Cytosplore EvoViewer: Visual Analytics of Conserved Evolutionary Patterns in multi-species single-cell sequencing data

Soumyadeep Basu [*,†] Morgan Wirthlin [‡] Jeroen Eggermont [*] Thomas Kroes [*]
Boudewijn Lelieveldt [*] Ed Lein [‡] Trygve Bakken [‡] Thomas Höllt [†]

Leiden University Medical Center [*], Delft University of Technology [†], Allen Institute for Brain Science [‡]

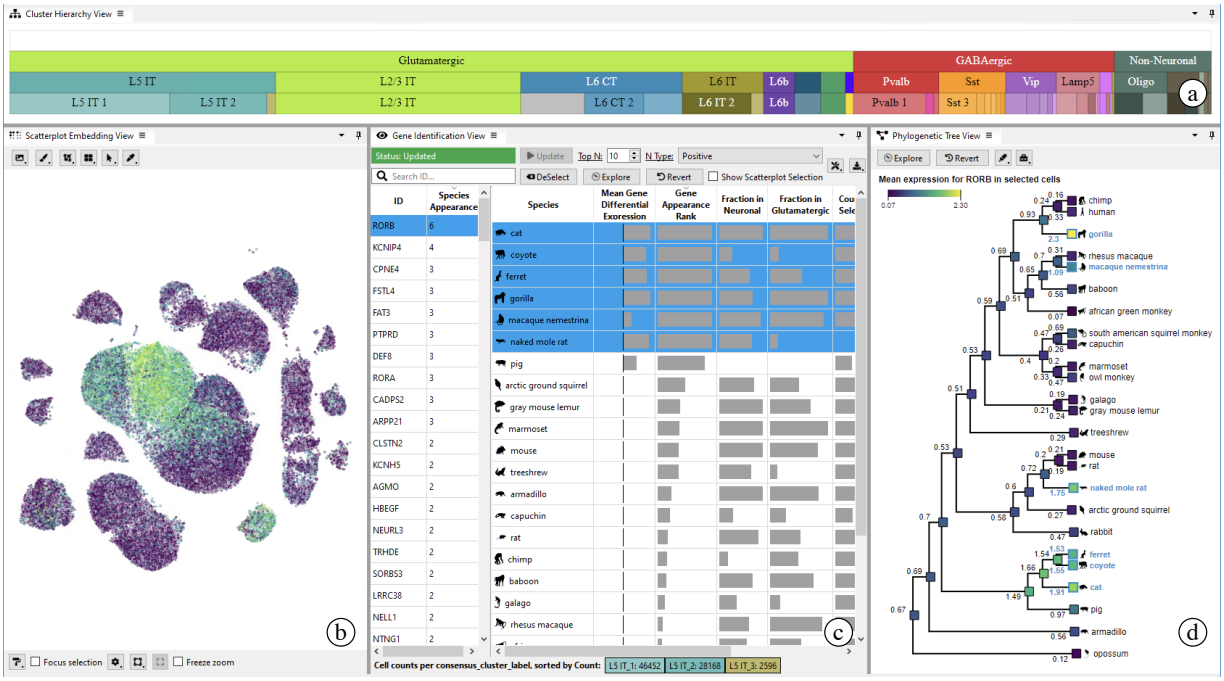

Figure 1: **Cytosplore EvoViewer** showing data on cell composition within the primary motor cortex of twenty-five mammalian species with four views: cell type hierarchy (a), single-cell embedding (b), detail table (c), and phylogenetic tree (d).

Abstract

Single-cell transcriptomics has enhanced our understanding of the brain's cellular composition. Biologists now analyze complex datasets to explore how marker genes influence biological processes, genetic variations, and phenotypic traits. A challenge is comparing these datasets across species to detect subtle differences or similarities to evolutionary development. Here, we present Cytosplore EvoViewer to facilitate examining relationships between transcriptomic datasets across species, simplifying the analysis of marker gene regulation and its impact on biological functions and integrating these findings with prior evolutionary knowledge or species-specific traits. We conducted a design study, including domain analysis, implementation of the results into Cytosplore EvoViewer, and an expert evaluation. Cytosplore EvoViewer offers valuable insights into genetic variations and evolutionary dynamics, helping to understand the diversity and the unity within diversity across species and their evolutionary development.

