Integration of single-cell transcriptomic atlases of the adult human midbrain

Chen Ji Rong Jiang
Supervised by: Dr. Onur Basak

Abstract

The cellular composition of the human brain has shown to be more heterogeneous than previously understood. The ventral midbrain is no exception. Advances in single-cell technologies has enable the creation of high-resolution transcriptomic datasets. However, current atlases lack representation of rare cell types, obscuring their distinction from noise in subsequent analyses. In this thesis, we present an integrated midbrain reference atlas that combines data from two midbrain atlases, three substantia nigra (SN) datasets and one ventral tegmental area (VTA) atlas. The additional data enhances the sensitivity of the dataset, allowing for more refined subtyping of the (ventral) midbrain. Additionally, we optimized the integration pipeline with harmonized realignments of the original snRNA reads, benchmarks of integration methods such as scVI, Harmony and Scanorama using batch correction metrics, and hyperparameter tuning on scVI. By investigating the hierarchical relationships between previously identified subtypes, we found agreements across datasets. Notably, we provided evidence for VTA-related combinatorial subtypes that express both gamma-aminobutyric acid (GABA) and dopamine neurotransmitters. Our findings support the existence of specialized subtypes that can be efficiently discerned through region specific atlases. We anticipate this reference genome to serve as a cornerstone for future research on the human midbrain and to expand over time with the inclusion of additional atlases.

Layman's Summary

The midbrain is a region in the brain that plays an essential role in various functions, from cognition to emotion. Its functional diversity is evident in the many different cell types that can be found in this region. In recent research, it has been shown that the previously identified cell types can be further divided into subtypes. However, the exact composition within the midbrain is not known yet. The transcriptome is the collection of genetic products within a cell. Datasets have been made by detecting these transcriptomes within the midbrain in order to investigate the composition within this region. A major obstacle to identifying the different subtypes is that these datasets are often not large enough to detect sparse subtypes with confidence. In this project, we combined multiple midbrain datasets in order to gain more data and determine the various subtypes within the midbrain. A problem in this process is technical variation in the way these datasets were formed, which influences the transcriptome data, often referred to as batch effect. Computational integration tools can be used to minimize these batch effects. We evaluated three different integration methods and determined that scVI works the best for our specific problem. Furthermore, we tuned and assessed different parameters of scVI to optimize this procedure. After combining the different datasets and controlling for batch effects, we used a computational tool called scHPL to form a hierarchy of previously discovered subtypes from the original datasets. This is necessary as each of these datasets has its own naming conventions. As a result, we can compare which subtypes are identical to each other and which can be further divided. This way, knowledge from the original datasets is combined, revealing a refined set of unique subtypes.

Additionally, we analyzed the results, specifically focusing on the ventral tegmental area (VTA). The VTA is an important region within the midbrain, but is often underrepresented within datasets. Our in-house dataset within this project was specifcally sampled from the VTA, allowing us to investigate its subtypes. Using the combined dataset and its hierarchy tree, we discovered evidence for unique subtypes only present in the VTA.

Overall, this project generated a reference atlas that elucidates the molecular composition of the midbrain, opening many more avenues of research and support the existence of VTA-specific subtypes.

Introduction

The ventral tegmental area (VTA) is a region located on the floor of the midbrain and has been associated with several psychiatric and movement disorders such as Parkinson's disease (PD).[1] The region mostly contains dopaminergic (DA) neurons which play an essential role in cognition, the reward system, and the regulation of emotion, motivation and behavior through the mesolimbic and mesocortical pathways.[2–4] The VTA further consists of GABAergic neurons that are involved in the reward system, aversion, and stress through inhibition of neighboring and distant DA neurons as well as excitatory glutamatergic neurons, that are associated with (defensive) behavior and reward.[3–7] In addition to these cell types, the existence of combinatorial neurons that express multiple neurotransmitters has been shown.[8, 9] Furthermore, the diverse functionality of the VTA is reflected in the heterogenous gene expression profiles, electrophysiological and molecular properties; projection patterns, and disease vulnerabilities of individual neurons within the VTA, indicating a more elaborate and complex collection of cell subtypes.[4, 10, 11] A critical step in understanding the underlying mechanism behind the VTA's functional pathways is the unbiased determination of its agents in higher resolution through single-cell and single-nuclei sc/snRNA sequencing analysis. Typically, in such an analysis, expression matrices are formed of individual cell transcriptomes. Subtypes are characterized by their specific transcriptomic profiles and can be visualized by dimensional reduction and clustering techniques such as TSNE and UMAP.

Recent advances in sc/snRNA sequencing technology have allowed cell type profiling at unprecedented throughput and resolution, resulting in several molecular atlases of adult human and mouse midbrain structures. This was initiated by La Manno et al. that identified subclasses of dopaminergic cells in the dopaminergic midbrain structures; the ventral tegmental area (VTA) and the substantia nigra (SN), by applying SMART-seq single-cell RNA sequencing (scRNA-seq).[12] Since then, whole brain atlases of adolescent[13] and adult mice[14, 15] as well as atlases of mouse dopaminergic neurons[16] have extended the list. Very recently, two comprehensive atlases of the whole adult mouse brain have been published that elucidate the heterogeneity within the midbrain at even higher resolution and scale as well as integrated with spatial information.[17, 18] Similarly, multiple transcriptomic atlases of the human midbrain have been published. Four atlases focused on the SN and one on the whole midbrain also show the heterogenenity within the midbrain and reveal various neurological disorder-associated transcriptomic profiles and cell subtypes.[19–23] More recently, Siletti et al. created the first single-cell atlas of the whole adult human brain, resulting in more than 3 million cells including 300,000 cells from the midbrain and allowing subtyping at even higher resolution.[24]

A common obstacle of these atlases is the undersampling of rare subtypes. Even within the most comprehensive atlas of the adult human brain, a supercluster was formed of suspected undersampled neurons, the splatter neurons. [24] To overcome this problem, large-scale integrations of multiple atlases have to be done in order to achieve a higher sensitivity that allows the detection of these subtypes. Steuernagel's Hypomap shows the application and value of such measures on hypothalamus atlases by characterizing type 2 diabetes-related neurons.[25] However, the integration of multiple atlases suffers greatly from batch effects due to experimental and technical differences. [26] Harmonized mapping of the raw sequencing data as well as proper usage of batch correction methods are essential to the comparability of multiple datasets. Tran et al., Luecken et al. and Steuernagel et al. have benchmarked many batch correction approaches and have shown several top-performing methods, but their rankings vary depending on the specific conditions of the integration such as the tissue type, cell count, and batch complexity. [25, 27, 28] Examples of such top-performing methods are Harmony, Scanorama, scVI, and scANVI.[29–32] Harmony performs better on simpler integration problems, where the datasets contain the same set of cell types. In contrast, scVI and Scanorama excel at the integration of more complex problems, where the datasets contain nonidentical biological variation. scANVI seems to perform well overall, but does require ground truth cell type labels to drive the integration, which is rarely readily available. Evaluating such integration methods is a challenge on its own. Various metrics have been proposed to assess an integration's batch-correction and bio-conservation performance such as k-nearest-neighbor batch-effect test (kBET) and Random Forest-mixing.[25, 33]

Most annotations within these datasets are created by performing dimensional reduction and clustering techniques such as PCA in combination with UMAP and Leiden clustering. This results in a highly subjective labeling that is dependent on the cells and resolution of the specific atlas. The lack of standardized labeling results in a high output of unique subtypes that are incomparable across atlases and hinder the assessment of cellular compositions. Michielsen et al. present a hierarchical progressive learning (scHPL) method that allows the continuous learning of a classification tree on diverse annotation resolutions from multiple datasets. This enables the identification of hierarchical relationships of labels across atlases.

In this thesis, we realigned and integrated one in-house ventral tegmental area (VTA), three substantia nigra (SN), and two whole midbrain human snRNA atlases in order to elucidate the cellular composition

of the VTA. Additionally, we showed the importance of a harmonized realignment procedure, evaluated various integration methods, and performed hyperparameter tuning to optimize the integration process. Finally, we determined the hierarchical relationship between previously identified neuron, oligodendrocyte and astrocyte subtypes, and found evidence for the existence of VTA-specific neuronal subtypes.

Results

Realignment validation

In order to determine if the technical variation in the alignment process causes problematic downstream batch effects, we performed a preliminary comparison of the scRNA mouse data from Zhong et al. The atlas was originally aligned to the Ensemble 93 mouse reference genome with the 10X Genomics alignment software, Cell Ranger v3.0. We aligned the same data to the Ensemble 98 reference genome with Cell Ranger v6.1. We selected these parameter values based on the alignment of our in-house mouse VTA dataset. After preprocessing, quality control, and normalization, the two alignments were concatenated, and clustering was performed. This process resulted in well-mixed clusters for the oligodendrocytes, however, a clear distinction between the two alignments was visible for the remaining cells, such as the post-mitotic neurons, indicating strong batch effects. (See figure S1)

Following the preliminary results, we tested the reference genome and software version separately to discern which technical variation drives the batch effects. Furthermore, we included intronic region mapping in the comparison, a parameter that has been confirmed by 10X Genomics to capture additional real biological information.[34] Without the `include-introns` parameter, reads from the dataset are exclusively mapped towards the exons of a gene. Enabling this feature allows the mapping towards the introns, which accounts for about 20 to 40% of the total reads in tests conducted by 10X Genomics.[34] We also included our in-house mouse VTA dataset to investigate the effect size in the context of a second dataset. For these tests, scRNA sequencing data was taken from 5000 and 5500 healthy mouse cells from our in-house mouse VTA dataset, made by Hey et al., and the midbrain dataset from Zhong et al., respectively. These two datasets were aligned using Cell Ranger v6.1 to its complementary mouse reference genome, mm10 (Ensemble 98), without mapping to intronic regions as a control. Then we compared the following alignment parameters with the identical datasets: 1) Cell Ranger v3.0 instead of v6.1 as alignment software version; 2) reference genome Ensemble 93 instead of the newer Ensemble 98; or 3) using the `include-introns` parameter instead of excluding it. For the reference genome and intronic mapping comparisons, the alignments cluster more based on their technical similarity rather than their identical counterparts. (See figure 1a-i) While the batch effects seem undetectable in the software version comparison. (See figure 1d) In order to quantify these batch effects, we calculated the pairwise distances between the identical cells from the two alignments, including all genes. (See figure 1) Of the three parameters, intronic mapping caused the largest distances with a median of 33.25. Interestingly, the median deviation caused by the software version (5.54) is higher than that of the reference genome (3.09), even though the UMAP does not indicate this. These findings suggest that the software version's batch effect does not impact the genes responsible for the primary cell-to-cell variation within the midbrain or VTA. However, it is plausible that these effects are more pronounced in other tissue types. Furthermore, filtering on highly variable genes (HVGs) negates the batch effect and results in the cells clustering towards their corresponding cells for the reference genome and software version parameters, but not for the intronic mapping. (see figure S2-S4) This suggests that the batch effect of these parameters primarily influences low-variance genes that are not the primary drivers of the cell-to-cell variation. In contrast, the batch effect caused by the intronic mapping parameter is mostly retained even after filtering for HVGs.

