BRIE2_tutorial

Anne Brussaard

2025-06-14

1 BRIE2 tutorial

Na het uitvoeren van de seurat tutorial, filtering onderzoek en analyse pipeline van seurat, stappen we over naar de volgende stap namelijk BRIE2. Om te begrijpen welke mogelijkheden BRIE2 heeft voor wordt eerst een tutorial gevolgd. Deze is afkomsting van de maker van BRIE (https://brie.readthedocs.io/en/latest/brie2_msEAE.html). In deze tutorial ligt de focus op hoe de uitkomsten van BRIE-qaunt kan worden gevisualiseerd. Hoe de input van BRIE2 wordt verkregen wordt niet behandeld.

1.1 Deelvraag: Kan ik door het volgen van een tutorial met BRIE2 data anlayse uitvoeren?

De data die is gebruikt voor het uitvoeren van deze tutorial is van Falcao et al, 2018. Het bevat 2208 muizencellen die onderzocht zijn met SMART-seq2 waarbij de ene helft experimentele auto0immuun encefalomyelitis (EAE) cellen zijn. Deze cellen bootsen multiple scleorsis na, de andere helft zijn controle cellen. Het wordt gebruikt om na te bootsen hoe BRIE2 gebruikt kan worden of differntiele splicing events tussen twee groepen cellen te detecteren.

2 Analyse

Dan worden de packages geladen die nodig zijn voor het uitvoeren van de analyse. dit gaat o.a. om umap, os, brie, numpy, pandas, scanpy en matplotlib.pypolot.

2.1 BRIE2 optie 1

Hierna volgt de eerste mogelijkheid van analyseren met BRIE2. Optie 1= differential splicing events. Deze modus gebruikt statistische analyse (regressie) om de splicing in de twee groepen te vergelijken. In deze modus wordt ook rekening gehouden met celtype zodat het effect apart kan worden bekeken per celtype. In het script wordt de output van brie2-quant (anndata) met de kwantificaties van isovormen gevisualiseerd. De output bevat een bestand met annotaties en een bestand met de BRIE2 parameters zoals de elbo_gain en cell_coeff (wordt bij het figuur verder toegelicht). Ook is er een bestand met geninformatie en een bestand met de input. Deze bestanden worden aan elkaar gekoppeld om verder te kunnen visualiseren.

De resultaten van optie 1 worden eerst weergeven in een volcano plot. In deze plot is Cell_coeff een statistische maat die aangeeft hoe sterk de splicing (PSI) veranderd tussen twee groepen. Een positieve Cell_coeff geeft aan dat de PSI waarden in de EAE-groep hoger is dan in de controle groep. PSI geeft hierin aan wat de kans is dat een bepaald splicing event voorkomt.De Elbo_gain die op de Y-as wordt weeergeven is een maat voor hoeveel bewijs er is van een verschil in splicing tussen de twee condities. Zie figuur 1. Op de X-as is dus de effect groote te zien dus hoe sterk de splicing verander door een conditie. positieve waarde

weizen op meer splicing voor de groep EAE. Op de Y-as is hoe de splicing kan worden verklaard door de ziekte conditie (hoe sterk is het bewijs) Je kan hieruit aflezen bij welke genen de splicing wordt verhoogd door EAE (positieve cell_coeff) en werke genen worden onderdrukt door EAE (negatieve cell_coeff) De rode stippen zijn geselecteerd omdat er veel "bewijs" is en dus intressant om verder naar te kijken.