The Cytosplore EvoViewer installer application can be downloaded from the Cytosplore Viewer website[1], and its source code is available on the ManiVault Studio GitHub[2].

Index Terms: Bioinformatics Visualization, Biomedical and Medical Visualization, High-Dimensional Data, Design Study

1 Introduction

Modern advancements in single-cell transcriptomics have significantly improved our understanding of the brain's complexity. This technology allows scientists to explore the relationships between gene function, biological processes, cellular diversity, and phenotypic traits in unprecedented detail. Understanding gene regulation across species is crucial for grasping biological functions and species-specific traits [35]. Recently, datasets from multiple species have been acquired to identify gene function similarities and differences, mapping these to evolutionary development and unique traits. Conserved traits across species suggest their evolutionary importance, while differences indicate adaptation. For example, recent studies reveal how cell types developed within species with large evolutionary distances, highlighting unique features of cell types and gene expression in their brains [25, 1].

The vast size and complexity of multi-species, single-cell transcriptomics datasets present significant challenges for effective

*LKEB, LUMC:{s.basu, j.eggermont, t.kroes, b.p.f.lelieveldt}@lumc.nl
†CGV, TU Delft: {s.basu-2, t.hollt-1}@tudelft.nl
‡Allen Inst.: {morgan.wirthlin, edl, trygveb}@alleninstitute.org

[1] https://viewer.cytosplore.org
[2] https://github.com/ManiVaultStudio/CytosploreEvoViewer

analysis and interpretation [11]. While these datasets offer valuable insights, biologists often require custom solutions to address specific problems [30]. Existing visualization tools for transcriptomic datasets typically focus on analyzing individual datasets in detail or comparing conditions, with few designed to address specific evolutionary biology questions or incorporate prior evolutionary knowledge.

In comparative transcriptomics, which centers on gene expression and species-specific cell type hierarchies, current tools struggle to enable comparisons among larger numbers of species. Interactive visualization tools have become essential for analyzing these complex biological datasets. Integrating evolutionary biology with single-cell transcriptomics provides a powerful framework for exploring how species traits evolved. Insights gained can lead to practical applications, such as understanding human-specific diseases or guiding conservation efforts for endangered species.

Our collaborators currently rely on various tools, including off-the-shelf general data analytics software (e.g., Igor Pro), but also have other in-house script-based analytics pipelines for molecular profiling (e.g., scrattch[3]). However, these tools address only specific aspects of the data analysis. An integrated solution for interactive exploration and comparison of the presented multi-species data in the context of evolutionary aspects remains lacking.

This paper presents a visual analytics design study to facilitate the comparative analysis of transcriptomic datasets from various species while integrating phylogenetic knowledge. Our approach enables biologists to conduct exploratory comparisons across species. The system was developed through iterative participatory design sessions with biologists, resulting in the Evolutionary Exploration Viewer (Cytosplore EvoViewer). Cytosplore EvoViewer allows users to identify similarities and differences in phylogenetic hierarchies, offering insights into evolutionary relationships at the single-cell level and highlighting gene expression variations across species.

The key contributions of this paper are:

- Domain Abstraction: We characterize the goals, tasks, and data, used in cross-species comparison, leading to an abstracted domain characterization. This characterization integrates evolutionary knowledge for cross-species comparison, introduces an iterative marker gene discovery and validation process, and links the gathered information to evolutionary relationships.
- Cytosplore EvoViewer: We iteratively designed and implemented a visual analysis framework for comparing large numbers of species following the domain abstraction.
- Pilot Evaluation: We conducted a pilot user study with domain experts validating Cytosplore EvoViewer's effectiveness in facilitating the identified tasks and goals.

