Bioinformatics, 2020,0-0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: 14 Dec 2020
Manuscript Category

Subject Section

Characterization of variable features by integration of scRNA-seq and scATAC-seq with MOFA+ in human and mouse brain

Tine Logghe[1,*], David Wouters[1,*], Jose Ignacio Alvira Larizgoitia[1,*] and Alexander Ian Taylor[1,*].

[1]Department of Bioinformatics.

*To whom correspondence should be addressed.

Abstract

Motivation: Integration of epigenomics and transcriptomics data into a single model is a daunting task. The computational tool "Multi-omics factor analysis" (MOFA) allows quick and unsupervised integration of multiple modalities of omics data into a single model explaining data variability. Application of MOFA to RNA-seq and ATAC-seq data originating from human and mouse brain cells produced two integrated models comprising factors encompassing features stemming from transcriptomics and epigenomics data.

Results: In both models, RNA-seq data accounted for the largest explanatory body of data. ATAC-seq data from distal regions had higher explanatory power than from promoter regions. Analysis of primary factors produced one factor of interest in each model. Gene set enrichment analysis of the dominant factors indicates that both models' variabilities are generally explained by genes involved in processes such as up or down-regulation of neurogenesis or neuron differentiation (e.g. SOX5&6, NRXN1). The mouse model involved significantly more genes promoting startle response (e.g. NRG1), whereas the human model yielded more genes involved in upregulation of neuron development (e.g. NTRK2). Motif enrichment of distal ATAC peak data produced motifs pertaining to brain function such as Hic1 or SOX13, serving as further validation for the validity of the models.

Supplementary information: Supplementary data are available at https://github.com/WoutDavid/MOFA2.

1 Introduction

A cell is a complex entity that consists of many different and unique internal processes, some of which can influence each other both positively or negatively. These processes are the subject of various omics disciplines or layers, for example genomic and epigenomic factors together contribute to the specific landscape of the cell. Therefore, analysing data one modality at a time is often not sufficient, as each will only deliver a partial profile of the cell. In order to fully be able to profile a cell and its characteristics, the ideal solution is

to integrate all available layers and analyse them together; in this way, more complete and thus reliable results can be achieved
(Argelaguet et al., 2018; Ma et al., 2020).

Recent technological advancements are enabling researchers to work at the single-cell level and obtain information on multiple omics layers at once. This process is referred to as single-cell multimodal omics (scMulti-omics). Since it was previously not possible to analyse the vast amounts of data originating from multiple omics disciplines in an integrative way, this new branch of technologies could potentially discover new or overlooked information in various areas of research (Butler et al., 2018; Ma et al., 2020).

One technique that applies such a multi-omics integrative approach is the MOFA, short for Multi-omics factor analysis. Given a few data matrices consisting of the multi-omics input data, the MOFA model can infer representative and directly interpretable factors, ready for a wide variety of downstream analyses such as variance decomposition, denoising, manifold learning, clustering or differentiation-trajectory inference.

In this project we will utilise the MOFA+ framework, which is an improved version of the formerly developed MOFA (Argelaguet et al., 2020; Argelaguet et al., 2018). The goal is to characterise the different factors that determine the brain cell phenotype both in mouse and in human cells and to examine whether there are factors that differ between the two species. The data that we will work on to research this are scRNA-seq (single cell RNA sequencing) and scATAC-sequencing (explores chromatin availability) results, which provide both a genomic and an epigenomic layer to integrate.

2 Methods

Data

The basic analysis of the raw single cell data was processed by CellRanger. Its output, and thus the input for our MOFA analysis, can be accessed at https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets
The data consists of two datasets that have been analysed completely separately, but since all performed steps are the same in both datasets, it must be noted all steps mentioned subsequently are performed on both datasets equally, but independently. Where different parameters or databases were used, it will be mentioned. The first dataset is a 5K fresh embryonic (E18) mouse brain dataset. More particularly, cells isolated from the cortex, hippocampus, and ventricular zone from the embryonic mouse brain. The second dataset consists of 3K frozen human brain cells, taken from the cerebellum. The entire analysis was performed in R, using Rstudio.

MOFA

Multi-Omics Factor Analysis is a model that provides a general framework for the integration of multi-omic datasets in an unsupervised manner, developed by Argelaguet et al. (2018 and 2020). It can be viewed as a versatile and rigorous generalization of PCA, capable of inferring an integrated, interpretable, linearly reduced and low-dimensional representation in terms of a small number of latent factors that capture the driving sources of variation across all data modalities.

Pipeline

Preprocessing and Quality Control was done on both data modalities separately, mostly using the Seurat (Stuart et al.) package. Seurat is currently home to almost all types of single-cell analysis. The MOFA framework easily allows transformation of a complete Seurat object into a mofa object that can be trained. These two arguments made the choice for Seurat as a data structure for our analysis elementary.

To avoid inclusion of outliers, for the human dataset, cells were filtered out that had more than 10K features in the RNA modality and more than 40K features in the ATAC modality. For cells overall, the percentage of mitochondrial pattern in genes was allowed to be up to 5%. The same filter was utilized on the mouse dataset, though with cutoffs at respectively 7K RNA features, 25K ATAC features, and 10% mitochondrial genes.