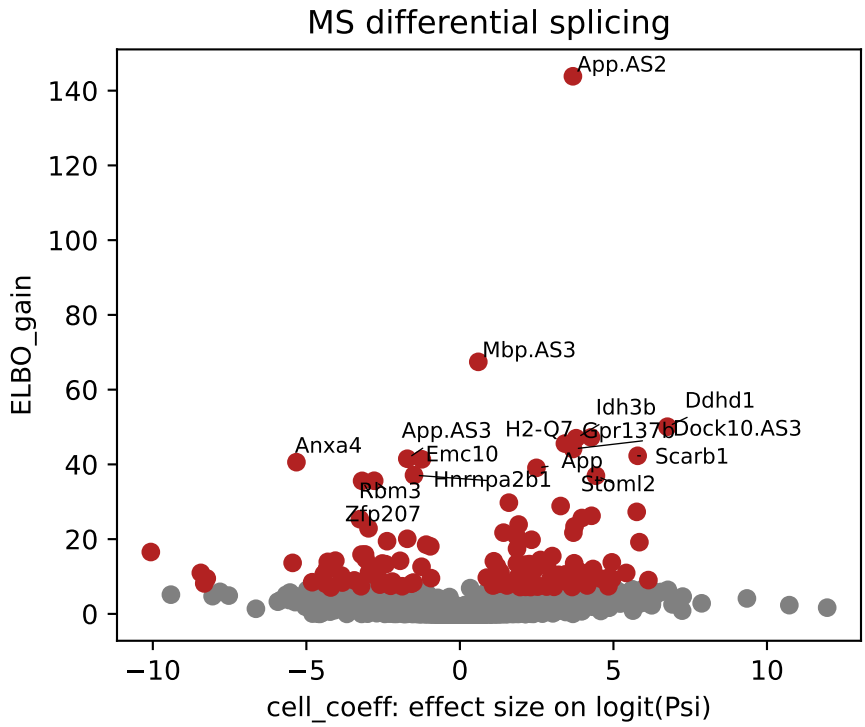

Figure 1: differential splicing events

Vervolgens worden de splicing events met een ELBO_gain van > 7 geselteerd dat zijn de differntiale splicing events. Er wordt geken naar hoeveel events dat zijn en hoeveel genen ze representeren zodat deze gebruikt kunnen worden in verdere analyse.

Daarna volgt de visulisatie van de ruwe counts van de differentiale splicing events (DSE). Dit wordt gedaan om te controleren of wat BRIE2 berekend ook zichtbaar is in de data. Dit is belangrijk omdat cell_coeff en elbo_gain berekend zijn met een Bayesiaans model en de ruwe data moet overeen komen met het model. Ook kan zo worden uitgesloten dat een gen significant lijkt omdat 1 cel een extreem hoge count heeft. Om dit te bepalen kijken we naar de verdeling van de splicing per groep. Deze plots laten de condities zien met de controle groep blauw en de EAE groep oranje. Elke stip is een enkele cel, de groote van de stip geeft aan wat de geschatte PSI waarden is (hoevaak wordt een exon geincludeerd).Steeds worden er twee isovormen vergeleken en zien we hoeveel reads er zijn voor de isovormen en in welke groep deze voorkomen. Zie figuur 2.

Sommige splicing events kunnen niet direct aan een isovorm worden toegewezen en hebben ambiguous (onduidelijke) reads. De oorzaken hiervan kunnen verschillen van lage kwaliteit, hoge sequencie gelijkenis, geen goed referentiegenoom, of overlappende genen. Met deze informatie kan verder gekeken worden naar welke pathways de genen betrokken zijn, welke celtype betrokken zijn en visualisatie maken per gen en event om duidelijk maken hoe de splicing verander per cel. Zie figuur 3.