2 Biological Background

Species evolve differently due to environmental pressures and the principle of survival of the fittest [8]. This process, driven by environmental factors, genetic mutations, and reproductive success, leads to diverse life forms with gradual changes in *traits* [20]. While each species has unique traits – such as humans developing complex language—many also share features, like opposable thumbs. Understanding this blend of uniqueness and similarity is interesting in evolutionary genetics [38].

In a recent study, our collaborators examined cells from the *primary motor cortex (M1)* in twenty-five species. The M1 is a specialized brain region essential for controlling movements and learning motor skills. *Motor skills* are the abilities required to perform coordinated movements and tasks, such as walking, writing, and playing sports. The M1 is also studied for its involvement in disorders like Parkinson's disease [39]. It contains many cell types with specific gene expression signatures, making it an excellent subject for studying evolution and gene regulation [1].

Researchers use *RNA sequencing* to study the genetic composition of tissue. RNA sequencing provides detailed data on the *transcriptome* – the complete set of RNA transcripts produced by the *genome*. These RNA transcripts are copies of genes used to make proteins, which perform various functions in the body, such as building tissues and regulating processes. The resulting *gene expression profiles* indicate which genes are active and to what extent. In particular, so-called *marker genes* – genes that are directly linked to specific cellular identity and function – play a key role in the analysis of RNA sequencing data [32]. A common way to identify marker genes is *differential gene expression analysis* [33, 7]. Instead of only focusing on whether a gene is expressed or not in a tissue, gene expression is compared between two samples of the data. Genes highly expressed in one sample but not in the other are typically good candidates for marker genes.

Nowadays, (differential) gene expression analysis is commonly carried out on the *single-cell level*, allowing for the examination of genetic information in individual cells. *Single-cell transcriptomics* offers detailed gene expression profiles at the single-cell level, with datasets reaching sizes of hundreds of thousands of cells times tens of thousands of genes. On the single-cell level, differential gene expression is typically calculated between two selected groups of cells, for example, to compare cells belonging to two cell types. A simple way to do so is to first calculate the mean expression μ of a gene g for a set of cells C

$$\mu(g,C) = \frac{1}{|C|}\sum_{c\in C} E(g,c), \tag{1}$$

where $E(g,c)$ is the expression of the gene in each single cell c of the set of cells C. The differential expression ΔE of gene g between two sets of cells C_1 and C_2 can then simply be defined as the difference between those two mean expressions:

$$\Delta E(g,C_1,C_2) = \mu(g,C_1) - \mu(g,C_2) \tag{2}$$

When looking for marker genes, we use the comparison of cells of a specific type with the rest of the data. C_1 in Equation 2 then becomes the set of cells of that type of interest while C_2 becomes the set of *all* cells A of the particular dataset without C_1, that is $C_2 = A \setminus C_1$. Positive values indicate higher expression in selected cells, while negative values indicate lower expression.

The same genes can have different functions in different species, meaning that similar traits might arise from different genes across those species. For cross-species comparisons, genes from other species are mapped to a reference species, like humans. This standardization ensures a consistent number of orthologous genes, typically between 10,000 and 15,000, allowing for meaningful comparisons of gene expression values across species. Further, our collaborators reduce the original gene set into a lower-dimensional latent space, which is normalized across species for data integration and comparison. This lower-dimensional space is then used, for example for cross-species cell type labeling.

For analyzing such integrated data from many species it is valuable to augment it with information on evolutionary development. Such information is commonly available as *phylogenetic tree*. Closely related species appear near each other in the tree and share many similarities, while more distantly related species are positioned further apart and generally exhibit greater differences. Despite this general pattern, closely related species can still have major differences due to unique adaptations or evolutionary pressures. Similarly, distantly related species might share some similarities due to convergent evolution, where unrelated species develop

[3] https://github.com/AllenInstitute/scrattch

similar traits in response to similar environmental challenges. For example, in the primate tree, humans and chimpanzees are close as they share a common ancestor, resulting in similarities in gene function and physical traits. Despite this, humans have complex language and tool use, while chimpanzees exhibit unique social behaviors and physical traits. Conversely, macaques (Old World monkeys from Africa and Asia) and howler monkeys (New World monkeys from Central and South America) are more distantly related in the tree, having diverged from their common ancestor much earlier. Macaques have flexible diets and social structures, while howler monkeys have strong tails adapted for moving through trees. Both species have adapted to life in trees, demonstrating convergent evolution.