Cell type annotation metadata was also added using functionality from the Seurat package, based solely on the RNA modality. Starting from the preprocessed scRNA data, we performed PCA, t-SNE and UMAP to acquire a representative lower-dimensional dataset. In this case, UMAP functioned best with our data and therefore UMAP-learnt dimensions were used to perform clustering. The resolution parameter for the clustering was set to 0.0015 in human and 0.025 in mouse, finding 9 and 7 distinct clusters respectively. Statistically significant biomarkers for each cluster were then found using the Seurat FindAllMarkers function. This information, together with the original tissue from which the cells originate, is the input of scCATCH, an R package developed by Shao et al. (2020) that uses available metadata of a scRNA-seq dataset to infer the cell types.

Peak annotation was added as metadata of the ATAC modality, with the purpose of splitting that modality into two different chromatin assays, namely “ATAC_distal" and “ATAC_promoter”. The idea here is that more robust biological conclusions can be made in the downstream analysis if the MOFA model is informed that they are different entities. Peak types were called by CellRanger and integrated into the rest of the metadata in R. While doing so, a GRanges object was made for each peak using the GenomicRanges package. This GRanges object was then used in combination with a position specific weight matrix extracted from the JASPAR2020 (Sandelin et al.) database and a full genome BioString object from UCSC in order to create a motif matrix using Seurat’s CreateMotifMatrix function. This motif matrix is added for downstream analysis.

RNA data was **normalized** using the LogNormalize method in Seurat, after which the data was **scaled** and centred to avoid confusion by the MOFA algorithm with outliers in expression values. To the same end, the ATAC data was normalized using the frequency-inverse document frequency (TF-IDF) normalization algorithm from the Signac package, which normalizes across cells, but also across peaks in order to assign higher values to rarer peaks.

5000 variable features for the RNA data were calculated using the vst **FindVariableFeatures** method in Seurat. For the ATAC-seq data, the Signac package (Adorf et al.) contains a similar function FindTopFeatures, that in lieu of a number of features, requires the definition of a minimum percentile of cells in which the top feature is required to be present. We went with $1/5^{th}$ of the total number of sequenced cells for each dataset.

After all of this preprocessing, we used mofa functionality to simply transform the formed Seurat object into a trainable mofa object. The model was trained with default parameters, with Gaussian likelihood set as distribution for all 3 modalities.

Training the model took approximately 30 minutes for each dataset, and took about 50 to 55 iterations each time. The model was trained without GPU boosting, on a commercial laptop with 8 GB DDR4 memory and an Intel Core i5-8300H 2.3Gz CPU.

Downstream analysis was almost entirely performed using MOFA framework functionality. For more detail, the reader is referred to the Results or Code availability section of this paper.

3 Results

The goal of this project is to determine factors contributing to the brain cell phenotype in both human and mouse so that similarities and differences can be studied. In order to do this, initial verification of the quality of the fit between our model and the data is required. After checking this, the data can be explored further and other analyses can be performed.

3.1 Human brain

3.1.1 Correlation, Variance & Factor Characterisation

A good model fit should not show excessive covariance between the different identified factors. Too much correlation would indicate immoderate similarity between factors, signifying that the explained variance will not significantly increase with the number of factors. The factors in the trained model did not show excessive correlation, meaning that the fit is good in this aspect (Fig. S1). Another component that needs to be checked to verify this is the variance explained by the model over the different views. The RNA data appeared to explain more than half of the variance (~50%), followed by ATAC-distal (~10%) and ATAC-promoter (<10%). When looking at the individual factors that contribute to the variance, it can be observed that Factor 1 explains the most over all views and therefore is the most interesting Factor to pursue analysis of (Fig. S2 & Fig. S3).

The next required step is to visualise the characterisation of the individual factors: their composition and what they associate with. Technical artefacts, for example, can be discovered by viewing the factors against the covariates (Fig. S4). Besides that, the MOFA analysis can be used to show if any of the factors account for the molecular variation in the data (e.g. can a factor distinguish different cell types, Fig. S5). Within one factor we can then look at the top contributing features to it; this can be done in an integrative manner by looking at separate views for the same factor. Identification of different features shows us the constitution of each factor. For example, in the RNA view of Factor 1, it can be observed that NRG3 has the highest positive weight in this factor (Fig. S6 - Fig.S8)

3.1.2 Gene Set Enrichment Analysis (GSEA)

After confirming that the model was satisfactory and prominent factors have been identified, a GSEA was performed to link the factors to genesets involved in specific pathways. For the human data, it was observed that the most positively enriched pathways were centred around important brain GO-terms (Ashburner et Al.) such as neuron development, neurogenesis and neuron differentiation

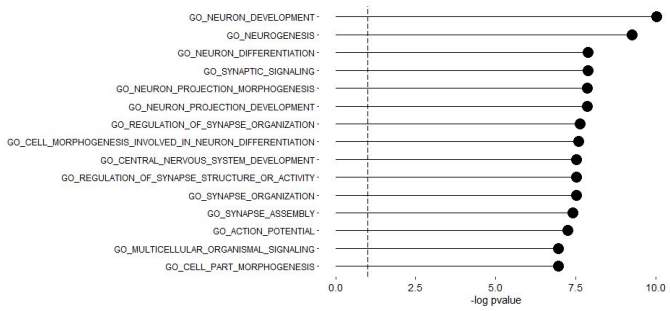


Fig. 1. GSEA: GO pathways for factor 1. This plot shows the positively enriched pathways for the first factor in order of significance.

The genes corresponding to each of these terms can also be visualised ordered by their weight (Fig. S11). The same analysis was performed for negatively enriched genes as well.