Siletti et al. used STARSolo v2.7.10a and a customized Ensemble 109 reference genome for their alignment. [24] In order to confirm the necessity of realignment for the Siletti dataset, similar tests were performed on these two parameters: 1) Cell Ranger v6.1 against STARSolo v2.7.10a and 2) the Ensemble 109 against the customized Ensemble 109. STARsolo produces comparable results as Cell Ranger in the UMAP, while the custom reference genome induces significant batch effects. (See figure 1j-o) Distance calculations of the software comparison indicate higher batch effects than is visible in the UMAP, similar to the software version comparison. In contrast, the customized reference genome induced extreme deviation for a subset of cells. Further investigation, shows these deviations are lowest in oligodendrocytes. Overall, our findings indicate that technical variation in alignment parameters introduces batch effects in downstream analysis. The differences in the `include-introns` parameter cause the largest batch effect in midbrain and VTA datasets, followed by the reference genome and software (version).

The original atlases were aligned with five different reference genomes and software (versions). Only one atlas used the `include-introns` parameter, but three other atlases used an alternative method

to include intronic reads. Furthermore, of the five reference genomes, two were customized. (See supplementary table 1) Given the technical variation in the original atlases, we decided to perform a harmonized realignment using the human reference genome Ensemble 109, Cell Ranger v6.1, and including the `include-introns` parameter.

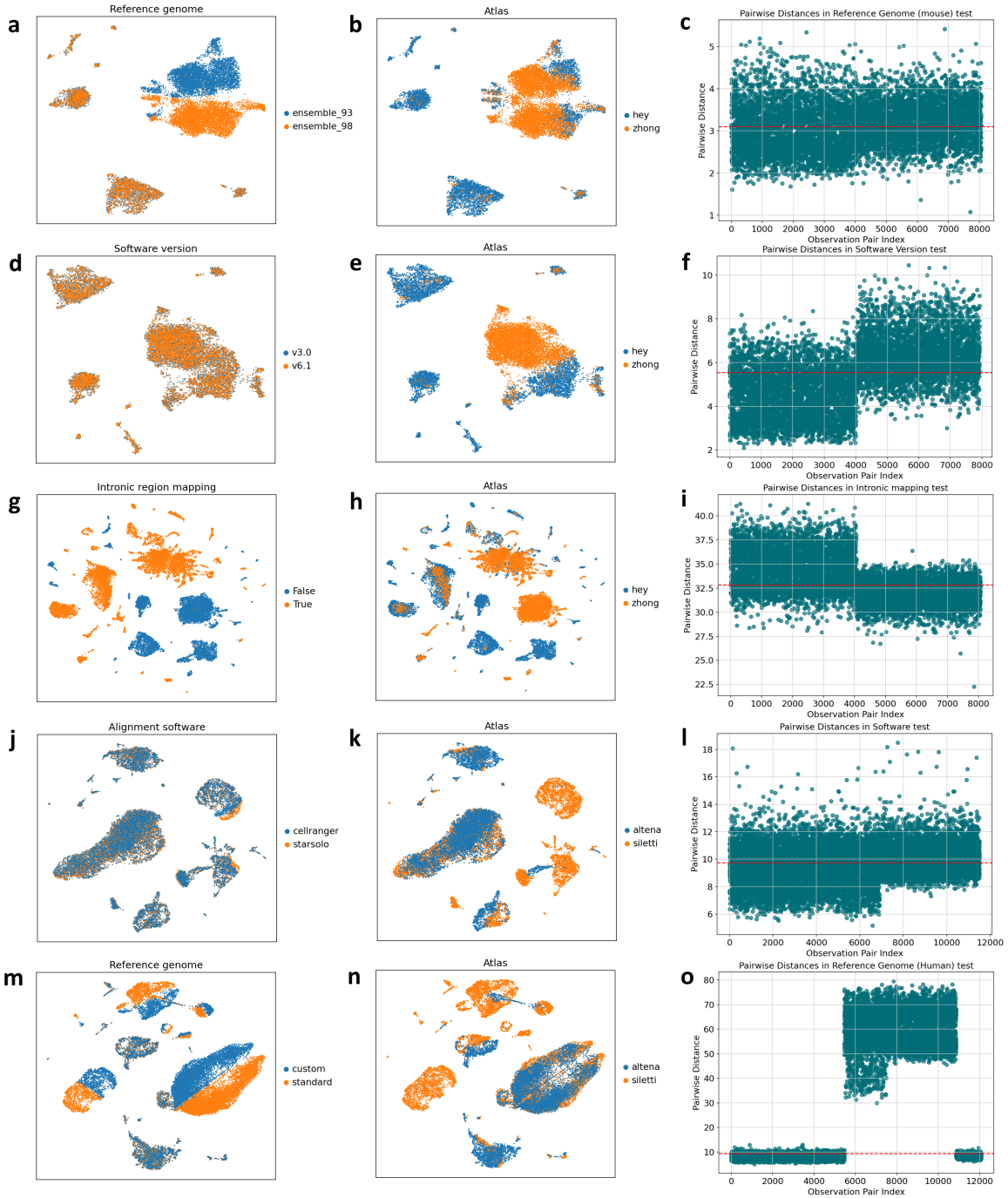

Figure 1: Alignment parameter comparison
a,d,g,j,m) UMAP of the parameter value comparison with colors representing the exact same dataset, but with different parameter values. More overlap suggests less batch effects. b,e,h,k,n) The same UMAP with color representing the two atlases used. c,f,i,l,o) Pairwise distance calculations between the two alignments. Higher values suggest stronger batch effects and median values are indicated with a red line. Ideal values would be 0. a-c) Comparison between the reference genome Ensemble 93 and 98. d-f) Comparison between software version Cell Ranger v6.1 and v3.0. g-i) Comparison between inclusion and exclusion of intronically mapped reads. j-l) Comparison between Cell Ranger v6.1 and STARSolo v2.7.10a. m-o) Comparison between reference genome Ensemble 109 and a custome version of Ensemble 109 used by Siletti et al.

Integration benchmarks

Initially, we assessed the performance of three integration methods – scVI, Harmony and Scanorama – using five batch correction metrics that measure the mixing of batch labels within clusters: silhouette batch, integration Local Inverse Simpson Index (iLISI), k-nearest-neighbor Batch-Effect Test (kBET), graph connectivity and Principal Component Regression (PCR) comparison, via the scib metrics package. (See figure S5a) Among these methods, scVI achieved the highest overall score (0.42), followed by Harmony (0.40) and Scanorama (0.35). In contrast, the unintegrated atlas received the lowest aggregate score (0.31). The key driver behind scVI's superior performance seems to be the PCR comparison metric (0.42). This metric assesses the extent to which the principal components (PCs), derived from the integrated latent space, improve the explained variance in the dataset compared to the unintegrated PCs in linear regression. For this metric, Harmony and Scanorama scored considerably lower (0.33 and 0.30, respectively), while the unintegrated was unable to be scored as its PCs are identical to before the integration. In the remaining metrics, all three integration methods scored similarly, except for the graph connectivity metric. While scVI, Harmony, and the unintegrated dataset had comparable scores(0.63, 0.65, and 0.64, respectively), Scanorama scored significantly lower (0.43), indicating more connected original cell labels. Therefore, we proceeded with scVI, and performed hyperparameter optimization for the final integration of all atlases.

We fine-tuned four essential hyperparameters – number of highly variable genes (HVGs), latent features, hidden layers, and hidden nodes – independently of each other. In addition to the batch correction metrics, five bio-conservation metrics were used to evaluate the hyperparameter values: Leiden Normalized Mutual Information (NMI), Leiden Adjusted Rand Index (ARI), KMeans ARI, and Silhouette label. These metrics assess the correct clustering of identical cell labels.

In order to use these metrics, we used scHPL to determine the hierarchical relationship of the lowest resolution labels, and harmonize the original cell annotations (See figure 4a). The labels were then collapsed to the first level of the tree and used for the bio-conservation metrics. Figure S5b shows the results of the hyperparameter tuning with the optimized values being 2000 HVGs, 90 latent features, 128 hidden nodes and 5 hidden layers. We proceeded with our integration using these hyperparameter values. Unless indicated otherwise, these values were used in subsequent integrations.

Reference atlas generation

To create one unified adult human midbrain atlas, we integrated three SN and two whole midbrain human snRNA atlases together with our in-house VTA-specific atlas. (See supplementary table 1) These datasets were processed through a Scanpy pipeline including quality control and normalization. The largest dataset is from Siletti et al., containing 287,523 midbrain cells post-filtering, followed by van Regteren Altena et al. (85,044 VTA cells), Welch et al. (67,467 SN cells), Wang et al. (43,477 SN cells), Smajic et al. (29,495 SN cells) and Agarwal et al. (9,214 SN cells). Overall, the integration resulted in a total of 522,220 nuclei after cell-level quality control. The integrated dataset is visualized in figure 2a-b, annotated with the atlas and original major cell type.

The harmonized alignment and quality control process retained numerous cells that were originally filtered out. This led to 121,490 unannotated cells in the integrated dataset (16.1%). (See figure 2c) Most of these cells originate from the Wang dataset as its annotation was not available at the time of this project.

Label prediction

In order to label the unannotated cells, a k-Nearest Neighbour (k-NN) model was trained on the lowest resolution labels, as they are the closest approximation to the true labels. We used the model to predict new annotations for the whole dataset including previously annotated and unannotated cells. Cells that failed to be classified were rejected, resulting in the final integrated reference dataset of 507763 nuclei. The model divided the cells into 6 superclusters, visualized using UMAP in figure 3a and S6. The majority of these cells are oligodendrocytes (45.4%) followed by neurons (31.0%), microglia (7.9%), astrocytes (7.7%), and oligodendrocyte precursor cells (6.0%). The remaining cells were classified as endothelial (1.1%), miscellaneous (0.5%), ependymal (0.1%), and pericytes (<0.1%). Figure 3b shows marker gene expression of the predicted supercluster labels.