3 Related Work

Computational methods are essential in life sciences for analyzing single-cell data, where visual analysis aids in interpreting complex information [10]. Single-cell omics compares gene and protein expressions across cells. This process is complicated by high dimensionality and data sparsity [3], which can obscure important signals [24]. Advances in single-cell sequencing have improved these techniques, providing deeper insights into biological complexity by studying gene expression at the individual cell level. This approach reveals cellular diversity and function but introduces challenges, such as ensuring adequate sequence coverage and performing robust statistical analyses. Researchers are developing new computational tools and algorithms that enhance data interpretation and integration to address these challenges.

Analyzing these complex biological datasets through visualization has become an established area of research. For instance, MulteeSum [29] visualizes gene expression in Drosophila embryos over time and space. Single Cell Explorer [13], iS-CellR [31], and ASAP [15] use similar methods for single-cell gene expression by employing dimensionality reduction. Cytosplore [21] combines dimensionality reduction and clustering to analyze single-cell protein expression. Similarly, Brainscope [22] offers a dual-dimensionality reduction approach, linking gene expression data with sample information. Tools like ImaCytE [36], Facetto [27], and Vitessce [26] focus on visualizing spatial imaging data of single cells. While several tools exist to explore these biological datasets in detail, there are fewer options for comparing omics datasets.

3.1 Comparative Visualization

Tools and methods exist to make cell-type comparisons among datasets. ClusterMap [14] helps find and compare cell types in single-cell RNA sequencing data. SCope [9] also compares cell types across datasets. Zhao et al. introduced DA-seq [42], which finds groups of cells that change in number between different conditions. Joint t-SNE [41] is another tool that helps compare multiple datasets by generating similar low-dimensional embeddings.

Comparative analysis is prevalent in genome visualization but is less common in quantitative omics. The Cytosplore Simian Viewer [2] addresses this gap by enabling the comparison of single-cell transcriptomics data across different species in a pairwise manner. With Cytosplore EvoViewer, we expand the comparison paradigm to twenty-five species, to help make comparisons across a whole phylogenetic tree. We tackle the limitations of the Simian Viewer by going beyond pairwise comparisons to compare several species at a time and add quantitative and visual cues on 'what to compare' with clear goals to explore and identify interesting similarities or differences in high-dimensional transcriptomic datasets.

4 Domain Analysis

In comparative transcriptomics, researchers aim to analyze and compare gene expression across species to identify differences and similarities in their transcriptomes.

Earlier work addresses specific challenges for comparing a small number of (three to five) of species [25]. However, the ultimate goal is to paint a comprehensive picture covering broad phylogenetic relations. For example, our collaborators are currently working on a study covering twenty-five different species. They aim to identify specific gene markers that set (related) species apart by mapping gene expression across known phylogenetic hierarchies, focusing on identifying genes within cell types contributing to each species' unique traits. They are especially interested in understanding which cell type properties differentiate a species or groups of species from each other, and how these differences fluctuate based on their phylogenetic relation. They also aim to determine where these changes occur in the species hierarchy which can offer valuable insights into the evolutionary processes behind species heterogeneity.

4.1 Design Process

Cytosplore EvoViewer was developed through a participatory design approach [23] in collaboration with domain experts of varying seniority levels (scientist, assistant investigator, and senior investigator) who are also coauthors of this paper. We held bi-weekly online meetings to identify their practical needs and challenges.

In the first phase, we derived the requirements for the system. We clarified the properties of the data and identified analysis goals and tasks through open-ended discussions as well as by observing one of the collaborators during an exemplary analysis session. Based on this information, we conceptualized a visual analytics system, which was implemented iteratively in the second phase. Initially, early prototypes and functionality were presented during meetings, and feedback from collaborators guided further improvements. Later prototypes were deployed with our collaborators, who then tried to integrate it into their day-to-day workflow. This real-world experience led to additional feedback. Finally, after several iterations, we evaluated the system with biologists from the same institution who had not participated in the design process (Sec. 6).

