van de cellen. De genen met positieve en negatieve bijdragen hebben tegenovergestelde expressiepatronen in de cellen: als genen met een positieve bijdrage hoog tot expressie komen in een groep cellen, zullen de genen met negatieve bijdrage juist laag tot expressie komen, en andersom.

Om een inzicht te krijgen in welke biologische processen of celtypen de dimensie eigenlijk presenteert, kun je informatie zoeken over de genen die hier worden weergeven:

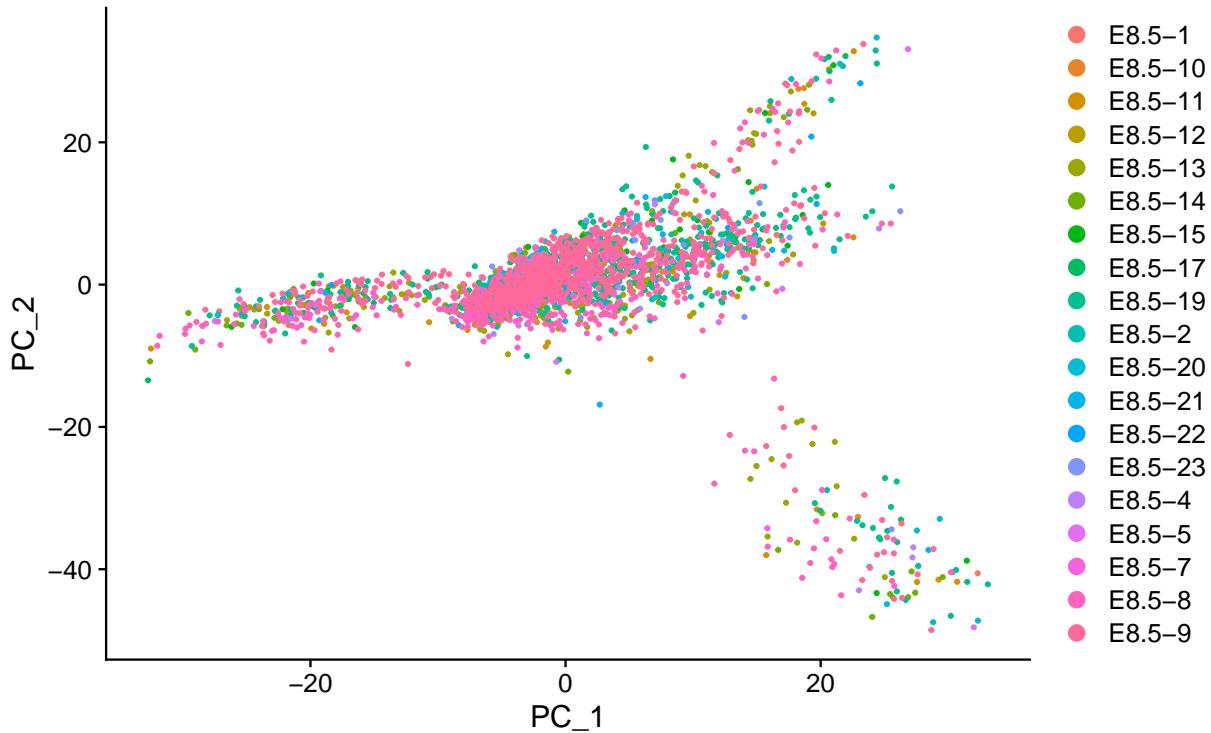
De genen die positief bijdragen aan PC1 omvatten bekende imprintingsgenen en genen betrokken bij groei en ontwikkeling, zoals *H19*, *Igf2*, *Peg10* en *Apoe*. Dit wijst erop dat PC1 mogelijk een biologisch verschil representeert gerelateerd aan imprinting en embryonale ontwikkeling.