Among these predictions, an untrivial amount of cells were reclassified. Indeed, across all datasets, previously annotated cells were given new supercluster labels, underlining the value of the realignment process. Sankey plots in figure 3c visualize the reclassification from the original labels to predicted superclusters. Overall, most previous labels align with the supercluster prediction. Figure S7a shows the proportion of atlases per supercluster label before and after the prediction. Across all atlases,

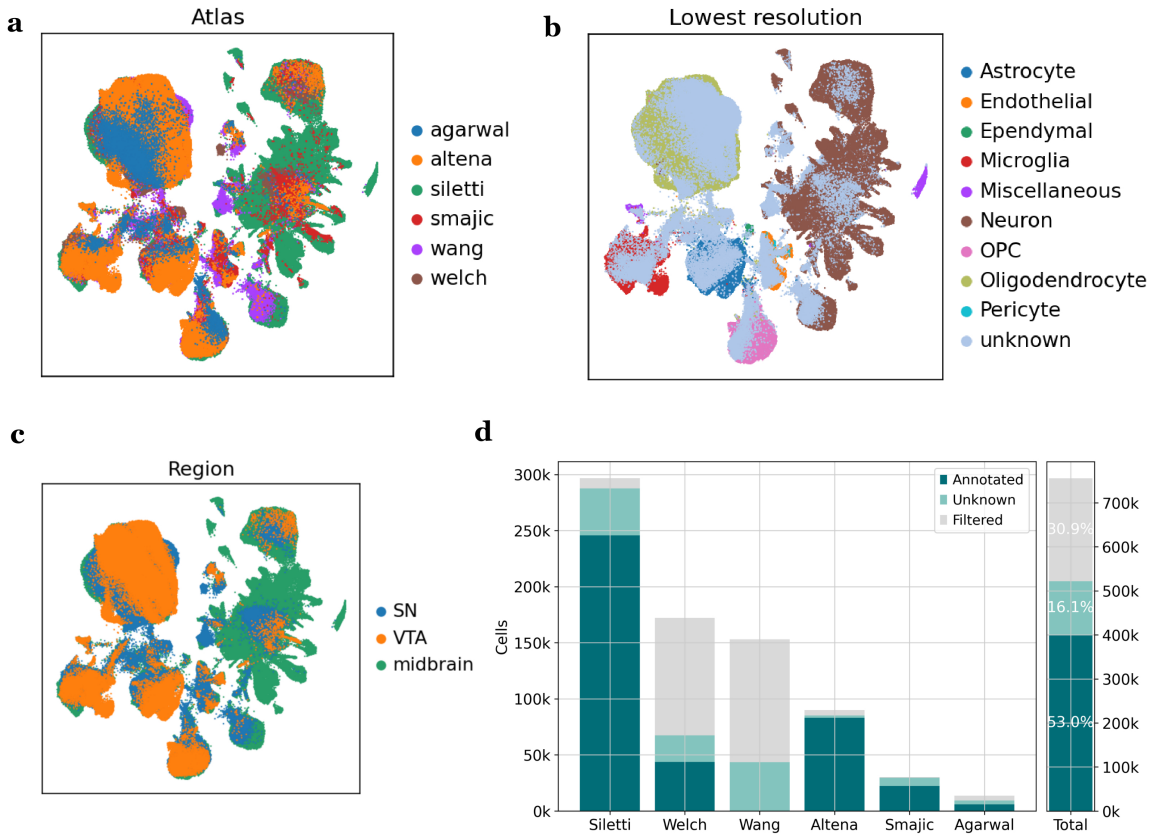

Figure 2: Integrated atlas
a) UMAP of the integrated dataset using scVI with the optimized hyperparameters, color represent the different original atlases. b) The same UMAP with colors representing the original lowest resolution labels collapsed to generic names. Note that unknown cells are also annotated. c) The same UMAP with colors representing the region of the original atlas. d) Proportions of annotated, unannotated and unknown cells in each atlas as well as the combined atlas. Note that the Wang atlas does not have original annotations available.

oligodendrocytes make up the highest partition of superclusters, except for the Siletti atlas which contained a higher proportion of neurons. Correspondingly, following the predictions, the proportion of oligodendrocytes exhibited the most significant increase, except in the case of the Siletti atlas, where neurons showed the most increase. It is possible that the prediction is influenced by the amount of supercluster-specific cells. However, this is unlikely at this resolution, given the high volume of cells. Furthermore, the proportion of unannotated cells is the highest in the Wang dataset, given its lack of available annotation. This is followed by Welch, Agarwal, Smajic, Siletti, and van Regteren Altena. The gain in new cells seems to increase with the technical differences between the original study and our process. Atlases such as Welch et al. and Agarwal et al. used older reference genomes and software, while van Regeteren Altena et al. and Siletti et al. used similar alignment parameter values. The proportion of atlases per supercluster label is visualized in figures S7b-c. The unannotated cells from the Wang atlas are distributed over the different superclusters after prediction. Overall, the atlas proportions of the prediction are very similar to the original annotations. As expected, the Siletti dataset made up the largest proportion of all predicted superclusters except for pericytes. These seem to be mostly from the Smajic dataset.

Hierarchical relationships of original subtypes

The study-specific label syntax hinders the direct comparison between datasets. With scHPL, we elucidated the relationship between these previously defined subtypes and visualized it in a hierarchical tree. (See figure 4) To get a higher resolution of the superclusters, subsets of predicted oligodendrocytes, neuronsz and astrocytes were integrated separately. From each subset, cells with original labels that align with the supercluster prediction were used to construct the hierarchy tree and labels with less

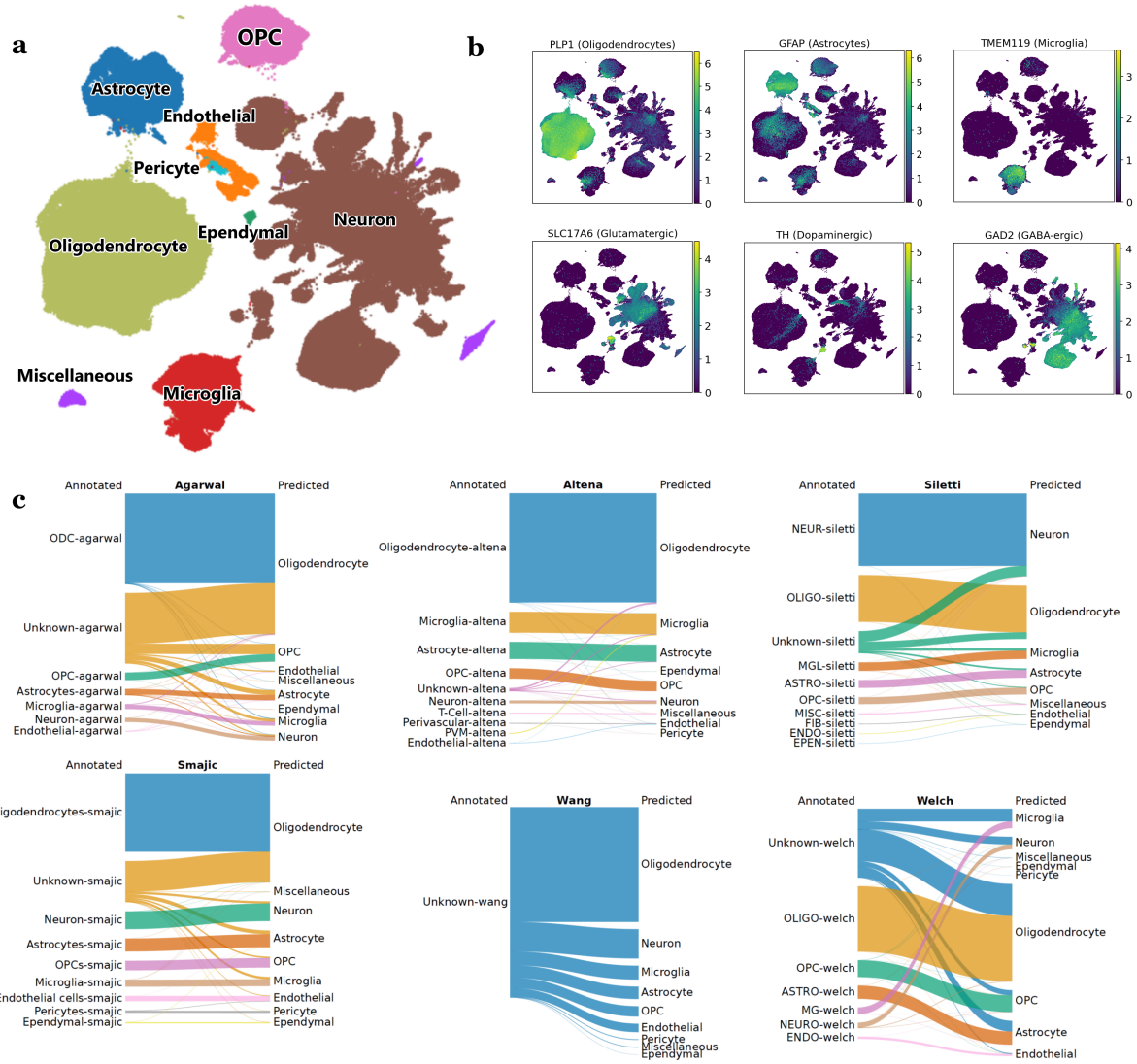

Figure 3: Final atlas
a) Integrated atlas filtered on successfully predicted supercluster labels. Prediction was performed with a k-NN model, trained on collapsed lowest resolution labels. b) The same UMAP annotated with normalized supercluster marker genes expression values. c) Sankey plots from original annotation to predicted annotations, including unknown cells.

than 5 cells were filtered out. The disparity in subtype sizes might introduce a bias, where labels with fewer cells are more inclined to become part of larger subtypes. To combat this, scHPL uses a `dynamic_neighbors` parameter that opts for the smallest value between the `n_neighbors` value and the smallest cell population. Furthermore, labels with a high reconstruction error are highlighted in figure 4. This indicates that the information from the original dataset might not be retained during the integration process.

The oligodendrocyte supercluster consists of 230,390 cells and contains 9 major label branches across all datasets. (See figure 4b) On the finest resolution, 33 leaves were found indicating unique labels. Due to its high resolution, most of these subtypes come from Siletti et al. and fall under other atlases such as Smajic et al. Similarly, the neuron supercluster consists of 157,244 cells. Its hierarchy tree contains 225 branches and a total of 1339 labels, 1291 of which came from Siletti et al. (See supplementary file 1) The hierarchical relationship of the neuron subtypes goes up to five levels deep, suggesting these subtypes exist across atlases, reinforcing their presence within the midbrain. Furthermore, the abundance of branches exclusively featuring labels from a single atlas suggests the possibility of distinct subtypes accessible only via region-specific sampling such as the van Regteren Altena atlas. The astrocyte supercluster consists of 39,179 cells, contains only three major branches, and in total 42 unique subtypes. (See figure 4c) Compared to the neurons, astrocytes and oligodendrocytes are a lot more unified.