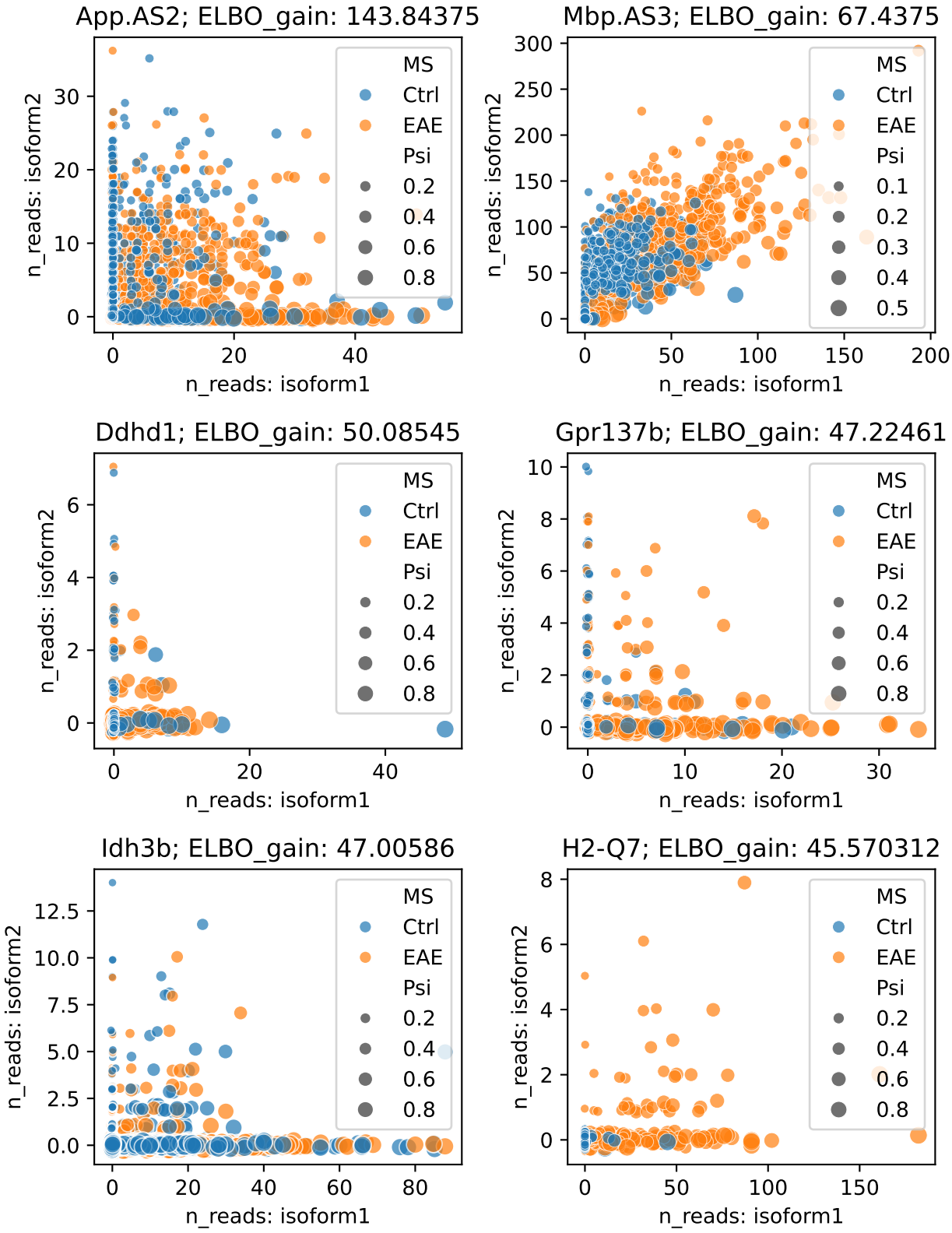

Figure 2: Ruwe counts DSE

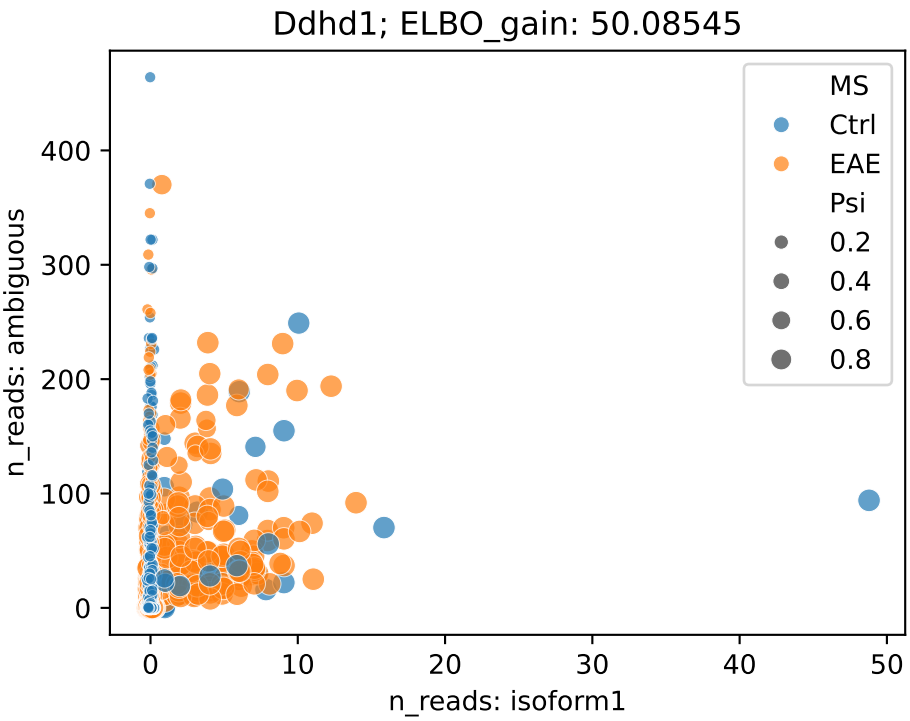

Figure 3: Verschillen in splicing van isovorm met onduidelijke reads

2.2 BRIE2 optie 2

Met de tweede optie van BRIE2 kan de kwantificatie van de splicing events uitgevoerd worden. In dit geval wordt dit gedaan met de aggregatie van de cellen om te voorkomen dat dit invloed heeft op je biologische hypothese. Dit houdt in dat de cellen als groepen worden geanalyseerd en niet als individule cellen zodat gekeken kan worden naar verschillende celtypes.

Ook worden de cellen gefilterd op basis van leesdiepte. In de onderstaande histogram zijn het aantal reads per cel te zien. Hierna worden alleen de cellen met voldoende reads (>3000 reads) meegenomen om te zorgen voor hogere betrouwbaarheid. Zie figuur 4. Er wordt een ondergrens voor het aantal reads ingesteld van 3000. De cellen die minder dan 3000 totale counts hebben worden verwijderd en deze filtering wordt ook toegepast op het oorspronkelijk adata-object zodat de cellen in beide objecten hetzelfde zijn.