4.2 Data

Transcriptomic datasets are represented as cell-by-gene matrices. Each row of the matrix corresponds to a cell and each column represents a gene. The values of each cell in the matrix indicate the expression levels of each gene for each cell. This format is commonly used in single-cell RNA sequencing (scRNA-seq) data analysis [35]. The number of genes can range from a few thousand to tens of thousands, depending on the organism and the sequencing technology. For example, human transcriptomic datasets typically represent around 20,000 to 25,000 genes. The exact number can also be affected by the quality of the sequencing data and the filtering criteria used during preprocessing.

Our collaborators collected over two million single-cell transcriptomes across twenty-five species resulting in a cell-by-gene matrix with 2,326,579 rows, each representing a single cell, and 7,754 columns, each representing a gene. Using an in-house pipeline, they built a three-level, hierarchical *consensus taxonomy*, assigning each cell a *Class*, *Subclass*, and *Cluster*, according to the gene expression and species. The cluster level of the hierarchy was still in flux at the time of writing and our tool was able to highlight some of the potential subdivisions during the evaluation (see Sec. 6.3). The hierarchy enables direct comparisons of cell type abundance at different levels and gene expression levels between species, from single-cell to cell-type level. Additionally, the dataset contains a phylogenetic tree and a metadata table with details on species and cell types, including prior evolutionary knowledge.

4.3 Domain Questions

Based on our discussions with the biologists, we identified the following questions that the biologists are mainly interested in:

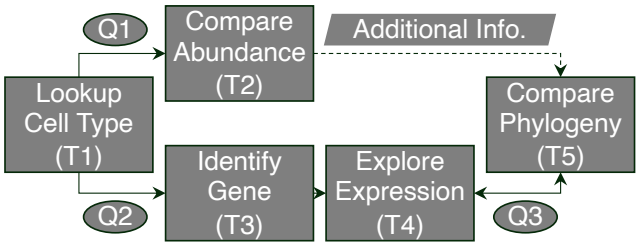

Figure 2: **The Cytosplore EvoViewer workflow**.

Q1 How do cell type abundances vary across species?

Comparing cell type abundance between species using transcriptomic data reveals differences in brain structure, function, and evolutionary traits. For example, capuchin monkeys have many specialized motor cortex neuronal cells for advanced motor skills, while squirrel monkeys have fewer L5 ET neurons with differential connectivity to the spinal cord, which correlates with their manual dexterity. Additionally, mice have more L5 IT cells than humans [19], highlighting distinct brain architectures and cognitive functions [12]. These comparisons provide valuable insights into brain development and evolution.

Q2 Which genes are potential marker genes for a cell type and how do they vary between species?

Marker genes help understand the uniqueness of cell types. Identifying which genes can serve as marker genes for a cell type and how they vary between species provides valuable insights. For example, the SST, PVALB, and VIP genes are markers for different types of interneurons [28], and comparing these genes across species helps researchers understand evolutionary differences in interneuron function, contributing to a broader understanding of health and disease mechanisms.

Q3 To what extent are marker genes related to subtrees of the phylogenetic tree or other prior evolutionary knowledge?

Studying marker genes within the context of the phylogenetic tree reveals their evolutionary trajectories across different species. For example, the FOXP2 gene, related to speech in humans, shows different changes in primates compared to other mammals. Similarly, the Pax6 gene is essential for eye formation across species, even though eye structures vary [34]. Mapping these genes onto the tree helps link them to specific evolutionary branches, offering insights into the development of unique traits and the relationship between marker genes, subtrees, and prior evolutionary knowledge.

Understanding these complex relationships between gene expression, and cell types across species provides insights into evolutionary biology, developmental biology, and gene regulation.

4.4 Tasks

We have identified the following set of tasks required to address the questions described in Sec. 4.3.