De genen die negatief bijdragen aan PC1 zijn voornamelijk gerelateerd aan rode bloedcel functie en ontwikkeling, waaronder verschillende hemoglobine-genen (*Hbb-y*, *Hba-a1*) en cytoskelet-gerelateerde genen (*Ank1*, *Sptb*). Dit suggereert dat PC1 de variatie onderscheidt tussen cellen met een bloedcel-achtige expressie en cellen met een imprintings- en groei-gerelateerde expressie.

Het opzoeken van de genen die bijdragen aan PC's helpt je om je data biologisch te interpreteren, waardoor je analyses meer betekenis krijgen.

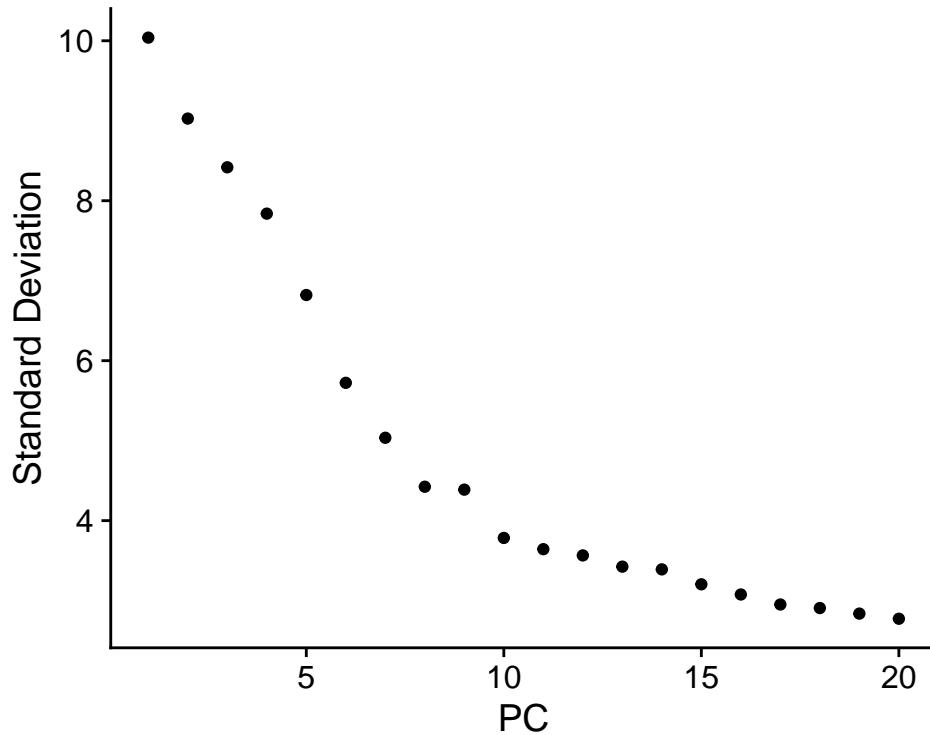


De Dimplot hieronder toont de cellen geprojecteerd in de ruimte van de belangrijkste hoofdcomponenten (PC's). Elke stip vertegenwoordigt een cel, en de positionering laat zien hoe vergelijkbaar de cellen zijn qua genexpressieprofiel. Cellen die dicht bij elkaar liggen, hebben een vergelijkbaar expressiepatroon, terwijl cellen die ver van elkaar liggen, verschillen in hun genexpressie. Deze visualisatie geeft een eerste in druk van de grote structuren en mogelijke clusters in de dataset. Dit is een globale kwaliteitscheck om te zien of er structuur in de data zit. Stel dat er helemaal geen groepen te zien zijn, dan is dat verdacht.



Vervolgens wordt er een ElbowPlot gemaakt om te bepalen hoeveel PC's worden meegenomen voor verdere analyse, clustering en UMAP. Deze plot helpt kiezen hoeveel PC's informatief genoeg zijn om te behouden. Op de x-as zie je het nummer van de PC, en op de y-as de hoeveelheid variantie die elk PC verklaard (hoeveel informatie die PC bevat). De lijn daalt geleidelijk: de eerste PC's bevatten veel informatie, latere steeds minder. Er wordt gekeken naar het punt waarop de lijn afvlakt - daar ontstaat een soort knik, ook wel de elleboog genoemd. Dat punt geeft aan hoeveel PC's er het beste kunnen worden meegenomen.