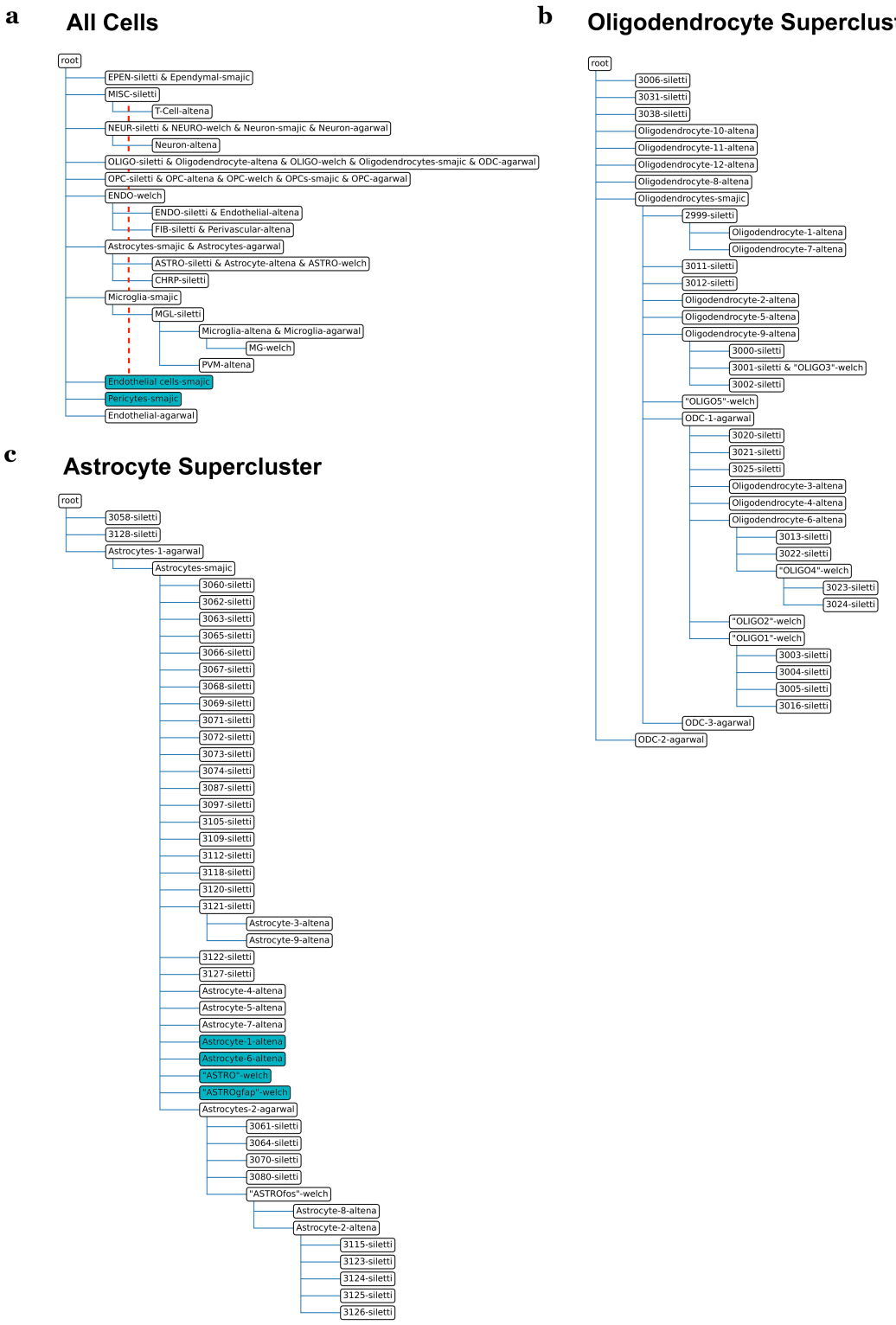

Figure 4: scHPL hierarchical tree
a) Hierarchy tree of all annotated cells, generated with scHPL and based on the original lowest resolution labels. b) Hierarchy tree of all annotated oligodendrocytes based on the original highest resolution labels. c) Hierarchy tree of all annotated astrocytes based on the original highest resolution labels.

Region-specific neuronal subtypes

As previously mentioned, branches within the hierarchy trees that contain only subtypes of particular regions imply the existence of region-specific subtypes. We examined the hierarchical relationships of the

22 neuronal subtypes within our VTA-specific dataset. An overview of the subtypes from Altena et al. and their hierarchical positions can be found in the supplementary table 5. Among these, 4 GABA and 1 unknown subtypes remained independent and did not encompass labels from other atlases. (See figure 5a-b) Additionally, 12 subtypes were identified as terminal nodes of subtypes from other atlases. One of these labels, 'Neuron-GABA-10-altena' aligns with another subtype, '650-siletti'. (See figure 5c) Siletti et al. highlighted potential contamination between dissections from adjacent tissue. Therefore, direct association between cell types and dissections should be avoided; rather, the annotation should be used as an estimate of cell type differences across neighboring regions. The cells annotated as '650-siletti' were primarily harvested from the super colliculus (SC), pretectal region (PTR), inferior colliculus (IC), and periaqueductal gray (PAG). (See figure 5d) Most of the dissections are located dorsally from the VTA but are not directly adjacent. These findings suggest that contamination from the VTA is unlikely and that the 'Neuron-GABA-10-altena'-subtype can be found in specific areas within the midbrain, but not exclusively in the VTA. Of the remaining terminal nodes, 4 subtypes named Neuron-GABA-DOPA-1 to 4, were categorized under the label '1366-siletti', which was originally annotated as splatter neuron. (See figure 5e) In Siletti et al., splatter neurons formed a heterogenous supercluster including neurons from most brain regions. The original study suggests that the splatter neuron cluster arises due to its distinction from telecephalon neurons, its inability to be organized by neurotransmitters, and its undersampling in the dataset. With the additional cells from van Regteren Altena et al., the cluster is supplemented and allows better characterization. The '1366-siletti' subtype was harvested primarily from the substantia nigra, red nucleus and nearby nuclei (SN-RN) as well as the periaqueductal gray (PAG). (See figure 5f) The SN-RN dissection is the dissection most prone to contamination with VTA neurons. Reinforcing the regional specificity of the 'Neuron-GABA-DOPA-n' subtypes. The other 11 terminal nodes are connected to low-resolution labels such as 'GABA-smajic' and 'Excitatory-smajic', implying their hierarchical position might stem from broader similarities and does not rule out their specificity to the VTA. The remaining 9 subtypes encompass multiple non-splatter labels from Siletti et al. affirming their presence within the midbrain and the possibility of further division. (See figure S8) Lastly, we focused on one of the highest resolution subtypes of the SN atlases, '"NEUROinh5"-welch'. This label encompassed two subtypes from the Siletti atlas, '1346-siletti' and '1347-siletti'. These subtypes are primarily from IC and SC dissections, suggesting that '"NEUROinh5"-welch' is likely spread around the midbrain. However, these subtypes can still be further subdivided and might contain subsets that are potentially SN-specific.

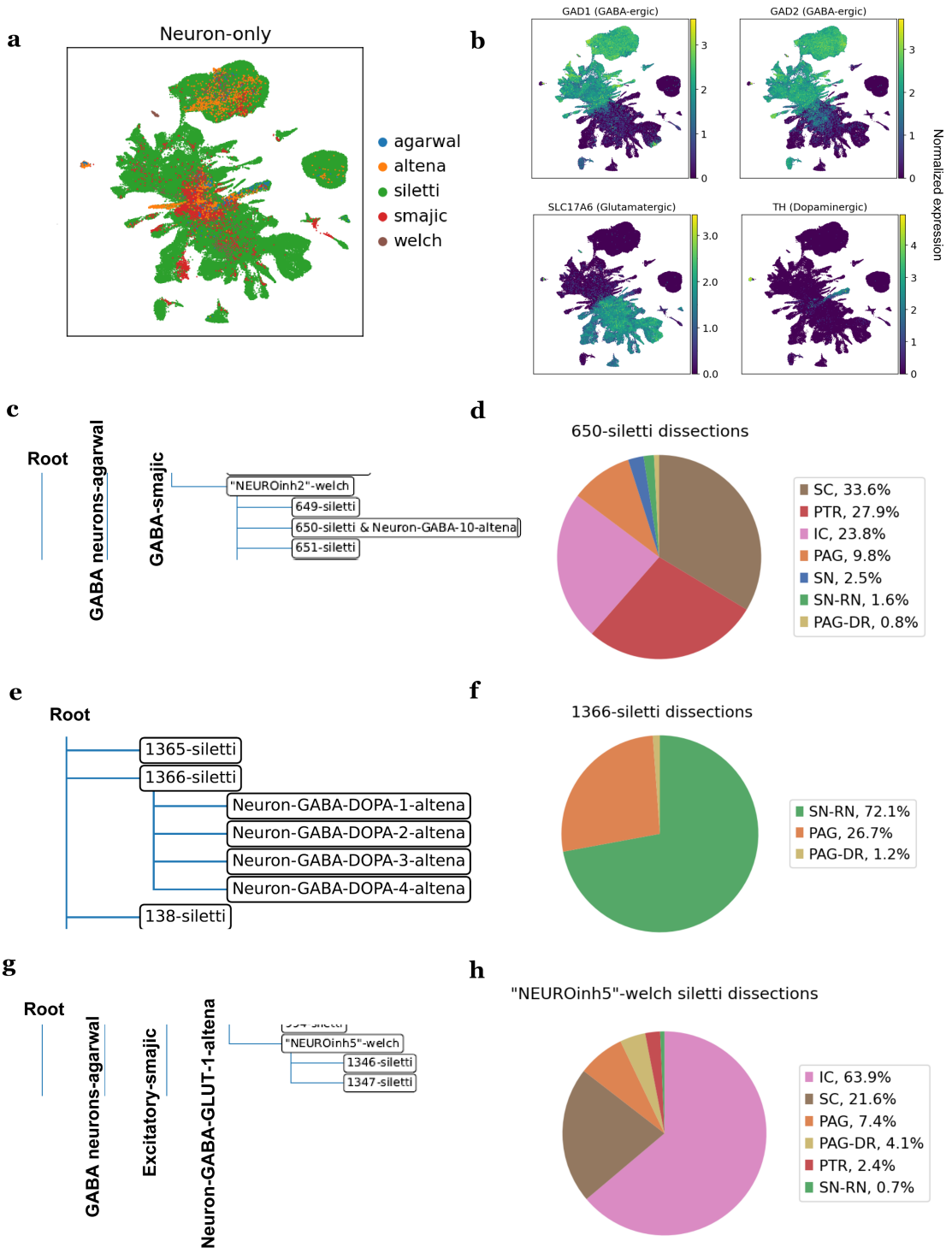

Figure 5: Region specific subtypes
a) UMAP of all neurons annotated with the original atlases. b) The same UMAP annotated with normalized neuronal marker gene expression values. c,e,g) Subset of the neuron hierarchy tree, generated with scHPL and based on the highest resolution labels. Previous branches are annotated with their label on the side. d,f,h) Original dissection proportions of the relevant Siletti subtypes corresponding to the hierarchy tree subset. c) Subset that indicates alignment between '650-siletti' and 'Neuron-GABA-10-altena'. d) Dissection proportions of '650-siletti' cells. (SC=superior colliculus; PTR=pretectal region; IC=inferior colliculus, PAG=periaqueductal gray and nearby nuclei; SN=substantia nigra; SN-RN= substantia nigra, red nucleus, and nearby nuclei; PAG-DR=periaqueductal gray and dorsal raphe nucleus) e) Subset with subtype '1366-siletti' that encompasses 'Neuron-GABA-DOPA-1 to 4-altena'. f) Dissection proportions of '1366-siletti' cells. g) Subset with subtype '"NEUROinh5"-welch' that encompasses '1346-siletti' and '1347-siletti'. h) Dissection proportions of '1346-siletti' and '1347-siletti' cells.

Discussion

In this thesis, we provide the largest human midbrain reference atlas to date, including a VTA-sampled dataset. This study underlines the importance of harmonized alignments and reveals batch effects caused by technical variations. In particular, the batch effect of the Cell Ranger's intronic mapping feature is non-trivial. Further investigation could provide insights into the deviations generated in the expression matrices. Indeed, the intronic mapping adds additional reads to genes with multiple poly-A strands within their introns, similar to gene length bias in traditional bulk RNA sequencing. [34] The feature results in expression values that are less representative of the transcriptome. Therefore, normalizing efforts will be essential in future searches for potential subtype markers. Insights into its batch effect characteristics will guide this process. Other alignment methods such as STARsolo allow the distinction between spliced, unspliced, and ambiguous reads by using `--soloFeatures Gene Velocyto`, but require additional computation time. [35, 36]

The harmonized alignments introduced additional data that increased the sensitivity of downstream analysis. This data is annotated with supercluster labels and can be included for whole atlas clustering to get a higher-resolution overview of the molecular composition within the midbrain. Newly discovered clusters can be further characterized with enriched marker genes, which may give insights into their functional properties. Furthermore, comparisons with the original annotations gives additional insights into subtype characteristics and validity.