T1 **lookup** cell types at different levels of the taxonomy.

T2 **compare** cell abundance of a cell type between species.

T3 **identify** genes with high differential expression for a specific cell type (marker genes) in one or more species.

T4 for such genes, **explore** the relevance of these genes for the same cell type in other species.

T5 **compare** species according to prior knowledge, such as traits or phylogeny according to differences in marker gene properties.

5 Cytosplore EvoViewer

Cytosplore EvoViewer is designed with projects such as the BRAIN Initiative Cell Atlas Network[4] and the BRAIN Initiative Cell Census Network[5] in mind. These initiatives largely use visualization for communication, i.e., browsing curated data and results through web-based interfaces. In contrast, Cytosplore EvoViewer focuses on interactive, in-depth exploration, supporting quick reference and detailed analysis. Therefore, we implemented Cytosplore EvoViewer using C++ in the ManiVault framework [40] to enable high-performance computations using desktop hardware without a complicated server back-end.

5.1 Workflow

Cytosplore EvoViewer supports the analysis process through a workflow as illustrated in Figure 2. The workflow utilizes four distinct views described below: Cell Type Hierarchy, Gene Details Table, Single Cell Embedding Scatterplot, and Phylogenetic Tree. Biologists typically begin their exploration by selecting a cell type of interest (**T1**) in the Cell Type Hierarchy. The next step involves investigating how the abundance of these cells varies across species (**T2**). Users can extend the cell abundance comparison to identify species with high abundance or verify species abundance based on prior knowledge, such as species phylogeny or traits. Following this, users aim to identify the most differentially expressed genes and potential marker gene properties within the selected cells (**T3**). All this information is available in the Gene Details Table. Then marker gene properties of the identified gene are verified mainly by visual inspection of the gene distribution in the Single Cell Embedding (**T4**). Once this information is gathered, users can map and extend their findings by comparing them with prior knowledge (**T5**) in the Phylogenetic Tree.

5.2 Cell Type Hierarchy

To support **T1** – lookup cell types at different levels of the taxonomy – we use an icicle plot (Figure 3). The input data includes the cell type taxonomy from Sec. 4.2 and the corresponding cell abundances for each group. The taxonomy is a three-level hierarchy, allowing any tree-view visualization. For ease of selection, we use a space-filling visualization with large targets with the icicle plot providing easy click targets on all levels compared to the treemaps' focus on leafs and fitting our layout better than a sunburst diagram. Nodes are sized proportional to cell counts across all species and color hues represent cell types, following conventions from existing literature [25, 1] to ensure easy recognition for experts. Labels for cell types are added where space allows and are otherwise provided on hover. Identifying smaller clusters can be challenging due to the many cell types at lower levels. Thus, we offer the option to zoom into individual levels, enabling users to select and analyze one or multiple cell types at any hierarchy level.

[4] https://www.portal.brain-bican.org
[5] https://www.biccn.org

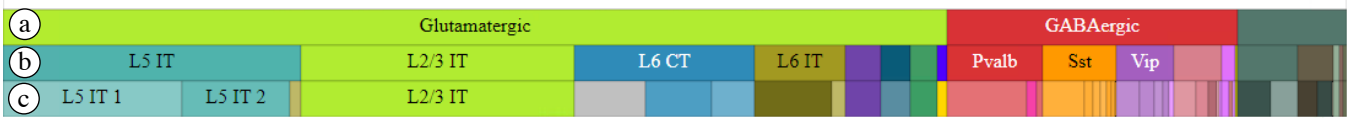

Figure 3: **Cluster Hierarchy View of the Primary Motor Cortex** visualized as an icicle plot. It displays cell types organized into three hierarchical levels: class (a), sub-class (b), and cluster (c). Each rectangle's size indicates cell counts, and colors denote different cell types. Hovering reveals cell type details while zooming, single, and multi-selections are supported for detailed analysis.

