ONE TREE, MANY CELLS: SINGLE-CELL DATA INTEGRATION ACROSS SPECIES

Anonymous authors
Paper under double-blind review

Throughout animal life, the brain is stunningly diverse, taking on myriad structural forms shaped over evolutionary time. Homology between distantly related taxa of anatomically specialized cells clearly exists (Lemaire et al., 2021), yet the extent to which cellular identities are conserved over evolutionary time is not well understood, as comparisons of functionally specialized brain regions are technically challenging or infeasible across neuroanatomically distinct species.

Comparative genomics, now possible at single-cell resolution, overcomes these constraints by profiling the fundamental units of the nervous system in a species- and anatomy-agnostic manner, thus allowing for direct comparisons of brain cell-type identities across diverse taxa. In particular, single-cell RNA-Sequencing (scRNA-Seq) allows for unprecedentedly deep profiling of cellular identities across species (Ofengeim et al., 2017), thus posing a unique question at the intersection of genomics, neurobiology, and computation: **How can brain cell populations, characterized at single-cell resolution, be meaningfully compared across distantly related species?**

One major challenge to such comparisons is finding a latent representation of brain cells from different species, such that similar cells cluster together, and variation across species can be attributed, at least in part, to signatures of evolutionary change in gene activity (Song et al., 2023). Previous approaches have relied on stringent homology detection methods like "one-to-one" orthology (Kon et al., 2022), though it is largely unknown how this approach performs as evolutionary distance between species increases. Orthogroups, or sets of genes from two or more species predicted to have descended from a common ancestral gene, better account for complex gene histories (Emms et al., 2019) and have also been used for integration purposes (Lemaire et al., 2021). However, like for one-to-one orthology, comparing gene expression alone across species misses other defining features of cell-type identities, such as regulatory relationships between transcription factors (TFs) and their targets (Lemaire et al., 2021). Recently, an integration model leveraging protein language models was released to explore similarities at the protein level to integrate across different species (Yanai et al., 2024). This approach aligns conceptually with orthogroups, as it clusters gene expression based on protein similarity, with orthogroups being defined from proteome data.

While expression is commonly used for joint embedding, network-level approaches such as gene regulatory network (GRN) inference have not yet been applied as a substrate for integration. Leveraging cell-type-specific regulatory programs could provide a meaningful cross-species comparison by taking advantage of similarities in TF sequence and binding site homologies, which tend to be highly conserved, even at greater evolutionary distances (Tanay et al., 2021).

To achieve this, we envision an innovative workflow for scRNA-Seq integration that captures the evolutionary plasticity of brain gene expression across diverse taxa. This workflow would leverage GRNs as a foundation for cross-species integration of single-cell brain data, aiming to unravel the evolution and diversification of brain cells across the animal kingdom. In this brief report, we set the foundations for such a mechanistic model by presenting the caveats of current integration models when applied to cross-species integration.

Datasets: We used public scRNA-Seq data from the GEO database for Human (GSE67835), Macaque (GSE233278), Mouse (GSE60361), Fly (GSE107451), and Honey Bee (GSE142044).

Orthofinder (Emms et al., 2019) was used to generate orthogroups based on the proteomes of human (*Homo sapien*), crab-eating macaque (*Macaca mulatta*), chimpanzee (*Pan troglodytes*, mouse (*Mus musculus*), rat (*Rattus norvegicus*), tropical clawed frog (*Xenopus laevis*), fruit fly (*Drosophila melanogastor*), Honey Bee (*Apis mellifera*), yellow fever mosquito (*Aedes aegypti*) and swiftwater hydra (*Hydra vulgaris*). The additional species listed here but not in "Datasets" above were included to better represent phylogenetic relationships and enhance the performance of OrthoFinder, in-line with best practices. To transform the scRNA-seq matrices from cells $\times$ genes to cells $\times$ orthogroups, we assigned each gene to its corresponding orthogroup and added the expression levels of all genes within the same orthogroup.