3.1.3 Motif Enrichment

The motifs to which transcription factors can bind can also be elucidated with the MOFA analysis. This analysis is similar to a gene enrichment, but pertains to epigenomic instead of genomic data.

The variance shows that the ATAC-distal data covers more variance than the ATAC-promoter, so the motif enrichment was performed on the distal peak data.

The results are realistic and expected. For example, the first motif corresponds to MAZ; inspection of gene function in UniProt shows that it encodes a Zinc-finger with specificity in just three tissues, these being kidney, liver and brain. ((Fig. S12)

The same was done for negatively enriched motifs: searching for the first motif, SOX13, provides the information that this transcription factor may be involved in the differentiation of oligodendroglia. (Fig. S13)

Overall, these results indicate close association with the brain or related processes, which is to be expected and thereby shows that the MOFA analysis was successful.

3.2 Mouse brain

3.2.1 Correlation & Variance

As previously mentioned, our factors should not display excessive correlation with each other. While some level of correlation was observed between certain factors, none of the observed correlations appear to be largely significant, indicating a satisfactory model fit ((Fig. S14).

Regarding variance decomposition, Factor 1 explains the highest amount of variability in the data, receiving input from all our sources of data (RNA and ATAC distal and promoter) (Fig. S15). For this reason, much like for our human model, Factor 1 will be considered our primary function in each step of analysis. Factor 2 receives a strong input from ATAC distal and a stronger input from ATAC promoter, and each subsequent Factor receives a gradually lower signal primarily from RNA-seq data. Overall, most of the variance explained (~30%) is explained by RNA-seq data, whereas around ~7% of the variance is explained by ATAC-seq data in distal regions, and ~5% is explained by ATAC-seq data in promoter regions (Fig. S16).

Again, we visualize individual factors' characteristics. We can compare factors against the covariates, which will uncover some technical artefacts. In this case, a strong association can be observed between Factor 2 and the number of expressed genes per cell (Fig. S17). We can also observe the contribution of various cell types to our Factor (Fig. S18), which we will look into later on.

Visualizing RNA feature weights provides a set of features according to their association with each factor. Performing this analysis with Factor 1 will provide GM29260 and Zbtb20 as strong positive features, and Tenm2, Dlg2 and Ptprd as some strong negative features (Fig. S19). Inspection of ATAC signature weights and patterns can also be performed through use of certain plots and heatmaps, though the results are difficult to interpret (we will need to resort to motif signatures) (Fig. S20-21).

With non-linear dimensionality reduction, such as UMAP, we can effectively separate results into cellular clusters. Examination of the contributions of each Factor to each cluster of cells can provide valuable insight into each factor. For instance, Factors 1 and 3 contribute most strongly to astrocyte cell types, whereas Factor 2 contributes the most to spiral ganglion neurons (Fig. S23). Results using only the RNA-Seq data or ATAC-seq data yield generally similar results. (Fig. S22)

3.2.2 Gene Set Enrichment Analysis

Fig. 4. GSEA: GO pathways for factor 1. This plot shows the positively enriched pathways for the first factor in order of significance.

Proceeding on to Gene Set Enrichment Analysis having a firm grasp of the relevant factors and their constituents, we can link these to GO pathways. Beginning by enriched pathways in Factor 1, the most significantly enriched pathway is for startle response, with a few other enriched pathways being for neurogenesis, glutamate secretion,…

Negative GSEAs do tend to indicate similar GOs, such as neurogenesis and regulation of nervous system development. As these cells originate from the brain, these results are fully expected and also serve as evidence supporting the validity of the model.

3.2.3 Motif Enrichment

With regards to motif enrichment, the ATAC-distal data was, again, responsible for more variance than the promoter data, hence the usage of distal peak data for motif enrichment.

Taking a cursory glance at these motifs, we can see that Hic1 (the top motif of positive Factor values) is a transcriptional repressor involved in the P53 pathway and is involved in the development of the head, face, limbs and ventral body wall. Atoh4 (in the negative Factor values) is a transcriptional regulator involved in the differentiation of neural cells. These results are consistent with the cortical origins of the cells (Fig.S25-S26).

4 Discussion

The MOFA analysis has revealed the importance of scRNA-seq data in both human and mouse datasets. It explains most of the variance and it is highly correlated with the most important factors. These inferred factors are poorly correlated between themselves, meaning that the orthogonality constraint, shared with PCA, is very well maintained).

The downstream analysis performed on the inferred latent factors revealed important enriched gene sets, mostly related to neuron development, neurogenesis and neuron differentiation in human cells, and to startle response, neurogenesis and glutamate secretion in mouse cells. Notable genes or gene families that have significant involvement in Factor 1 of both models include SOX5 and SOX6, NRXN1 (present in both models, NRXN3 also present in the mouse model), NRG1 and NRG3, NTRK2, Robo1 and Robo2 among others. This means that the most important gene sets captured by the latent factors are those ensuring correct genesis and development of the brain tissue (more represented in the human dataset), and those that maintain correct functioning of the brain, favouring sensitivity and responsiveness of neurons (more represented in the mouse dataset).

A number of motifs were enriched in both datasets and were also well captured by the inferred latent factors. The ATAC-distal part has been capable of explaining more variance than its ATAC-promoter counterpart, however, the latter is not negligible and has also proven valuable in this study. Most of the motifs found possess a known and very important biological role (Hic1, Atoh4 or SOX13) and thus are useful to validate the trained MOFA models, considering them a sort of quality control.