In this thesis, an unconventional approach was employed to explore the regional-specific subtypes within the midbrain. Our investigations of the hierarchical relationships within the reference genome support the presence of VTA-specific neurons and provide insights into previously uncharacterized subtypes. These efforts are crucial for integrating knowledge across atlases. However, this method is limited by deviations in the original subtype clusters as a result of the realignment and integration process. Moreover, previous subtypes were established through clustering within the original datasets with a limited amount of cells. Cells that were connected in these datasets might not exhibit the same connectivity within the reference atlas. To enhance the analysis of the region-specific subtypes, more conventional approaches such as reclustering could be adopted. This allows subtyping based on the integrated latent space, uncovering novel subtypes, and incorporating the newly found cells. Additionally, undersampled subtypes such as the splatter neurons can be supplemented by the additional data and distinguish themselves more clearly.

We systematically evaluated three batch correction methods and performed hyperparameter tuning on scVI for our final integration. This integration process utilizes a conventional k-NN model combined with the novel tree training tool from scHPL to harmonize labels across atlases and resolutions. scVI and scHPL both support continuous, allowing new datasets to be easily integrated into the reference genome and hierarchy tree. Nevertheless, this approach can be further streamlined through additional validation and exploration of various alternatives such as scANVI. This pipeline tunes hyperparameter values on the whole dataset. Using the same values for less complex subsets might lead to over-correction. Supercluster-specific optimizations will allow for better integration of less complex subsets. Furthermore, hyperparameter tuning was performed in a time and resource-efficient manner, only considering four hyperparameters, three different values, and each hyperparameter was evaluated independently. This process can be expanded with more relevant hyperparameters and smaller intervals between values. The expansion will guide a more accurate assessment of these hyperparameters and improve the final integration.

In summary, we combined multiple snRNA datasets to form a unified reference atlas by harmonizing atlases and optimizing the integration process. Our work provides a foundation for future research on the midbrain and over time will reveal numerous insights on its cellular composition and functional components through targeted investigations of midbrain subtypes.

Data avaiability

All raw count tables, notebooks and scripts for each step within the pipeline will be available on the UMCU archive.

Methods

Validating realignment

10X Genomics has confirmed that the intronically mapped reads within their libraries are valid measures of mRNA across all their platforms and propose several mechanisms. Indeed, the pre-RNA targeting

reads represent cell-to-cell variation within the scRNA pipeline. 10X Genomics's technical note reported across various human scRNA datasets that about 15-40% of UMIs are intronic when using the Single Cell 3' Gene Expression v3.1 (Dual Index) kit. This fraction increases to about 20-60% in nuclei assays. The additional reads increase the sensitivity for cell detection and provide detection of genes without exonic UMI. It is important to note that the intronic reads are able to assign multiple UMIs to a single transcript and that there is a bias towards longer genes and genes that contain more internal poly-A sequences.

The 10X Genomics dedicated `-incude_intron` feature has only been available since Cell Ranger v5.0, so atlases using older versions won't contain any intronic reads or used an alternative approach to include the intronic reads such as using a pre-RNA based reference genome. Differences in such alignment parameters introduce various batch effects that can overshadow the biological variation in downstream analyses. To determine the necessity of a harmonized alignment for all datasets over the use of provided count tables, the batch effects of software, reference genome, and intronic mapping parameters were evaluated.

The preliminary test was conducted on 5000 healthy mouse cells from Zhong et al. and compares the alignment with Cell Ranger v3 and v6 using their respective reference genome. The resulting alignments were processed as follows using Scanpy and default parameters unless otherwise specified: count tables were loaded as Anndata objects with Scanpy, metadata such as atlas, sample, and gender was added and quality control metrics were calculated with the `calculate_qc_metrics` function. Filtering was performed on cell total UMI and genes (500-15000 cells; 500-5000 genes), the proportion of mitochondrial and ribosomal UMIs (¡3%;¡3%), doublet score with Scrublet (¿0.2) and cell count for genes (¿3). Normalization was performed on the cell level with the `normalize_total` function and followed by the `log1p` function. Finally, the datasets were concatenated for downstream analysis by performing dimensional reduction with PCA, forming the 15-nearest neighbor graph, and plotting the UMAP results. (See figure S1)

For the follow-up tests, 5000 and 5500 healthy mouse cells were taken from Hey et al. and Zhong et al. respectively. These two datasets were aligned using Cell Ranger v6.1 to its complementary mouse reference genome, mm10 (Ensemble 98), without mapping to intronic regions as control. Then we compared the following alignment parameters with the identical datasets: 1) Cell Ranger v3.0 instead of v6.1, 2) reference genome Ensemble 93 instead of Ensemble 98, or 3) including mapping to intronic regions instead of excluding it. The data was processed similarly to the preliminary test but with sample-specific filtering thresholds. Additionally, in a separate processing pipeline, the samples were filtered on highly variable genes (HVGs), using the `highly_variable_genes` function, unwanted sources of variation were regressed out with the `regress_out` function on the cells' total counts and counts were scaled using the `scale` function with a maximum of 10.

Data collection

Supplementary table 1 contains an overview of all atlases included in the integration and their metadata. For two atlases, Smajic et al. and Agarwal et al., the fastq files were downloaded using `fastq-dump` and healthy donor SRR identifiers. In order to access the R2 fastq files from Welch et al., the 10X BAM files were downloaded directly from the SRA identifiers and subsequently converted to fastq files with the dedicated 10X `bamtofastq` function. Fastq files from Wang et al. were downloaded through the dedicated Human Cell Atlas Data Explorer and filtered on healthy donors. Lastly, midbrain fastq files from Siletti et al. were manually selected and downloaded from the NEMO Archive.

Three levels of whole midbrain cell-type annotations from Siletti et al. were extracted from the neural and non-neuronal h5ad files on the dedicated Cell X Gene webpage. One level from Smajic et al. and two levels of SN annotation from Agarwal et al. were taken from GEO and supplementary data files, respectively. The high-level annotation for the data from Welch et al. was provided by the author. Cell-type annotations for the Wang et al. dataset were not available at the time of this thesis. Lastly, the VTA-specific fastq files and matching three levels of annotation of our in-house dataset were readily available.

Realignment

From the fastq files, alignments were performed with Cell Ranger v6.1 using the Ensemble 109 human reference genome, including the `--include-introns` parameter. Additionally, the `--expect-cells` parameter was set to the number of cells found in the samples of the original studies.

Quality control

Cell Ranger's filtered count tables were preprocessed with Scanpy (v1.9.6). For each atlas, we annotated the cells with metadata features including species, atlas, and sample-id. Quality control metrics were added with the `calculate_qc_metrics` function. Cells with less than 500 reads were filtered out automatically through Cell Ranger. Mitochondrial genes and genes with less than three reads were filtered out as well as cells with more than 10% mitochondrial or ribosomal reads. Doublets were detected and filtered using `scrublet`. Expected doublet rates were manually calculated with the guidelines from 10X. [REF] Normalization was performed on cell-level with `normalize_total` using a total sum of 10000. Subsequently, `log1p` was used to condense the range of values and reduce the influence of highly expressed genes. After adding the original cell-type annotations, the datasets were combined as input for the integration process.

Integration

Preliminary integrations with scVI, Harmony, and Scanorama were performed using 4000 HVGs. Furthermore, scVI's `n_latent` parameter was set to 90 for fair comparison as Harmony and Scanorama integration result in 90 latent features as well. The final integration was performed with scVI using 2000 HVGs, 90 latent features, 5 hidden layers, and 128 hidden nodes.

Benchmarks

Evaluation of the scVI, Harmony and Scanorama were performed using Scib's `benchmarker` function with all five batch correction metrics. Due to the function's constraint, a single bio-conservation had to be used, but these results can be filtered out. We continued with scVI as our integration method and performed hyperparameter tuning on four hyperparameters: 1) `n_hvgs` (2000, 3000, 4000), 2) `n_latent` (10, 50, 90), 3) `n_hidden_layers` (1, 3, 5) and 4) `n_hidden_nodes` (64, 128, 256). Optimization was performed independently, while keeping the best-performing values of the previous hyperparameter. An overview of hyperparameters and tested values can be found in supplementary table 4.

Label harmonization

Original annotation was divided into three levels, `cell_type_lvl1-3` one being the lowest resolution and three being the highest. Cells with only one level of annotation are represented by that label at each level. (See supplementary table 2) Further processing of these labels leads to four resolutions of annotations: 1) lowest, 2) low, 3) mid, and 4) high. The low-resolution labels are equivalent to `cell_type_lvl1` and high-resolution labels are equivalent to `cell_type_lvl3`. The first-level cell labels of Siletti et al. are at an incomparable resolution, considering the other atlases. These labels were collapsed into the lowest resolution labels. Finally, the mid-resolution labels are equivalent to `cell_type_lvl3`, except the labels from the Siletti dataset were kept on `cell_type_lvl2`. Whole dataset hierarchy tree formation was performed using scHPL's `train_tree` function with the lowest resolution labels. Subsequent trees of neuron, oligodendrocyte, and astrocyte superclusters were formed using high-resolution labels. Harmonization of labels to the first level within the tree was done using a custom `harmonize_label` function. (See supplementary file 2)

References

[1] Kathleen Carmichael et al. "Diverse midbrain dopaminergic neuron subtypes and implications for complex clinical symptoms of Parkinson's disease". In: *Ageing and Neurodegenerative Diseases* (2021). DOI: 10.20517/and.2021.07.

[2] R. Alison Adcock et al. "Reward-Motivated Learning: Mesolimbic Activation Precedes Memory Formation". In: *Neuron* 50 (2006), pp. 507–517. ISSN: 08966273. DOI: 10.1016/j.neuron.2006.03.036.

[3] Eric J. Nestler and William A. Carlezon. "The Mesolimbic Dopamine Reward Circuit in Depression". In: *Biological Psychiatry* 59 (2006), pp. 1151–1159. ISSN: 00063223. DOI: 10.1016/j.biopsych.2005.09.018.

[4] Kelly R. Tan et al. "GABA Neurons of the VTA Drive Conditioned Place Aversion". In: *Neuron* 73 (2012), pp. 1173–1183. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.02.015.

[5] Natalia Omelchenko and Susan R. Sesack. "Ultrastructural analysis of local collaterals of rat ventral tegmental area neurons: GABA phenotype and synapses onto dopamine and GABA cells". In: *Synapse* 63 (2009), pp. 895–906. ISSN: 08874476. DOI: 10.1002/syn.20668.

[6] Ruud Van Zessen et al. "Activation of VTA GABA Neurons Disrupts Reward Consumption". In: *Neuron* 73 (2012), pp. 1184–1194. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.02.016.

[7] Hui Ling Wang et al. "Rewarding effects of optical stimulation of ventral tegmental area glutamatergic neurons". In: *Journal of Neuroscience* 35 (2015), pp. 15948–15954. ISSN: 15292401. DOI: 10.1523/JNEUROSCI.3428-15.2015.

[8] David H. Root et al. "Selective Brain Distribution and Distinctive Synaptic Architecture of Dual Glutamatergic-GABAergic Neurons". In: *Cell* 23 (2018), pp. 3465–3479. DOI: 10.1016/j.celrep.2018.05.063.