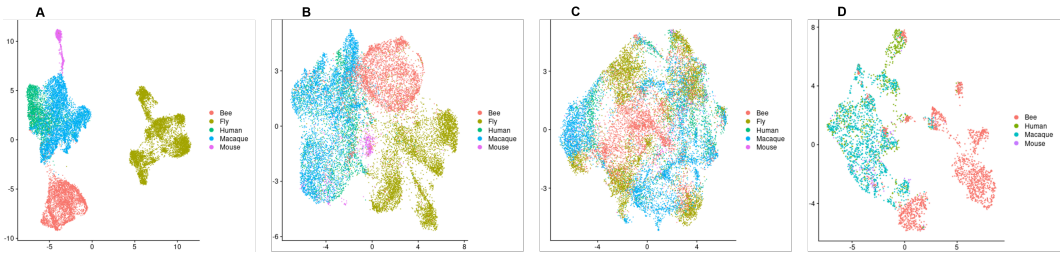

Figure 1: UMAP of Human, Macaque, Mouse, Fly and bee single cell data using A) No integration method B) Harmony and C) Seurat. D) UMAP of Human, Macaque, Mouse and Bee astrocyte single cell data integrated using Harmony.

Integration models for single-cell data: We used Seurat (Stuart et al., 2019) to preprocess and normalize the datasets, followed by integration using both the Harmony (Korsunsky et al., 2019) and Seurat. We then subsetted only the astrocyte populations and performed integration using Harmony.

CURRENT INTEGRATION MODELS DOES NOT CAPTURE BIOLOGICAL RELATIONSHIPS.

Since we could not identify single-copy orthologs across all datasets (human, mouse, macaque, fruit fly and honey bee; see Methods), we computed orthogroups (i.e., sets of one or more genes for each species that are predicted to have descended from a common ancestral gene) as a lower-dimensional representation (Fig. 1A), increasing the overlap to 199 common features. However, we still observed very low overlap across species indicating that the computed orthogroups do not fully capture the transcriptional identity of cells (Fig. 1A). Regardless of the technical variations therein, human, mouse, and macaque cells tend to cluster together, while fruit fly and honey bee, as the only invertebrate species in this analysis, remain transcriptionally distinct. Additionally, the evaluated state-of-the-art integration methods (Harmony and Seurat) yielded completely different levels of overlap in the low-dimensional embeddings (Fig. 1B, Fig. 1C).

We focused our preliminary analyses on astrocytes, a subtype of glia that likely share a common evolutionary origin across vertebrates and invertebrates (Falcone, 2022).Therefore, we leveraged prior knowledge of conserved transcriptional programs that characterize astrocytes and astrocyte-like glia across taxa. We specifically explored honey bee astrocytes, as fruit fly glia annotations were insufficient for the targeted analyses. Integration results (Fig. 1D) demonstrate a clear overlap between the human, macaque, and mouse datasets, while the honey bee data generate separate clusters. We will further refine our approach to quantifying this separation and determining if it is consistent with the known evolutionary relationships across the included species.

Due to the substantial variability in integration results across different methods and the limited understanding of cross-species cell-type relationships, a comprehensive analysis of existing integration techniques is essential. We hypothesize that an optimal low-dimensional embedding exists that more accurately captures the underlying biology of evolutionary relationships across species. We hope that this approach will enable us to better study brain evolution across taxa, shedding light on the conserved and divergent aspects of brain function and structure.

REFERENCES

D. M. Emms et al. *Genome Biol.*, 20:238, 2019.
C. Falcone. *Frontiers in cell and developmental biology*, 10:931311., 2022.
T. Kon et al. *Commun. Biol.*, 5:1404, 2022.
I. Korsunsky et al. *Nature Methods*, 16:1289–1296, 2019.
L.A. Lemaire et al. *Sci. Adv.*, 7, 2021.
D. Ofengeim et al. *Trends Mol. Med.*, 23:563–576, 2017.
Y. Song et al. *Nat. Commun.*, 14, 2023.
T. Stuart et al. *Cell*, 177:1888–1902, 2019.
A. Tanay et al. *Trends Genet.*, 2021.
R. Yanai et al. *Nature methods*, 21:1492–1500, 2024.