5 Conclusions

- MOFA has proven to be a very valuable tool in the analysis of multimodal biological data. The models can be trained quickly, are scalable, directly interpretable and provide useful insight into the inner mechanisms of the system at hand. The trained models have been capable of integrating and learning a low-dimensional and useful representation of the multi-omics data for both human and mouse.
- The scRNA-seq data explains most of the variance of the multi-omics dataset, but the information provided by the ATAC-seq data is not negligible and can be integrated very well.
- The biomarkers, enriched gene sets and motifs found are very relevant and play a key part in the correct development and functioning of both the human and mouse brain. They have also been useful for inferring the cell types involved in these processes.

6 Data availability

Both datasets of the human data (3K) and mouse data (5K) can be found using the following links, for mouse and human respectively. The data is available both in raw fastq format or in CellRanger output format, that contains processed count matrices as well as peak calling data. The analysis pipeline that can be found in the Code Availability section starts using the CellRanger output.

7 Code availability

The entire analysis pipeline used in this paper can be found following this github link. For more information you can read the README.md file.

Acknowledgements

We thank Mr. De Winter and M(r)s. Floc'hlay for their help and guidance in bringing this project to a successful conclusion.

Conflict of Interests:

The authors declare to have no conflicts of interest.

References

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, *21*(1), 1-17. DOI for nicer reference: 10.1186/s13059-020-02015-1

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., ... & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, *14*(6), e8124.

8 Appendix

8.1 Human data

8.1.1 Correlation, variance & factor characterisation

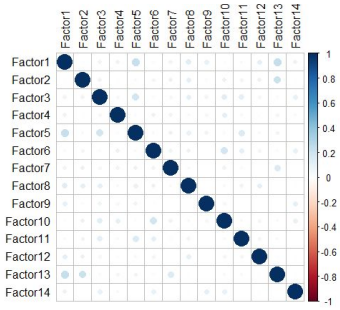


Fig. S1. Correlation plot between factors. Factors should not be overly correlated to each other, this plot indicates that that is indeed not the case.

Fig. S2. Variance Decomposition of factors. The plot shows the contribution of the separate factors to the view. Factor 1 is important in each view and therefore interesting to study. Factor 2 is also a good contributor for the RNA view and in a lesser extent to the ATAC views. Factor 3 has no contribution in the RNA view but does have more contribution to the ATAC views than factor 2. These 3 factors are therefore the main factors that this analysis focuses on.

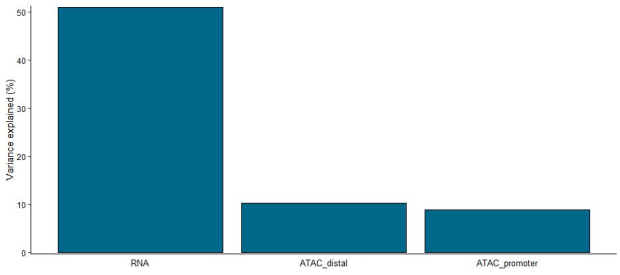


Fig. S3. Total variance. The RNA data explains most of the variance, followed by ATAC-distal and at last ATAC-promoter.

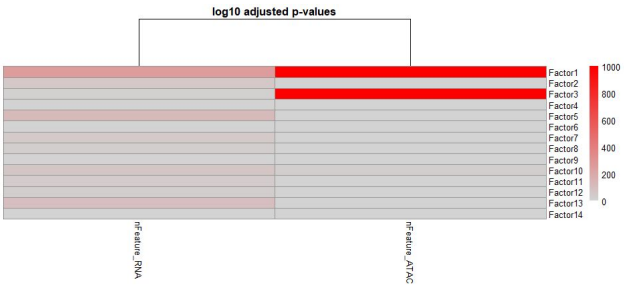


Fig. S4. Association between factors and covariates. Factors 1 and 3 are strongly associated with the peak number in the ATAC view. This is possibly due to technical issues and could be solved by applying a better or different normalisation.

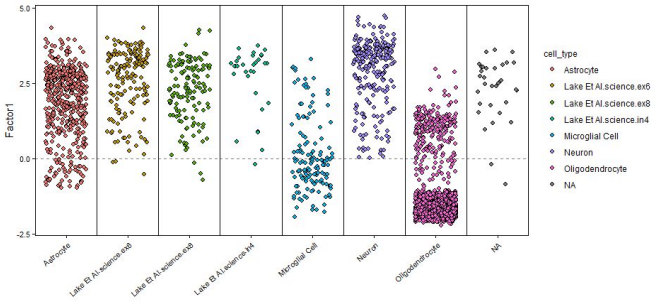


Fig. S5. Visualisation of factor values grouped by cell-type for factor 1. Factor 1 is able to differentiate a big portion of the oligodendrocytes from other cell types (purple scatter in the previous to last column).

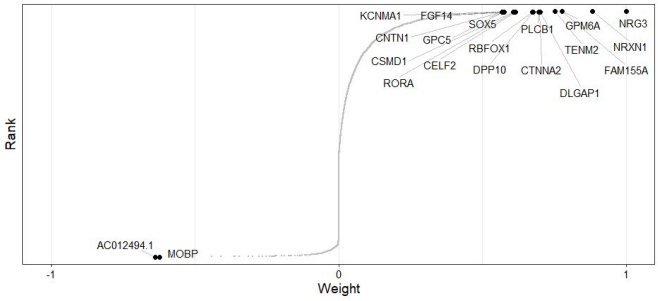


Fig. S6. Visualisation of the contribution of top feature weights in RNA view for factor 1. Positive weights are on the upper right and negative ones on the bottom left.