[9] David H. Root et al. "Distinct Signaling by Ventral Tegmental Area Glutamate, GABA, and Combinatorial Glutamate-GABA Neurons in Motivated Behavior". In: *Cell Reports* 32.9 (2020), p. 108094. DOI: https://doi.org/10.1016/j.celrep.2020.108094.

[10] Stephan Lammel, Byung Kook Lim, and Robert C. Malenka. "Reward and aversion in a heterogeneous midbrain dopamine system". In: *Neuropharmacology* 76 (2014), pp. 351–359. ISSN: 00283908. DOI: 10.1016/j.neuropharm.2013.03.019.

[11] Marisela Morales and Elyssa B. Margolis. "Ventral tegmental area: Cellular heterogeneity, connectivity and behaviour". In: *Nature Reviews Neuroscience* 18 (2017), pp. 73–85. ISSN: 14710048. DOI: 10.1038/nrn.2016.165.

[12] Gioele La Manno et al. "Molecular architecture of the developing mouse brain". In: *Nature* 596 (2021), pp. 92–96. ISSN: 14764687. DOI: 10.1038/s41586-021-03775-x.

[13] Amit Zeisel et al. "Molecular Architecture of the Mouse Nervous System". In: *Cell* 174 (2018), 999–1014.e22. ISSN: 10974172. DOI: 10.1016/j.cell.2018.06.021.

[14] Arpiar Saunders et al. "Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain". In: *Cell* 174 (2018), 1015–1030.e16. ISSN: 10974172. DOI: 10.1016/j.cell.2018.07.028.

[15] Jixing Zhong et al. "Single-cell brain atlas of Parkinson's disease mouse model". In: *Journal of Genetics and Genomics* 48.4 (2021), pp. 277–288. ISSN: 1673-8527. DOI: https://doi.org/10.1016/j.jgg.2021.01.003.

[16] Paul W. Hook et al. "Single-Cell RNA-Seq of Mouse Dopaminergic Neurons Informs Candidate Gene Selection for Sporadic Parkinson Disease". In: *American journal of human genetics* 102.3 (2018), pp. 427–446.

[17] Zizhen Yao et al. "A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain". In: *bioRxiv* (2023). DOI: 10.1101/2023.03.06.531121.

[18] Jonah Langlieb et al. "The cell type composition of the adult mouse brain revealed by single cell and spatial genomics". In: *bioRxiv* (2023). DOI: 10.1101/2023.03.06.531307.

[19] Joshua D. Welch et al. "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity". In: *Cell* 177.7 (2019), 1873–1887.e17. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2019.05.006.

[20] Semra Smajić et al. "Single-cell sequencing of human midbrain reveals glial activation and a Parkinson-specific neuronal state". In: *Brain* 145.3 (2021), pp. 964–978. ISSN: 0006-8950. DOI: 10.1093/brain/awab446.

[21] Tushar Kamath et al. "Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease". In: *Nature Neuroscience* 25 (2022), pp. 588–595. ISSN: 15461726. DOI: 10.1038/s41593-022-01061-1.

[22] Qian Wang et al. "Single-cell transcriptomic atlas of the human substantia nigra in Parkinson's disease". In: *BioRxiv* **Preprint**.[cited 2023 Sep 20]] (2022).

[23] Devika Agarwal et al. "A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders". In: *Nature Communications* 11.4183 (2020).

[24] Kimberly Siletti et al. "Transcriptomic diversity of cell types across the adult human brain". In: *Science* 382.6667 (2023), eadd7046. DOI: 10.1126/science.add7046.

[25] Lukas Steuernagel et al. "HypoMap—a unified single-cell gene expression atlas of the murine hypothalamus". In: *Nature Metabolism* 4.10 (2022), pp. 1402–1419. DOI: https://doi.org/10.1038/s42255-022-00657-y.

[26] Stephanie C Hicks et al. “Missing data and technical variability in single-cell RNA-sequencing experiments”. In: *Biostatistics* 19.4 (2017), pp. 562–578.

[27] Hoa Thi Nhu Tran et al. “A benchmark of batch-effect correction methods for single-cell RNA sequencing data”. In: *Genome Biology* 21.1 (2020). DOI: https://doi.org/10.1186/s13059-019-1850-9.

[28] Malte D. Luecken et al. “Benchmarking atlas-level data integration in single-cell genomics”. In: *Nature Methods* 19.1 (2021), pp. 41–50. DOI: https://doi.org/10.1038/s41592-021-01336-8.

[29] Ilya Korsunsky et al. “Fast, sensitive and accurate integration of single-cell data with Harmony”. In: *Nature Methods* 16.12 (2019), pp. 1289–1296. DOI: https://doi.org/10.1038/s41592-019-0619-0.

[30] Brian Hie, Bryan Bryson, and Bonnie Berger. “Efficient integration of heterogeneous single-cell transcriptomes using Scanorama”. In: *Nature Biotechnology* 37.6 (2019), pp. 685–691. DOI: https://doi.org/10.1038/s41587-019-0113-3.

[31] Romain Lopez et al. “Deep generative modeling for single-cell transcriptomics”. In: *Nature Methods* 15.12 (2018), pp. 1053–1058. DOI: https://doi.org/10.1038/s41592-018-0229-2.

[32] Chenling Xu et al. “Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models”. In: *Molecular Systems Biology* 17.1 (2021). DOI: https://doi.org/10.15252/msb.20209620.

[33] Maren Büttner et al. “A test metric for assessing single-cell RNA-seq batch correction”. In: *Nature Methods* 16.1 (2018), pp. 43–49. DOI: https://doi.org/10.1038/s41592-018-0254-1.

[34] “Interpreting Single Cell Gene Expression Data With and Without Intronic Reads”. In: *10X Genomics* CG000554 (2022). URL: https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/interpreting-single-cell-gene-expression-data-with-and-without-intronic-reads.

[35] Benjamin Kaminow, Dinar Yunusov, and Alexander Dobin. In: *STARsolo: Accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-Seq Data* (2021). DOI: 10.1101/2021.05.05.442755.

[36] Gioele La Manno et al. “RNA velocity of single cells”. In: *Nature* 560.7719 (2018), pp. 494–498. DOI: https://doi.org/10.1038/s41586-018-0414-6. URL: https://www.nature.com/articles/s41586-018-0414-6.

Supplementary Figures

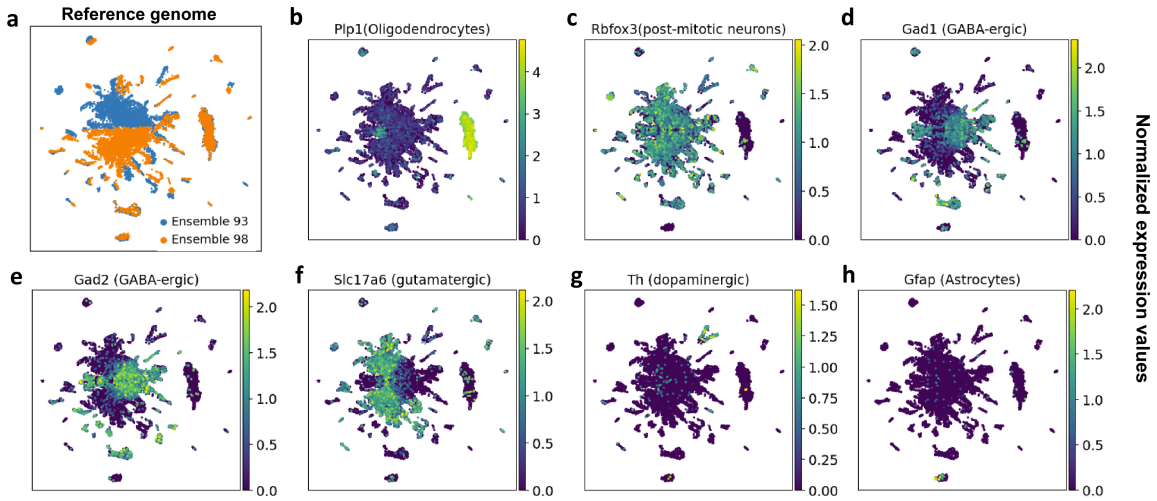

Figure S1: Preliminary alignment parameter comparison
a) UMAP of the preliminary alignment comparison between Cell Ranger v3.0 with reference genome Ensemble 93 and Cell Ranger v6.1 with reference genome Ensemble 98. Sample data from Zhong et al. was used. b-h) The same UMAP annotated with normalized marker gene expression values

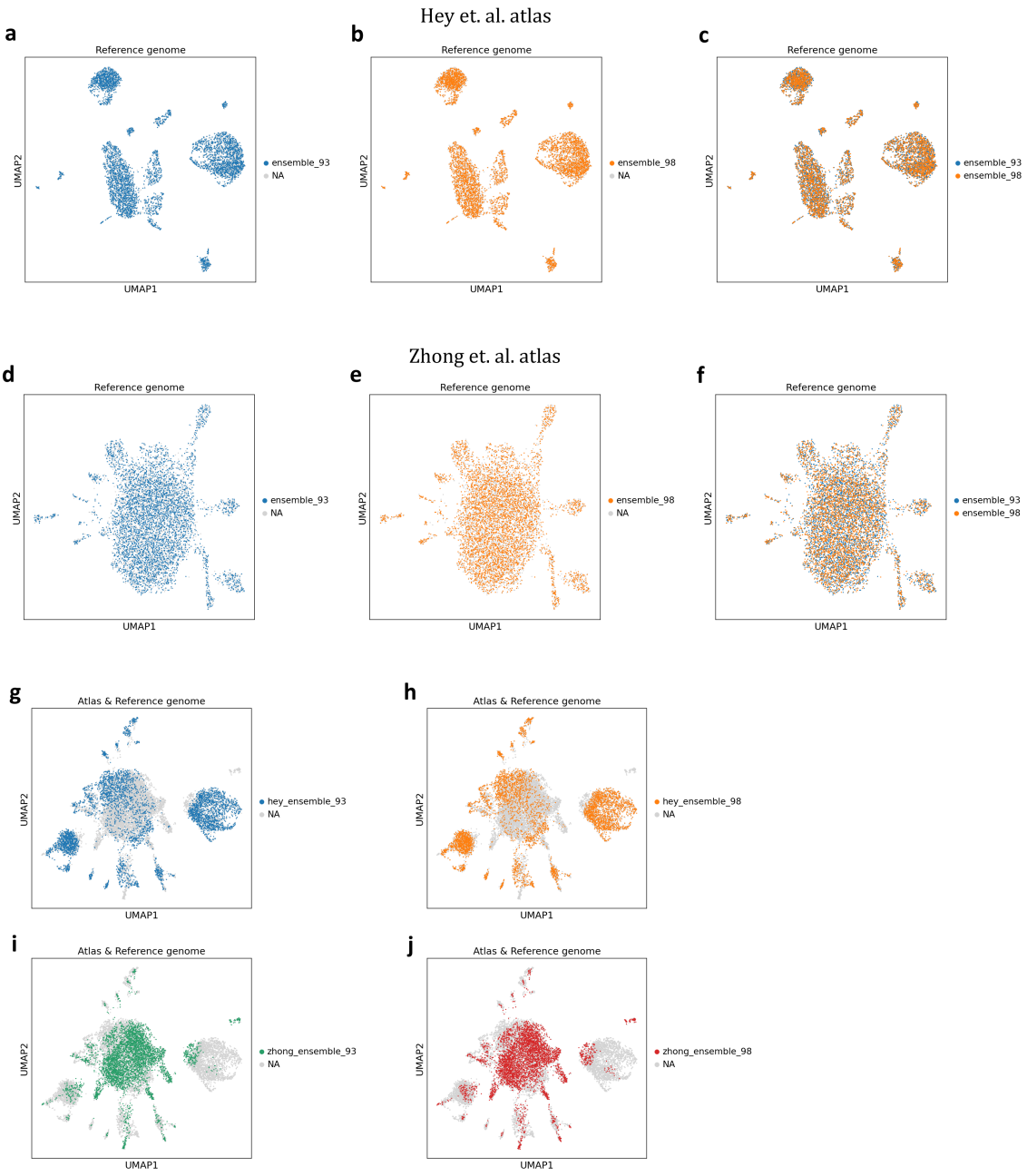

Figure S2: Reference genome parameter comparison with HVGs
a-c) Reference genome parameter comparison between Ensemble 93 and Ensemble 98 of only data from Hey et al. d-f) Reference genome parameter comparison between Ensemble 93 and Ensemble 98 of only data from Zhong et al. g-j) Subplots of the reference genome parameter comparison between Ensemble 93 and Ensemble 98, including data from both Hey et al. and Zhong et al.