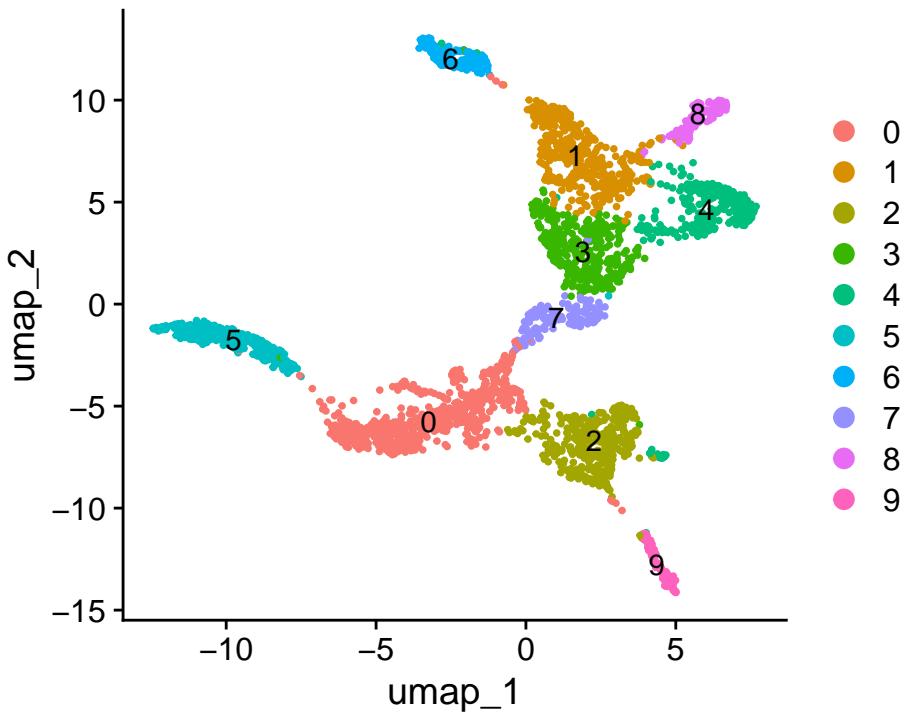
Er is besloten om PC1 t/m PC8 mee te nemen voor verdere analyse.



### 0.2.10 Clustering van cellen

De volgende stap is het groeperen van cellen die vergelijkbare genexpressieprofielen hebben. Dit noemen we clustering. Clustering helpt om biologisch relevante celtypes of -toestanden te onderscheiden zonder vooraf te weten wat die precies zijn. Door te clusteren krijgen we inzicht in de biologische structuur van de data. De clusters vormen de basis voor vervolganalyse, zoals: - het herkennen en benoemen van celtypes, - het ontdekken van nieuwe of zeldzame celpopulaties, - of het vergelijken van celgroepen tussen condities.

```
## Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck
##
## Number of nodes: 2994
## Number of edges: 89010
##
## Running Louvain algorithm...
## Maximum modularity in 10 random starts: 0.9027
## Number of communities: 10
## Elapsed time: 0 seconds
```