Daarna worden de splicing fenotypes gevisualiseerd op basis van gen expressie in een Umap. Eerst met de verdeling van clusters en visualisatie van de EAE groep en controle groep. Iedere punt in de UMAP is een cel en zo wordt de verdeling van de cellen over de condities bekeken. Te zien is dat er 3 PC's zijn met een grote ratio, deze zullen gebruik worden voor verdere analyse. Zie figuur 5.

Vervolgens worden de genen App.As2, Mbp.AS3 en Emc10 weergeven in een scatterplot en violin plot. Voor de genen App.AS2, Mbp.AS3 en Emc10 wordt de geschatte splicing weergeven. Hierin laat de kleur zien wat de geschatte PSI waarde is en kan gekeken worden waar bepaalde splicing events meer of minder voorkomen. De genen App.As2, Mbp.As3 en Emc10 worden hier weergeven omdat deze biologisch relevant zijn. Dit zijn namelijk genen die betrokken zijn bij eiwit vouwing en myeline afbraak en dus een belangrijk onderdeel van het centrale zenuwstelsel. Het zien of er verschil is in deze genen bij gezond weefsel en "Alzheimer" weefsel kan biologisch zeer intressant zijn. In de UMAPs wordt weergeven wat de verdeling van PSI waarden is van de 3 genen. Zie figuur 6.

In de violinplots wordt per groep en conditie de verdeling in PSI waarde weergeven. Hierin wordt de PSI