5.3 Cell Abundance and Gene Identification

We provide an enriched table view (Figure 4) that includes attributes derived from single-cell data (Sec. 4.2) to support the comparison of cell type abundance across species (**T2**), the identification of marker genes within species (**T3**), and the assessment of their relevance across species (**T4**). The table view is split into two parts: the left side (Figure 4a), which focuses on genes, and the main panel on the right (Figure 4b), which displays species, cell abundance (Sec. 5.3.1), and gene expression details (Sec. 5.3.2). This design separates the selection and detailed information views, allowing users to iteratively identify genes (**T3**) in the left view and explore their relevance across species (**T4**) in the right view, making it logically organized and accessible.

5.3.1 Cell Type Abundance

T2 – compare cell abundance of a cell type between species – is important to understand differences in brain function and cognitive abilities [18]. Our collaborators are especially interested in studying how the relative abundance of neuronal cells varies across species, as these cells are essential for brain activity and do not scale proportionally with brain size. Depending on their specific goals, they may also consider absolute counts.

We derive three new attributes from the data for each species, based on the active cell selection; 1) absolute number, 2) relative abundance as a fraction of neuronal or non-neuronal cells (depending on the lineage of the selection), and 3) relative abundance as fraction of the parent in the cell type hierarchy. While cell selection primarily occurs through the cluster hierarchy (Sec. 5.2), analysts can also make arbitrary selections (e.g., in the scatterplot, Sec. 5.4). Thus, we calculate those values on the fly.

We use the species-focused panel of the table view for abundance presentation. We add species icons to the species column for easy identification. Each of the three derived attributes and the total number of cells per species are displayed in separate columns, sorted by absolute abundance by default. Values are visualized as bars to help users identify patterns. The width of each bar reflects the range between the minimum and maximum values in the column. To view the exact value of a cell, users can hover over it or expand the column for a detailed view.

For arbitrary selections and selections on higher levels of the hierarchy, we provide a list of clusters from the lowest hierarchy level, including cell count at the bottom of the table (Figure 4c).

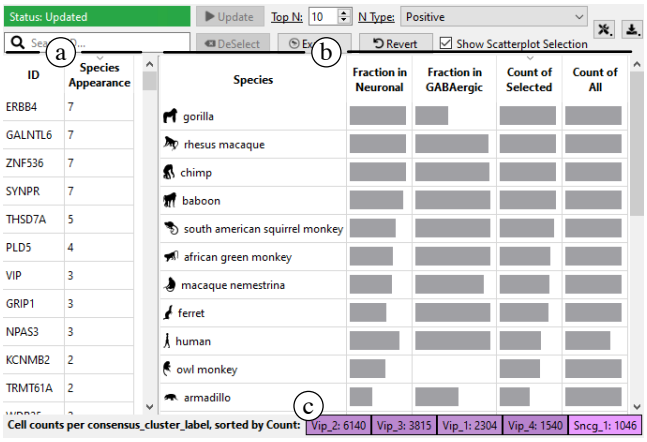

Figure 4: **Table view** for cell type abundance and gene expression exploration. The left panel (a) is gene-focused and lists all genes that are marker gene candidates in at least one species, the right panel (b) is focused on species and used for detailed analysis of cell abundance and differential expression. The bottom row (c) shows the distribution of the selected cells across different cell types.

5.3.2 Gene Identification

Cells express thousands of genes that define their characteristics and functions, contributing to similarity or uniqueness. However, not all genes are expressed in every cell, and gene expression varies among species. Therefore, it is essential to identify relevant genes (**T3**) and explore which species express them (**T4**).

To support **T3** – identify genes with high differential expression for a specific cell type (marker genes) in one or more species – we perform gene ranking according to their differential expression (Sec. 2) within each species. For the active cell selection, we calculate the mean differential expression of the selection versus the rest of the cells, according to Equation 2 for each species and gene. We then rank all genes for each species based on their maximum (absolute) differential expression. Finally, we list genes that are among the top n in at least one species in the left panel of the table (Figure 5a). Here, n is a user-defined parameter that may vary between cell types and genes. Generally, we consider genes with high ranks as potential markers for species requiring further exploration.