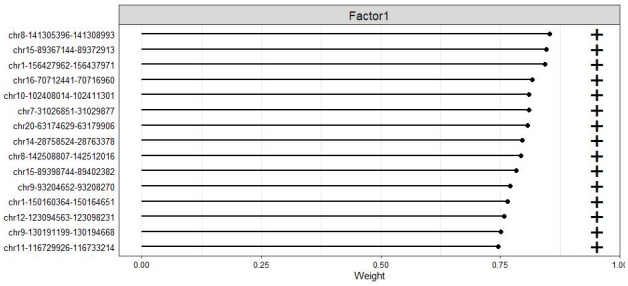


Fig. S7. Visualisation of the contribution of top feature weights in ATAC-distal view for factor 1. This plot has the same concept as the previous graph, the weights are only shown as a list instead of a graph. The sign on the right indicates whether the contribution is positive or negative.

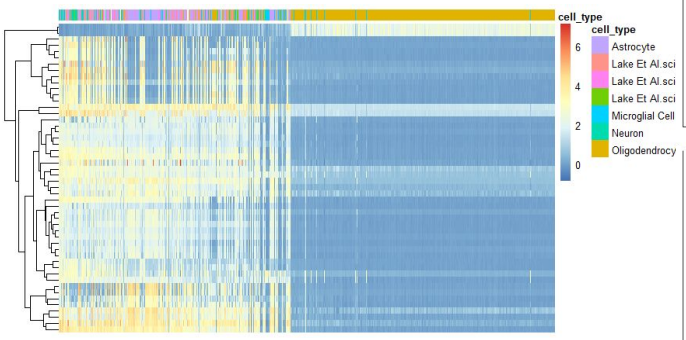


Fig. S8. Heatmap of first 50 features for factor 1 in RNA view. The heatmap is showing a clear difference between features for oligodendrocytes (which make up a big part of the data) and other cell types.

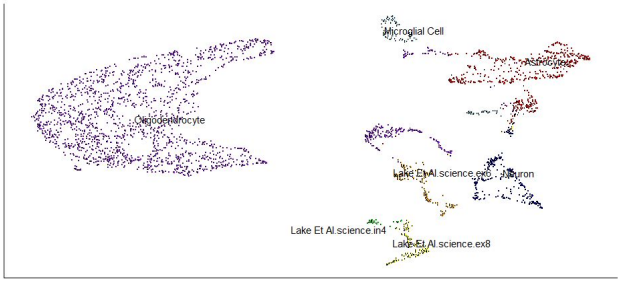


Fig. S9. UMAP of the MOFA factors. UMAP plots (human dataset), high dimensional data is represented and structure is kept in a low-dimensional way. In this graph the MOFA factors were plotted and reveal the clustering of the cell types based on the factors.

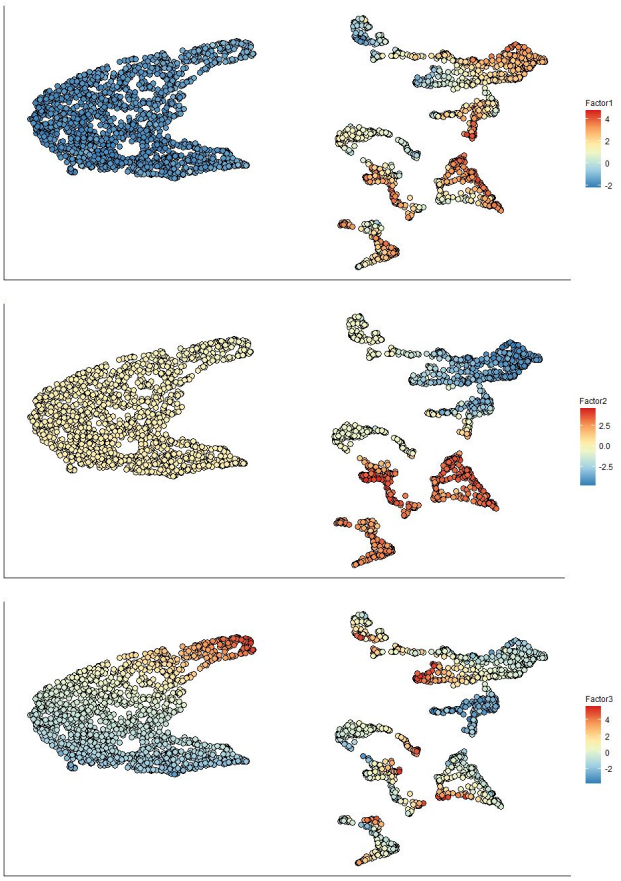


Fig. S9. Contribution of the different factors to the UMAP. *Top:* Factor 1, *Middle:* Factor 2, *Bottom:* Factor 3. Factor 1 contributes in a negative way to oligodendrocytes and microglial cells, the effect of factor 2 is neutral on oligodendrocytes and negative on astrocytes while factor 3 is showing mixed results. For the other cell types, the effects are mixed within each group.

Fig. S10. Contribution of the separated view factors. *Top:* Factor 1 in RNA view, *Bottom:* Factor 1 in ATAC-distal view. These plots are the breakdown of Fig. S9 into the different contributing views. When putting them together, we should therefore approximate this figure.