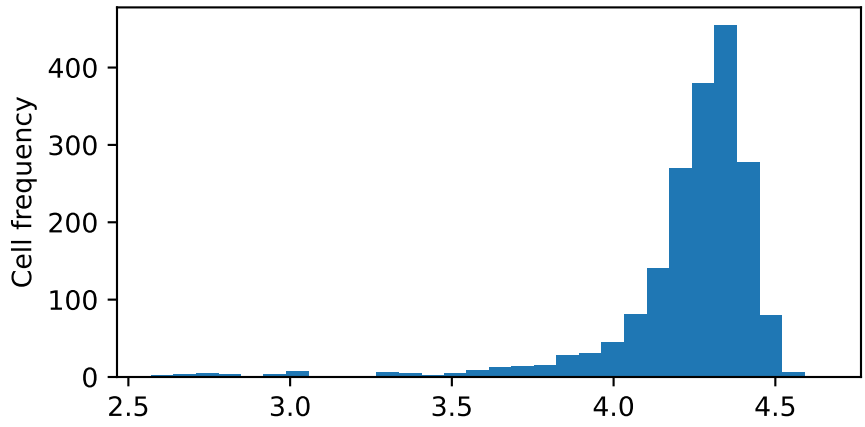

Figure 4: cell frequency

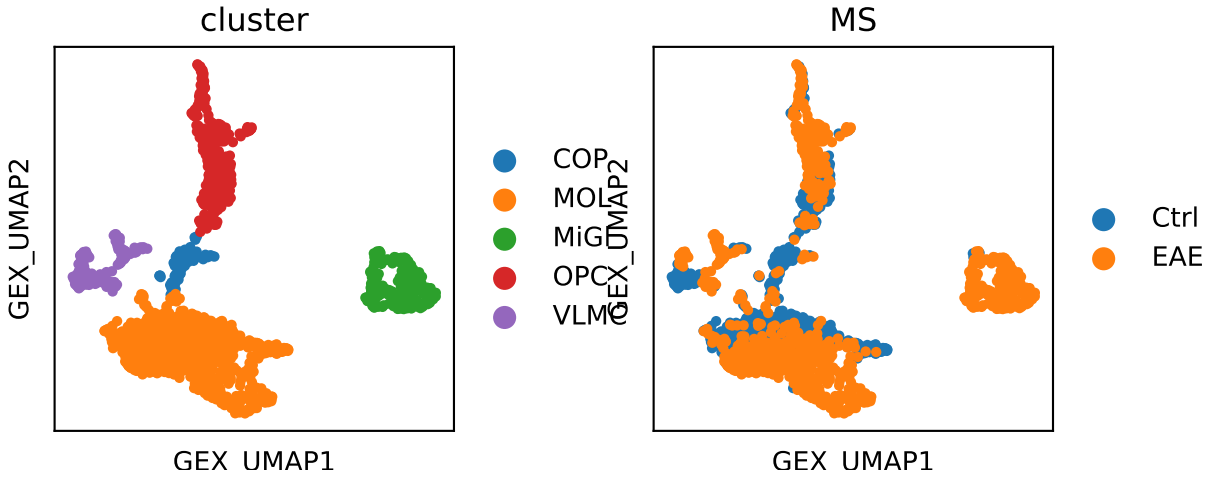

Figure 5: phenotype in gene expression UMAP

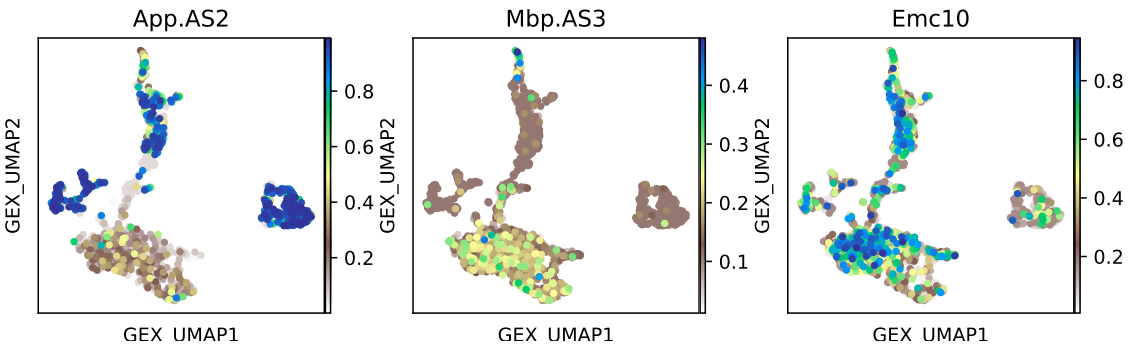

Figure 6: UMAP PSI

per cluster weergeven. Zie figuur 7.