*dit moet ik nog toevoegen*

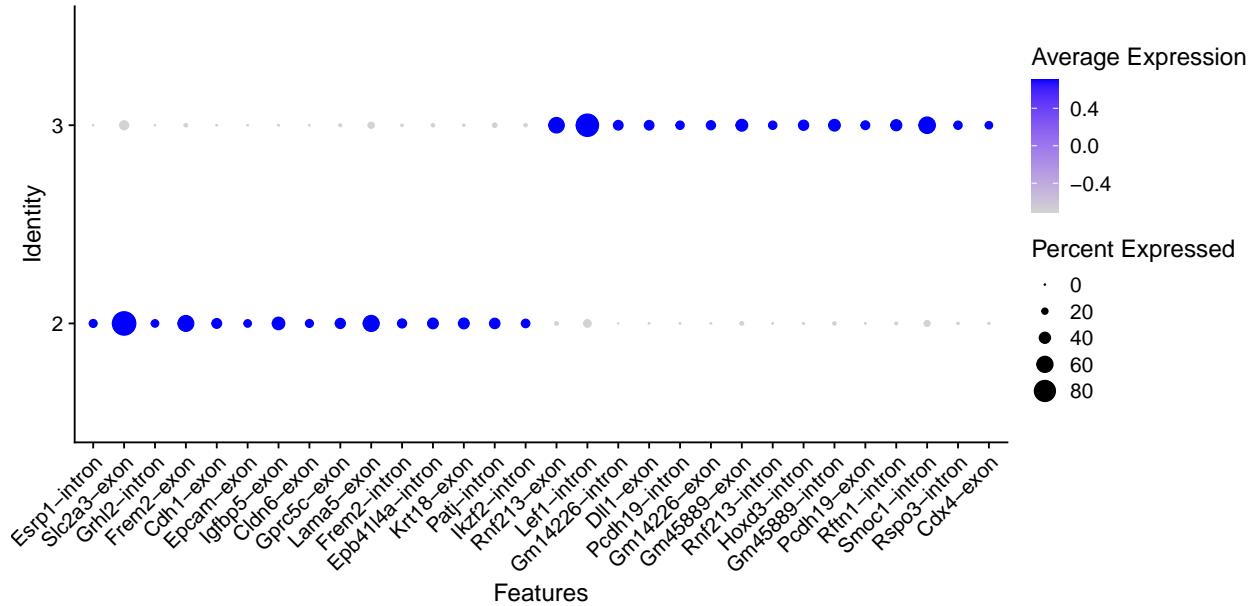
```
## # A tibble: 30 x 7
## # Groups:   cluster [10]
##       p_val avg_log2FC pct.1 pct.2 p_val_adj cluster gene
##       <dbl>     <dbl> <dbl> <dbl>    <dbl> <fct> <chr>
## 1 1.03e-135      4.42 0.337 0.023 2.81e-131 0     Nrcam-intron
## 2 1.21e-151      4.38 0.352 0.018 3.29e-147 0     Rfx4-intron
## 3 6.51e-129      3.97 0.346 0.03  1.78e-124 0     Ptprn2-intron
## 4 9.34e-138      6.45 0.273 0.006 2.55e-133 1     Ebf2-intron
## 5 3.63e-147      4.22 0.39  0.027 9.93e-143 1     Col26a1-intron
## 6 3.36e-156      3.95 0.462 0.043 9.18e-152 1     Foxp2-intron
## 7 1.26e-121      4.93 0.269 0.01  3.45e-117 2     Esrp1-intron
## 8 4.89e-109      4.86 0.253 0.011 1.34e-104 2     Grhl2-intron
## 9 5.75e- 97       4.13 0.253 0.015 1.57e- 92 2     Epcam-exon
## 10 2.25e-110      4.80 0.286 0.017 6.14e-106 3     Pcdh19-intron
## 11 9.04e-112      4.40 0.336 0.028 2.47e-107 3     Dll1-exon
## 12 3.42e-129      4.28 0.339 0.021 9.36e-125 3     Gm14226-intron
## 13 1.29e-163      5.59 0.425 0.025 3.51e-159 4     Col1a1-exon
## 14 2.22e-151      4.66 0.357 0.015 6.05e-147 4     Hand1-exon
## 15 1.27e- 93      4.63 0.305 0.028 3.46e- 89 4     Col3a1-exon
## 16 3.55e-168      8.86 0.28  0     9.71e-164 5     Spta1-intron
## 17 0               8.64 0.593 0.004 0     5     Slc4a1-exon
## 18 6.34e-182      8.02 0.309 0.001 1.73e-177 5     Nmmat3-intron
## 19 8.16e-163      9.31 0.282 0.001 2.23e-158 6     Stab2-intron
## 20 1.85e-311      8.78 0.526 0.002 5.06e-307 6     Cdh5-exon
## 21 2.75e-215      8.36 0.359 0.001 7.52e-211 6     Erg-intron
## 22 1.73e-137      4.91 0.446 0.022 4.72e-133 7     Nkx1-2-exon
## 23 3.58e-101      4.55 0.331 0.016 9.79e- 97 7     EphA5-exon
```

```

## 24 3.75e-205      4.50 0.811 0.062 1.02e-200 7      EphA5-intron
## 25 0              9.66 0.77  0.005 0          8      Ttn-intron
## 26 5.64e-316     9.62 0.603 0.006 1.54e-311 8      Myh6-exon
## 27 2.82e-318     9.56 0.905 0.036 7.71e-314 8      Ttn-exon
## 28 1.94e-275     9.54 0.439 0          5.29e-271 9      Cdhr2-exon
## 29 1.99e-304     9.33 0.5    0.001 5.44e-300 9      Afp-intron
## 30 2.49e-185     8.47 0.307 0.001 6.81e-181 9      Gm29721-intron

## character(0)