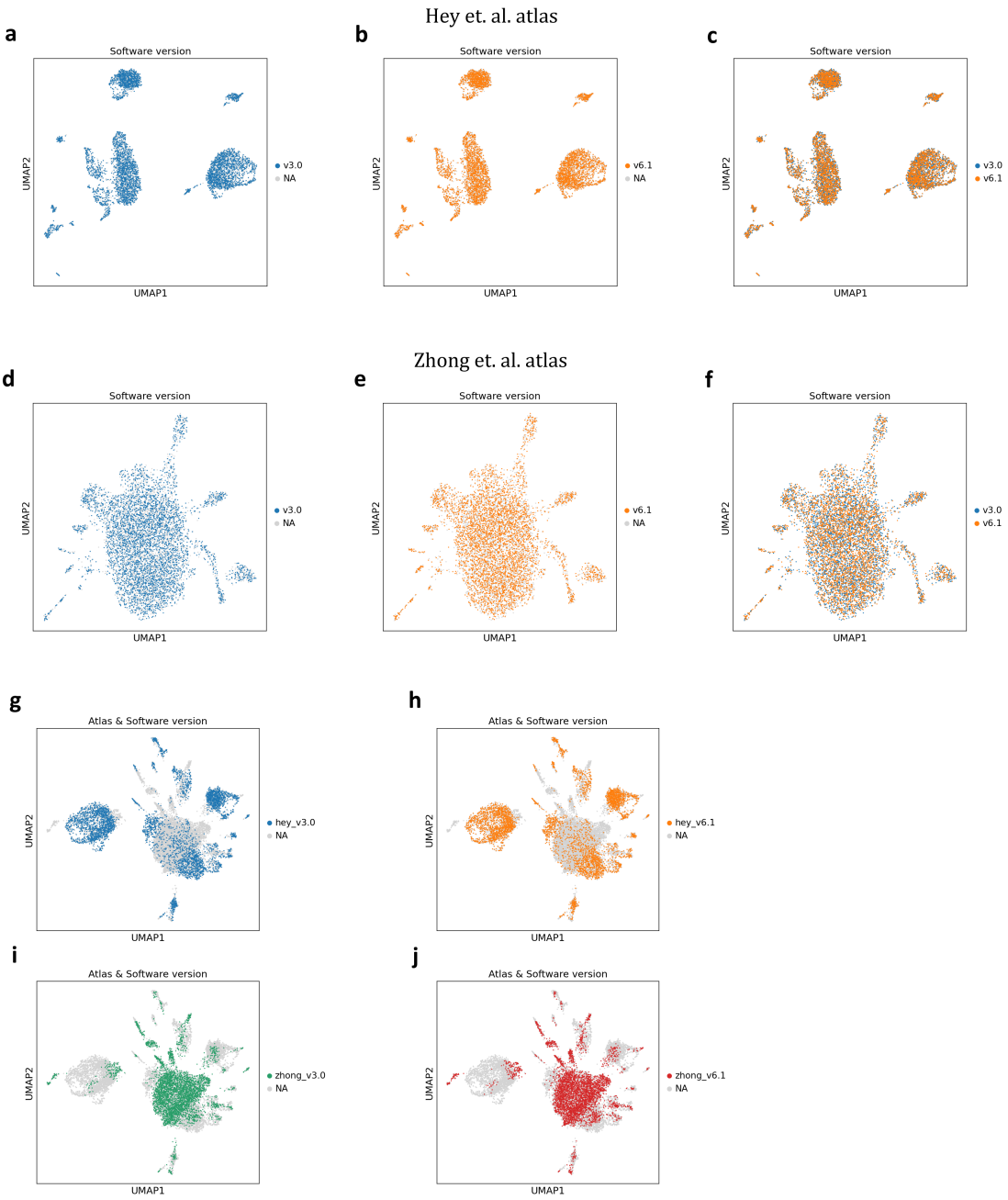

Figure S3: Software version parameter comparison with HVGs
a-c) Software version parameter comparison between Cell Ranger v3.0 and Cell Ranger v6.1 of only data from Hey et al. d-f) Software version parameter comparison between Cell Ranger v3.0 and Cell Ranger v6.1 of only data from Zhong et al. g-j) Subplots of the software version parameter comparison between Cell Ranger v3.0 and Cell Ranger v6.1, including data from both Hey et al. and Zhong et al.

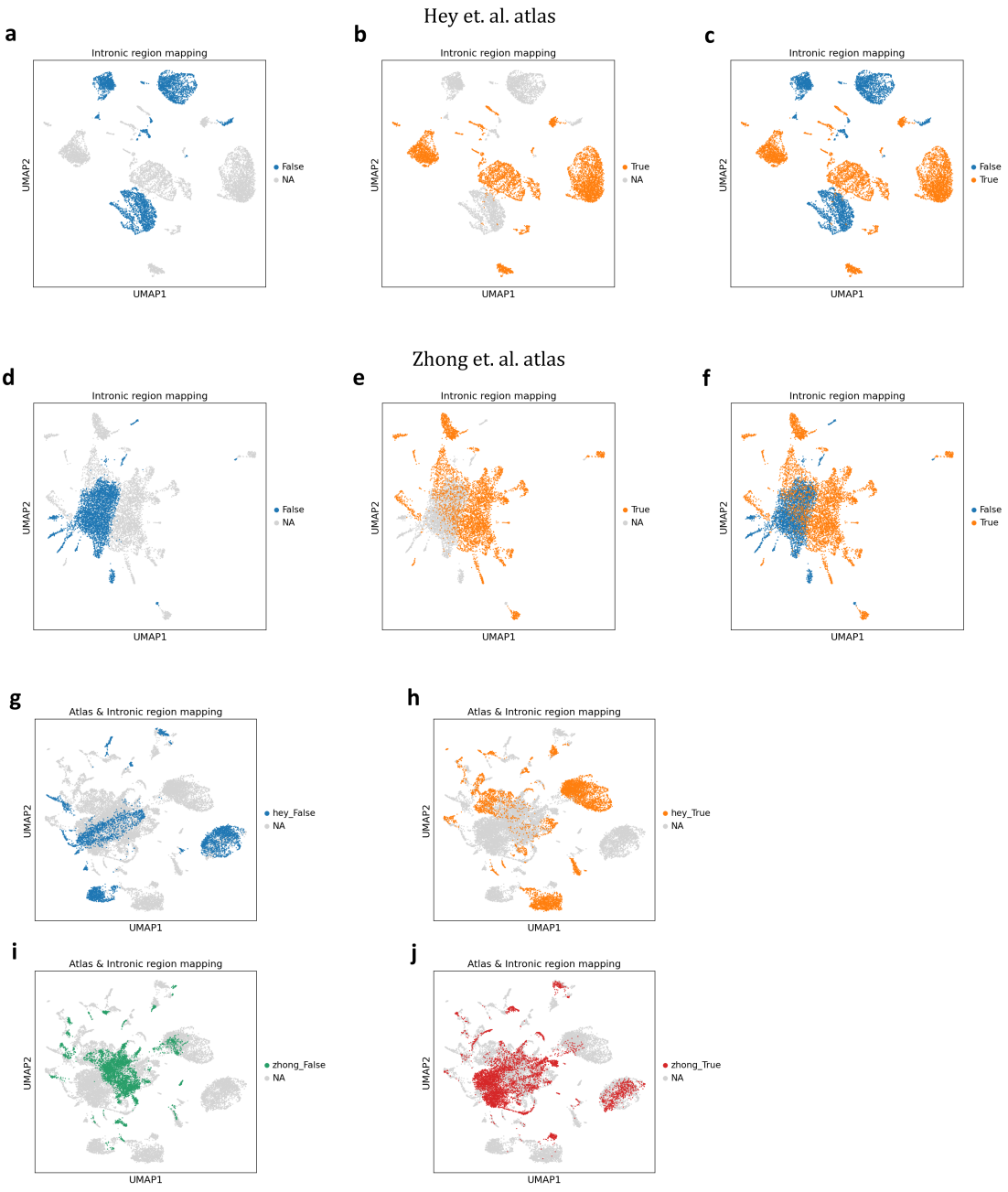

Figure S4: Intronic region mapping parameter comparison with HVGs
a-c) Intronic region mapping parameter comparison between True and False of only data from Hey et al. d-f) Intronic region mapping parameter comparison between True and False of only data from Zhong et al. g-j) Subplots of the intronic region mapping parameter comparison between True and False, including data from both Hey et al. and Zhong et al.

Figure S5: Integration and hyperparameter evaluation
a) Evaluation of integration methods scVI, Harmony and Scanorama using five batch correction metrics. (Green indicates the best score and purple indicates the worst score) b) Evaluation of scVI integration with different hyperparameter values using five batch correction and five bio-conservation metrics. From top to bottom the hyperparameter values tested were: amount of highly variable genes (2000, 3000 and 4000), amount of latent space features (10, 50 and 90), amount of hidden nodes (64, 128 and 256), and amount of hidden layers (1, 3 and 5). The best performing values are on top of their respective grouping.