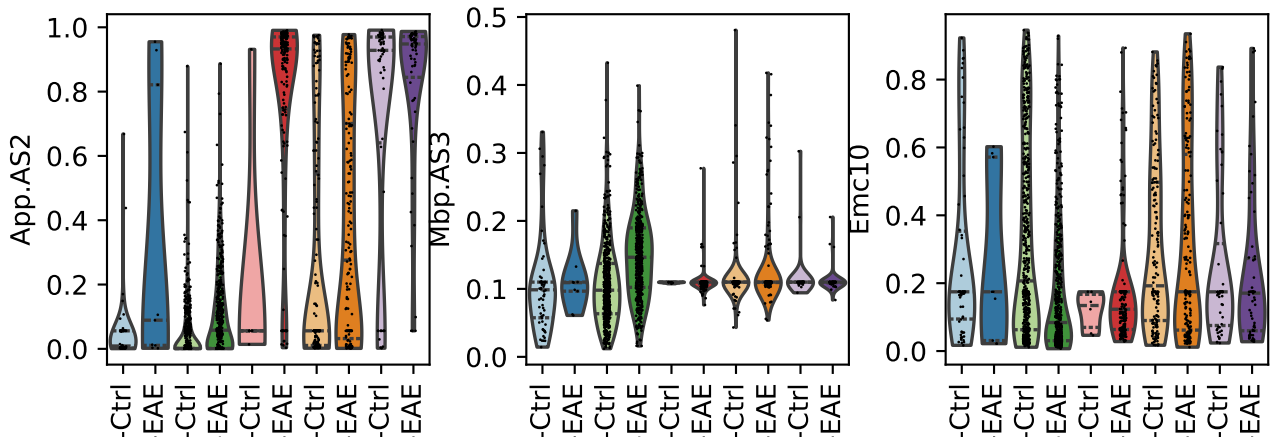

Figure 7: violin plot PSI

Voor verdere downstream analyse worden alleen de splicing events meegenomen die gedetecteerd zijn als differentaile splicing events. Deze UMAP laat de verdeling van cellen zien. Zie figuur 8.

Ook wordt de variance ratio tussen de PSI waarden berekend. Om dit te doen is eerst de splicing data omgezet naar componenten (PC). Dit houdt in dat er gekeken wordt welke PSI waarden verschillen tussen de twee groepen en statistisch relevant zijn. Dit kan vervolgens gebruikt worden om clusters te maken op basis van splicing in plaats van expressie. Zie figuur 9.

In de onderstaande UMAP is te zien wat de clusters zijn op basis van PSI waarden. Hieruit kan ook gezien worden of de splicing overeenkomt met de MS/controle. Zie figuur 10.

Hier wordt dit uitgezet voor de genen App.AS2, Mbp.AS3 en Emc10. Zie figuur 11.

Daarna wordt de UMAP nogmaals gemaakt, dit keer bevat de UMAP alleen cellen met een betrouwbare PSI, voor een naukeurige visualisatie van de splicing. Zie figuur 12.

Vervolgens zijn de eerste 3 PC geplot tegen de oorspronkelijke genexpressie UMAP. Dit kan laten zien of de splicing overeenkomt met het expressie cluster. Zie figuur 13.

2.3 Conclusie

In deze tutorial is BRIE2 toegepast op een dataset van muizencellen om differentiële splicing te analyseren tussen controle- en EAE-cellen. Zowel op individueel celniveau als op geaggregeerd niveau kon BRIE2 onderscheid maken tussen splicing events.

2.4 Discussie

Deze tutorial geeft veel informatie over het visualiseren maar niet over hoe de output van BRIE2 is verkregen. Het doel van mijn project is juist het opzetten van een pipeline in BRIE2 waarin visualisatie de laatste stap is. Deze tutorial geeft ook geen uitleg van de gemaakte stappen of waarom er bepaalde parameters zijn gekozen. Dit maakt dat het moeilijk is te begrijpen wat plaatjes betekenen en kennis toe te passen op de eigen data. Daarom is het belangrijk dat ik na het volgen van deze tutorial mij meer ga verdiepen in het verkrijgen van de output van BRIE2. De uitleg van deze tutorial is beperkt, dit heeft te maken met de uitleg die beschikbaar is vanuit de tutorial. Met behulp van internet heb ik geprobeerd de uitleg uit te breiden maar het is moeilijk om de juiste informatie te vinden.