```



### 0.3 Conclusie en Discussie *mis nog niet af, moet nog worden toegevoegd en herschreven*

Normaliseren: SCTransform is vooral beter als je data complexer is, met variabele sequencing depth, veel lage expressiewaarden en technische variatie — precies het geval bij VASA-seq. Daarom zie je in moderne pipelines dat SCTransform de voorkeur krijgt.

HVG = 3000: In embryonale datasets is de celdiversiteit doorgaans hoger vanwege de aanwezigheid van diverse celtypen en ontwikkelingsstadia. Dit zorgt voor complexere en subtielere veranderingen in genexpressiepatronen. Door meer variabele genen te selecteren (3000 in plaats van de gebruikelijke 2000) kunnen we deze biologische diversiteit van embryonale ontwikkeling beter vastleggen. Bovendien helpt een grotere genenset om zwakkere signalen op te vangen en verbetert het de detectie van fijne celtypen en subpopulaties.

Na het toepassen van SCTransform() is automatisch een selectie gemaakt van de meest variabele genen binnen de dataset. Dit zijn genen waarvan de expressie sterk verschilt tussen cellen, en ze zijn belangrijk omdat ze vaak bijdragen aan het onderscheiden van celtypen in latere analyses (zoals clustering of trajectanalyse). Om te controleren of deze selectie biologisch logisch is — en niet alleen technische ruis bevat — heb ik de top 10 meest variabele genen bekeken. Deze controle geeft inzicht in welke genen de meeste bijdrage leveren aan de variatie tussen cellen in de dataset. De topgenen bevatten onder andere: - Hemoglobinegenen (Hbb-bh1, Hbb-y, Hba-x): typisch actief in vroege rode bloedcellen in het embryo - Spier- en hartgerelateerde genen (Ttn, Myh6, Actc1, Ryr2): betrokken bij hartontwikkeling, wat past bij cellen in deze embryonale fase - Genen betrokken bij transport en metabolisme (zoals Apob, Cubn, Slc8a1): kunnen wijzen op actieve cellen met functies in vet- of iontransport

extra regressie stap. genen: Tijdens de normalisatie met SCTransform() heb ik het percentage mitochondriale genexpressie (percent.mt) meegeïncorporeerd als regressiefactor. dit betekent niet dat mitochondriale genen zijn verwijderd, maar dat hun invloed op de expressieniveaus van andere genen is gecorrigeerd. zo wordt voorkomen dat variatie door celstress of technische redenen de clustering en downstream analyses beïnvloed.