8.1.2 GSEA

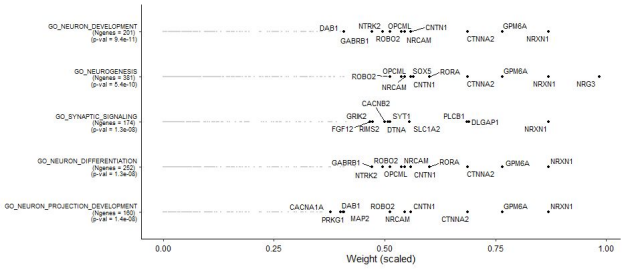


Fig. S11. Contribution of features towards positively enriched GO pathways of factor 1. This plot shows the top contributing features for the different positively enriched GO pathways.

8.2 Mouse data

8.1.1 Correlation, variance & factor characterisation

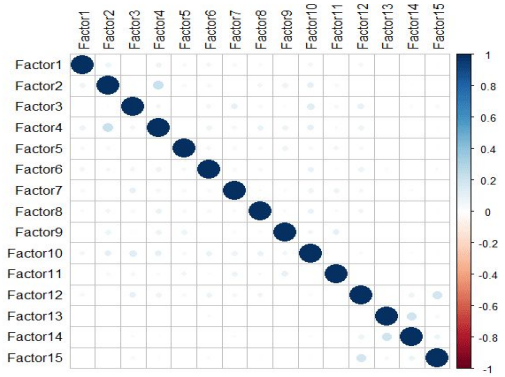


Fig. S14. Correlation plot between factors. Factors should not be overly correlated to each other, this plot indicates that that is indeed not the case.

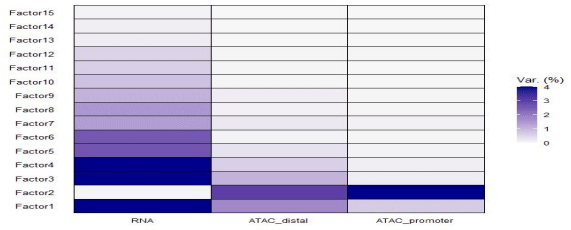


Fig. S15. Variance Decomposition of factors. The plot shows the contribution of the separate factors to the view. Factor 1 is important in each view and therefore interesting to study. Factor 2 a good contributor for the ATAC views but not the RNA views. Factors 3 and 4 have strong contributions for the RNA views and lesser contributions to the ATAC views. The first 3 factors will be the main subjects of the analysis.

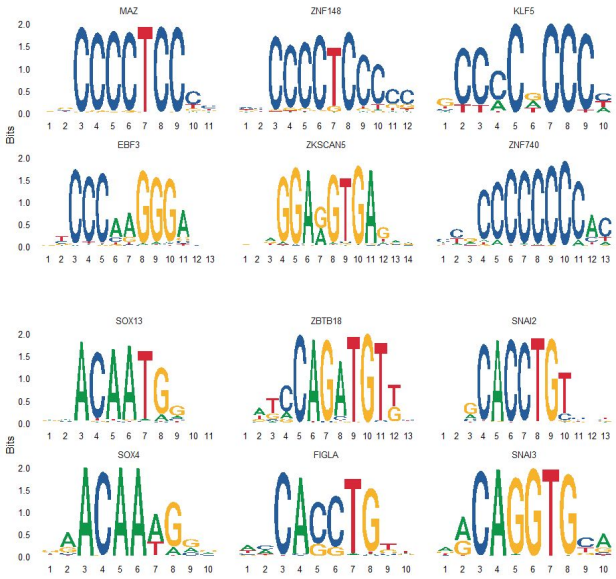


Fig. S12 (top) and S13(Bottom). Positively (S12) and negatively (S13) enriched motifs for factor 1. These show the 6 most significant positively and negatively enriched motifs for Factor 1.

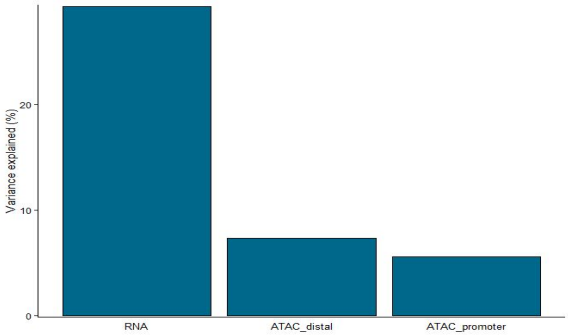


Fig. S16. Total variance. The RNA data explains most of the variance, followed by ATAC-distal and at last ATAC-promoter.

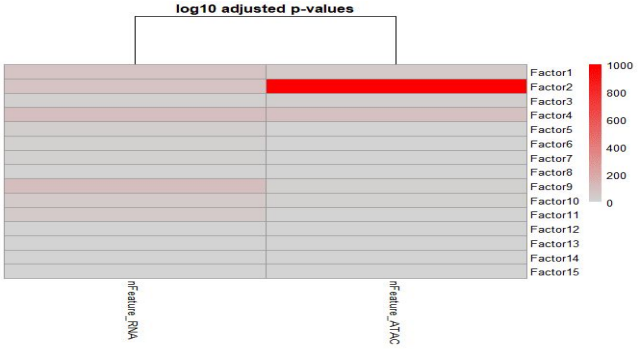


Fig. S17. Association between factors and covariates. Factor 2 is strongly associated with the peak number in the ATAC view, possibly because of a technical issue solved by normalization.