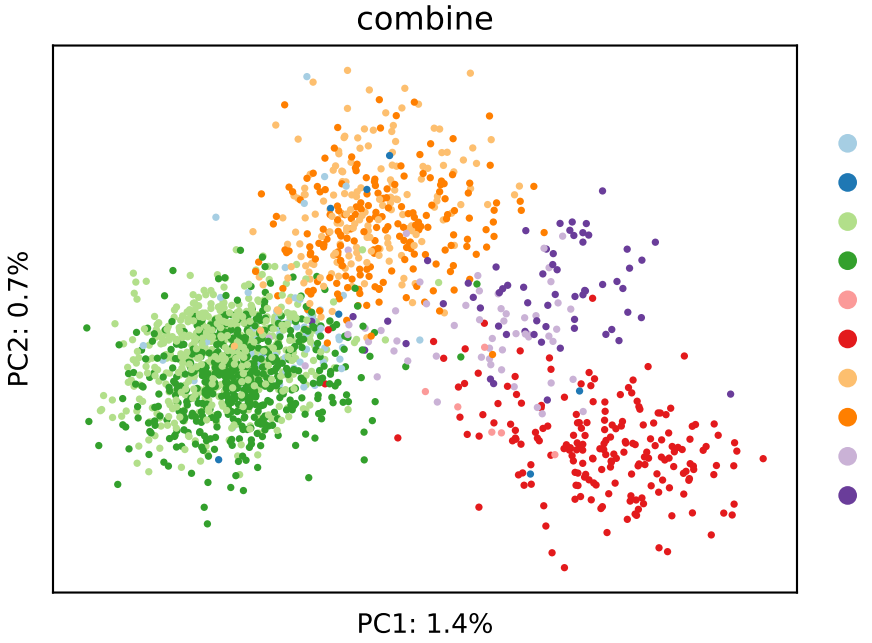

Figure 8: clusters PC

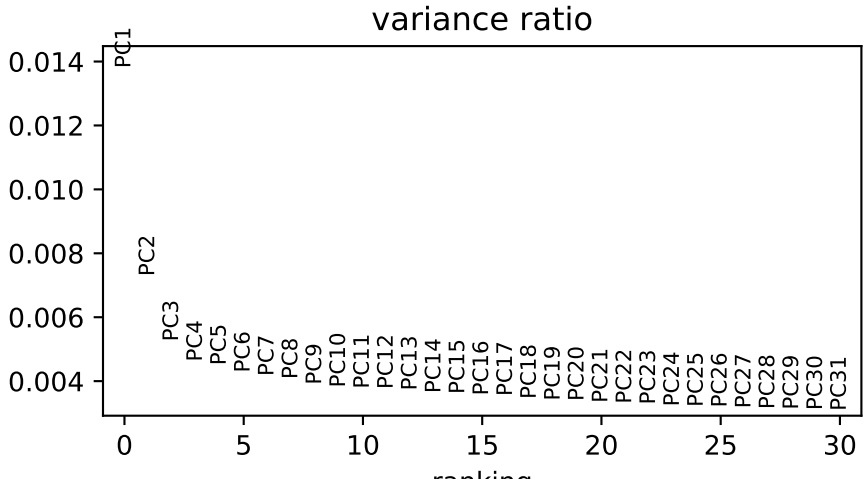

Figure 9: variance ratio

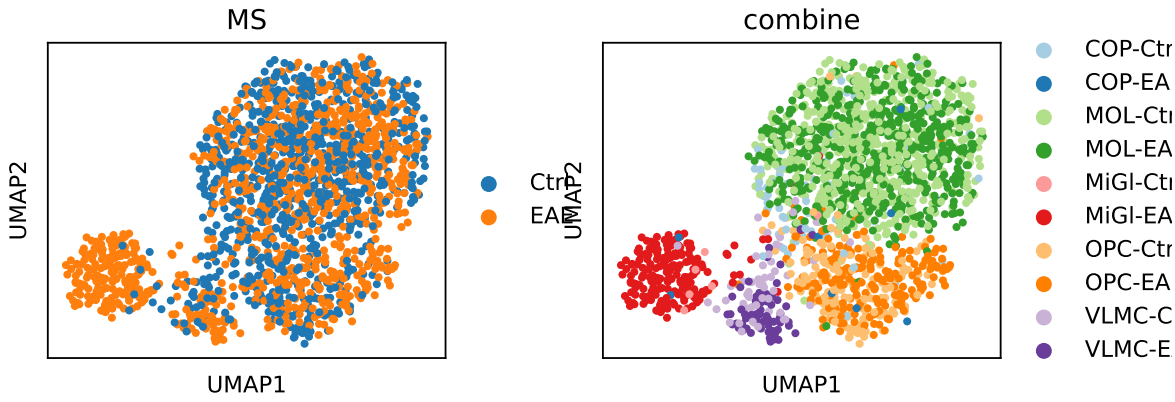

Figure 10: UMAP combine