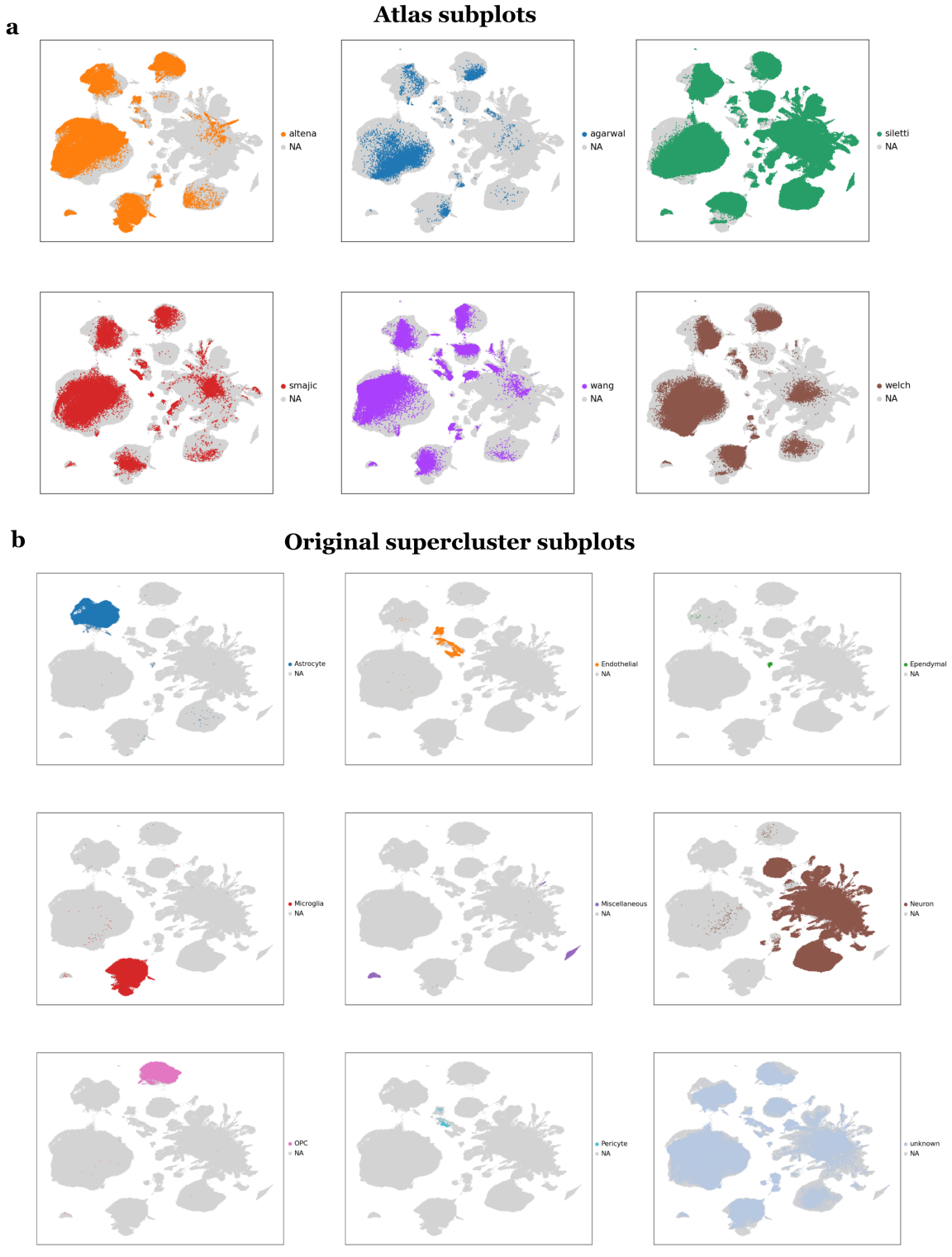

Figure S6: Final atlas subplots
a) UMAP subplots of the final integrated atlas filtered on predicted superclusters annotated with atlases. b) The same UMAP subplots, annotated with original superclusters.

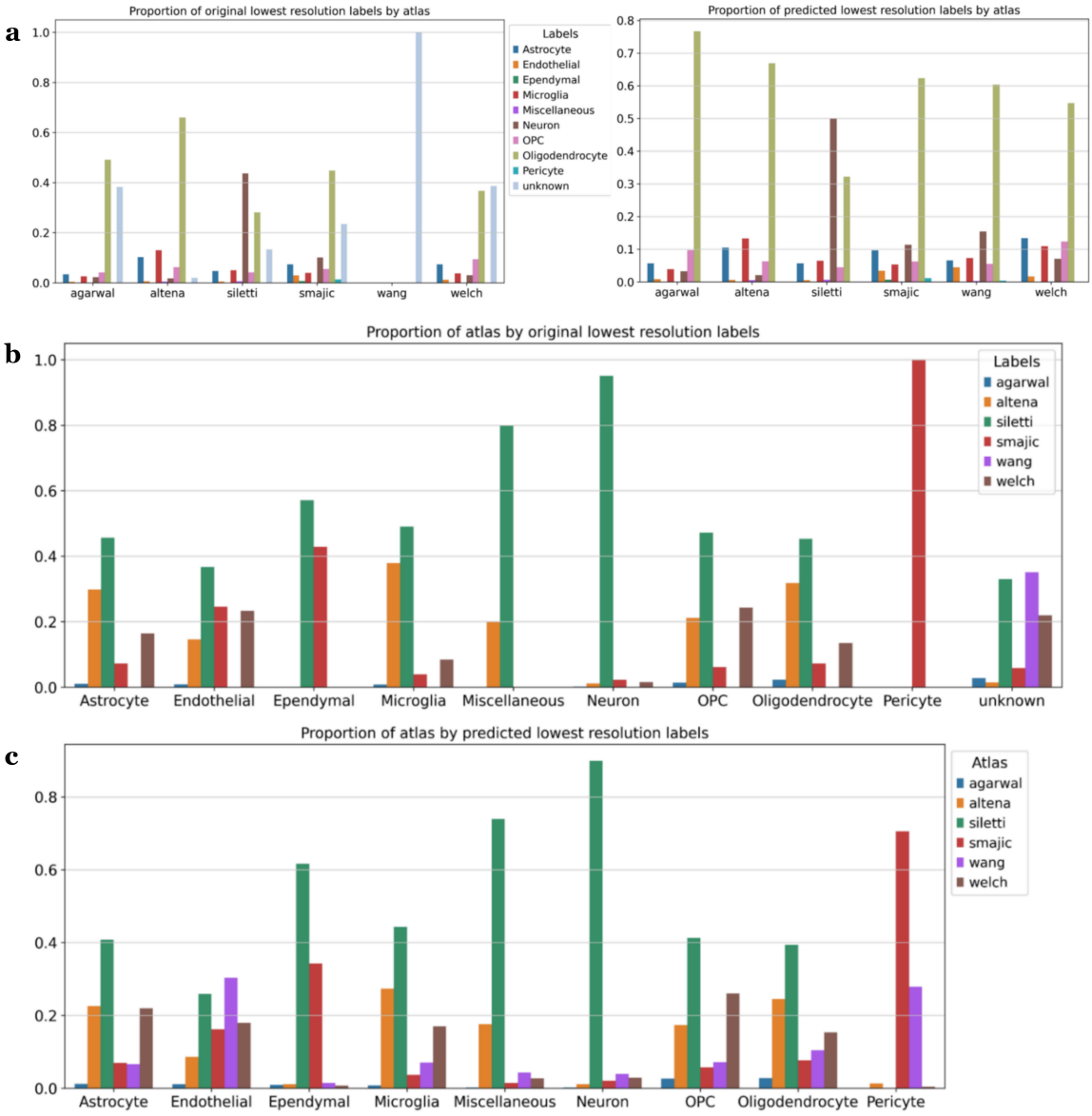

Figure S7: Proportions of major celltypes and atlases
a) Proportions of supercluster labels in each atlas before (left) and after (right) supercluster prediction. b) Proportions of atlases in each supercluster label before prediction. c) Proportions of atlases in each supercluster label after prediction.

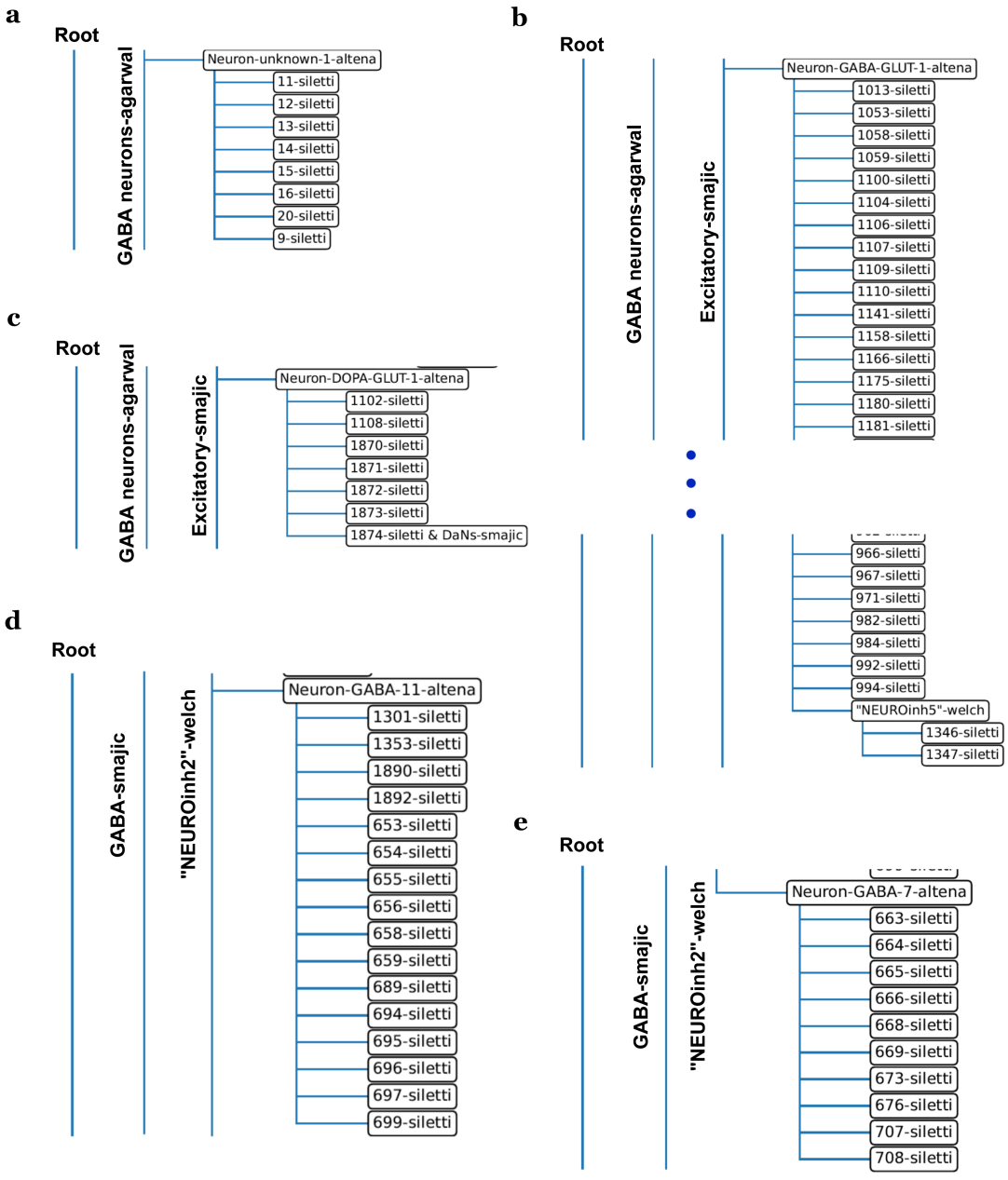

Figure S8: Non terminal subtypes of van Regteren Altena et al.
a) Hierarchy tree subset of 'Neuron-unknown-1-altena'. b) Hierarchy tree subset of 'Neuron-GABA-GLUT-1-altena'. c) Hierarchy tree subset of 'Neuron- DOPA-GLUT-1-altena'. d) Hierarchy tree subset of 'Neuron-GABA-11-altena'. e) Hierarchy tree subset of 'Neuron-GABA-7-altena'.