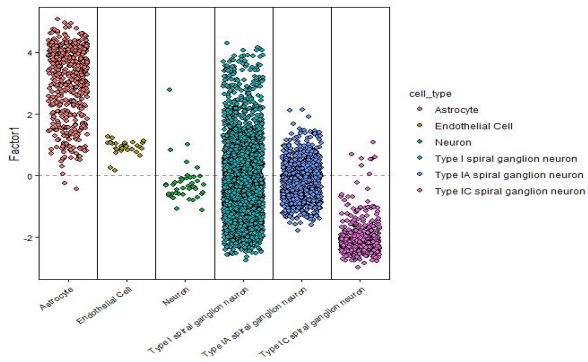


Fig. S18. Visualisation of factor values grouped by cell-type for factor 1. Factor 1 differentiates mostly spiral ganglion neurons, particularly Type 1 spiral ganglion neurons.

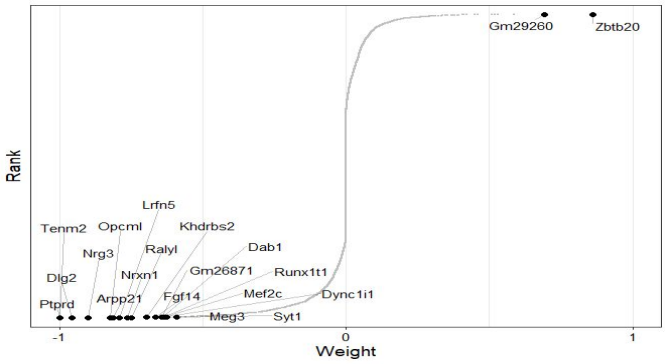


Fig. S19. Visualisation of the contribution of top feature weights in RNA view for factor 1. Positive weights are on the upper right and negative ones on the bottom left.

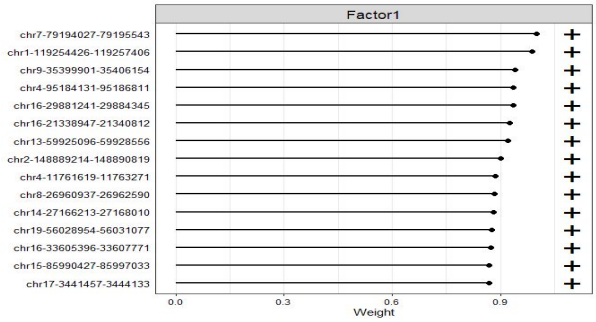


Fig. S20. Visualisation of the contribution of top feature weights in ATAC-distal view for factor 1. This plot has the same concept as the previous graph, the weights are only shown as a list instead of a graph. The sign on the right indicates whether the contribution is positive or negative. Not very easy to interpret.

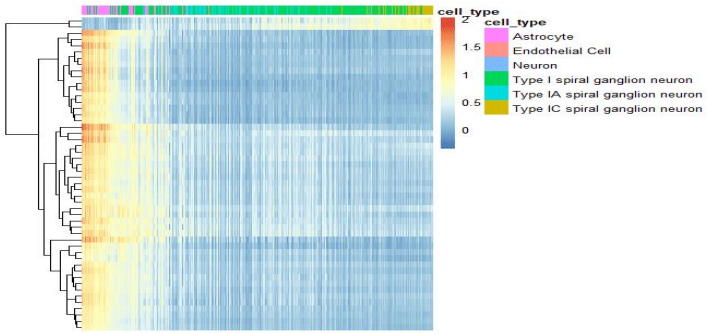


Fig. S21. Heatmap of first 50 features for factor 1 in RNA view. The heatmap is showing a clear difference between features for spiral ganglion neurons (particularly of Type I) and other cell types.

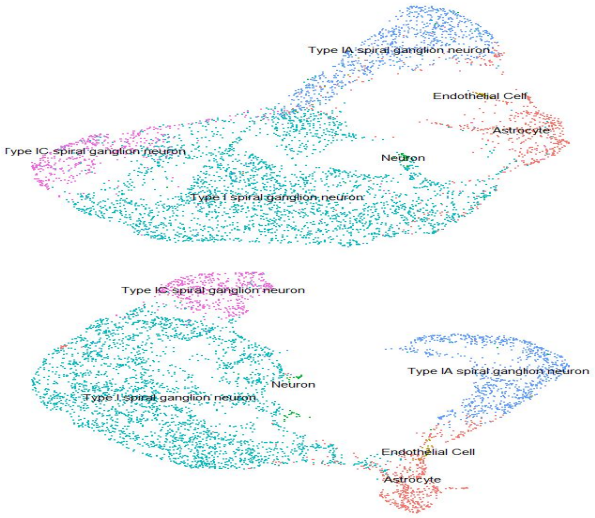
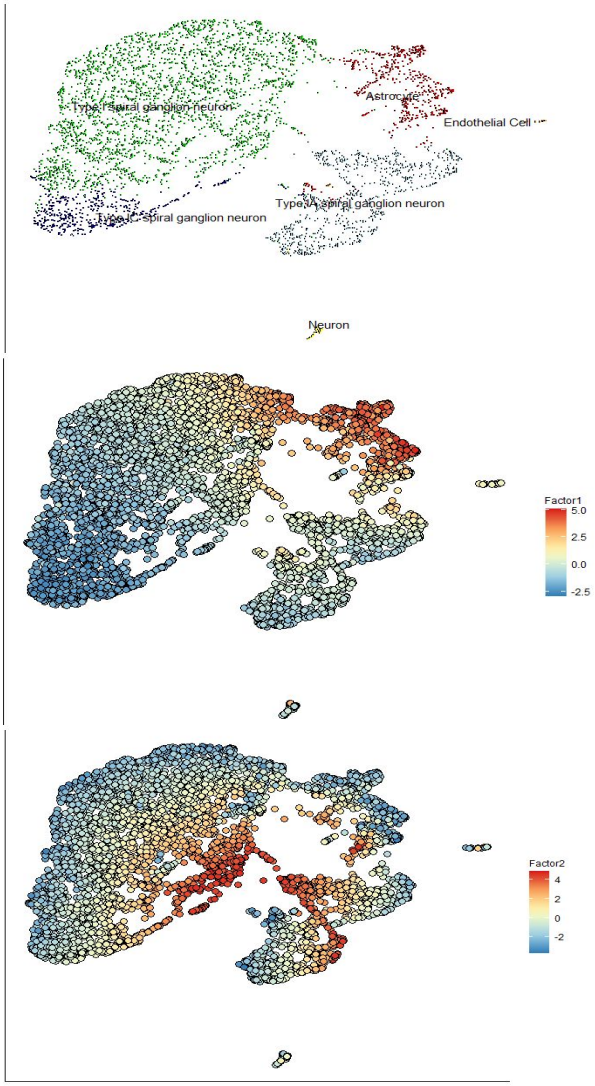