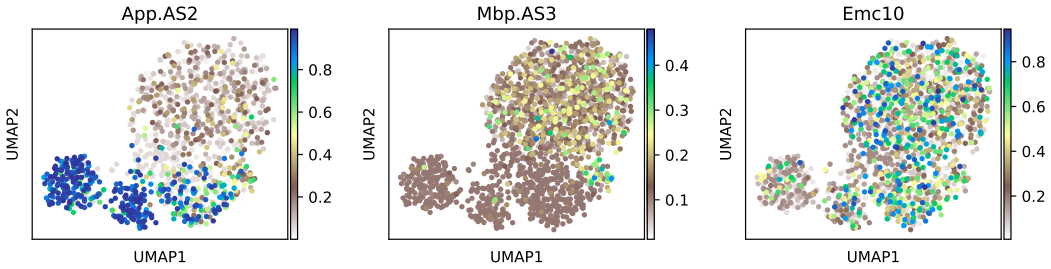

Figure 11: Visualisatie PSI value

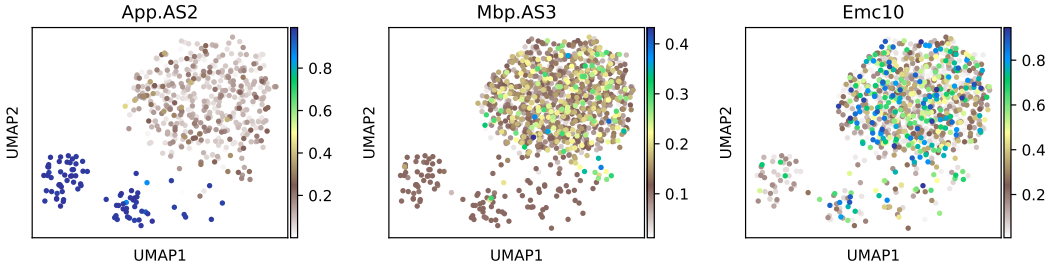

Figure 12: UMAP met betrouwbare PSI

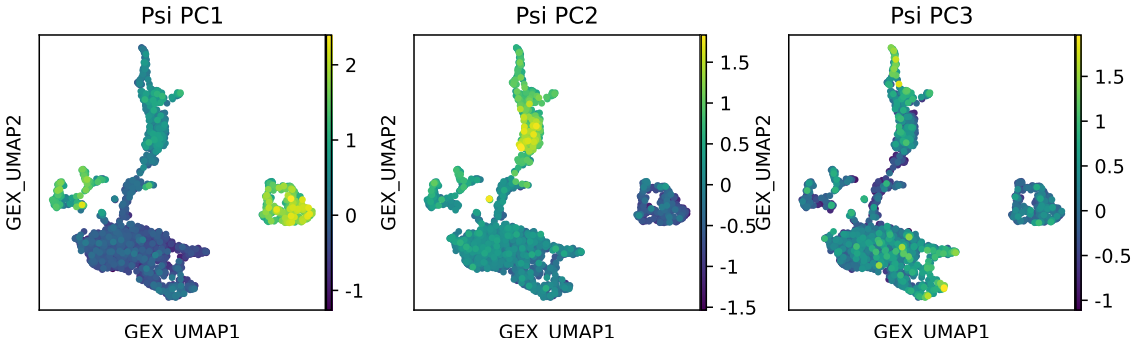

Figure 13: scatter betrouwbare PSI