Fig. S22. Contribution of the separated view factors. *Top:* Factor 1 in RNA view, *Bottom:* Factor 1 in ATAC-distal view.



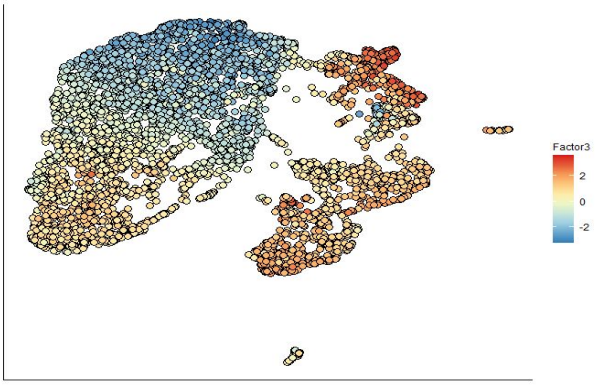


Fig. S23. Contribution of the different factors to the UMAP. *Top:* Factor 1, *Middle:* Factor 2, *Bottom:* Factor 3. Factor 1 contributes in a negative way to Type IC spinal ganglion neurons and positively to astrocytes, the effect of factor 2 is strongly positive on some Type 1 and Type 1A spiral ganglia neurons and neutral to negative on other cell types, while factor 3 is generally positive for astrocytes, generally negative for Type I spiral ganglion neurons and mildly positive for all other cell types.

8.2.2 GSEA

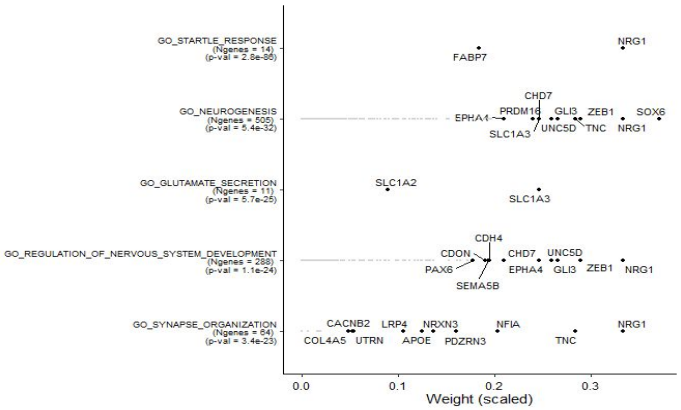


Fig. S24. GSEA: GO pathways for factor 1. This plot shows the positively enriched pathways for the first factor in order of significance. Startle response is by far the highest, followed by a number of expected processes within the brain.

8.2.3 Motif Enrichment

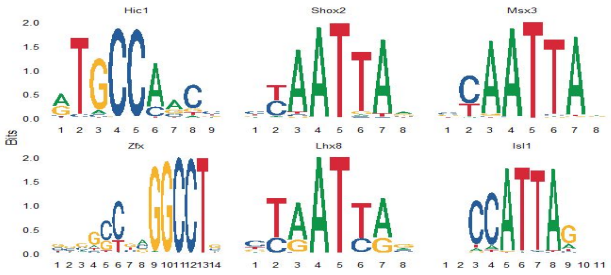


Fig. S25. Motif enrichment: Positively enriched motifs for factor 1. This plot shows the 6 most significant positively enriched motifs for the first factor.

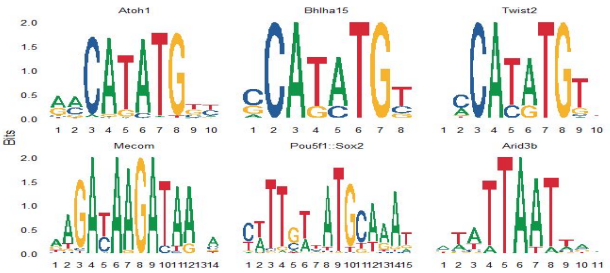


Fig. S26. Motif enrichment: Negatively enriched motifs for factor 1. This plot shows the 6 most significant negatively enriched motifs for the first factor.