To support **T4** – for potential marker genes, explore the relevance of these genes for the same cell type in other species – we aggregate the number of species where each gene ranks in the top n and sort the left table panel accordingly. This enables quick identification of genes relevant to individual species and those that may indicate developmental conservation between species.

Using the gene ranking overview, analysts can explore detailed species information in the right table panel (Figure 5). This panel provides detailed insights on selected genes for each species, including abundance, rank, differential expression values, and mean expression in selected versus non-selected cells for raw comparison. We sort the species table by the rank of the selected gene and automatically select those species where the gene ranks in the top n for differential expression (Figure 5b). We use logarithmic scaling to visually enhance high ranks (smaller value) and employ an inverse rank bar chart to show larger bars for high-ranking genes. In some cases, a gene important to a species may rank just below the threshold, but since we display all species ordered by rank, such genes will appear near the top of the list, just below the highlighted species (e.g., Figure 5c). This allows users to include previously missed items or focus on specific genes of interest across species.

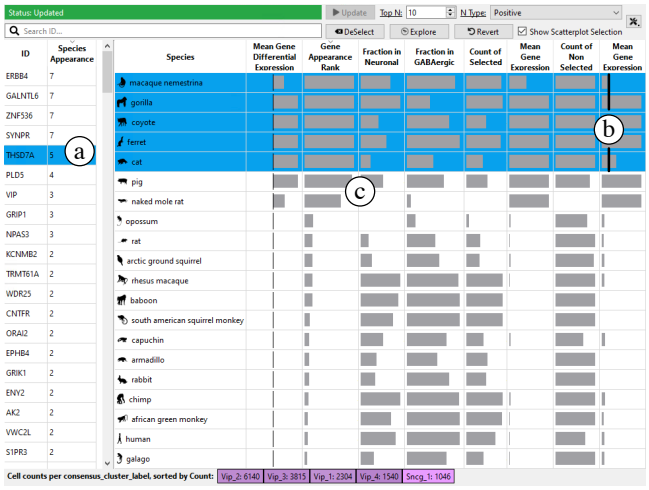

Figure 5: **Table view in gene expression mode.** After selecting a gene of interest (a) in the left panel the species table on the right is sorted according to the rank of the gene in the species. Species for which the gene ranks above a user-defined threshold are selected and highlighted (b) automatically. Showing the rank of the gene (c) allows expanding the selection on the fly during analysis.

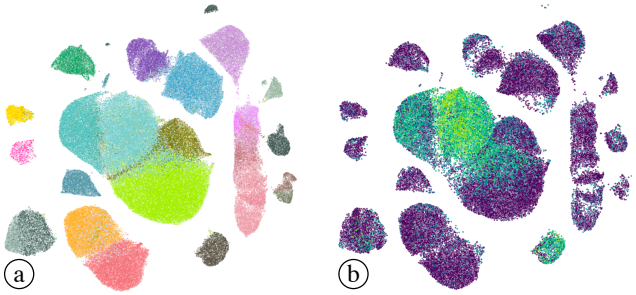

Figure 6: **Scatterplot of Two-Dimensional Embeddings** showing cells arranged by similarity, with colors indicating cell type (a) and gene expression (b), to visually verify differential expression.

5.4 Single Cell Embedding

To further support tasks identifying marker genes within species (**T3**) and assessing gene relevance across species (**T4**), we show a UMAP embedding of the cell-by-gene matrix in a scatterplot (Figure 6). A point in the plot represents a cell in the dataset and is placed based on similarity to others. The plot offers two coloring modes: in cell-selection mode (Figure 6a), we use the same cell-type coloring as described in Sec. 5.2. This coloring helps orient the user and identify potentially related cell types. In marker gene mode (Figure 6b), we color points based on the expression of the selected gene in the corresponding cell. Given that gene expression is a quantitative, sequential attribute, we use the Viridis colormap. This coloring mode is useful for quickly assessing whether a gene is specific to a cell type, consistent across a cluster, or distinct enough to be a marker gene. Shown cells are automatically filtered based on the selected species in the table view, enabling detailed analysis both within and across species.

5.5 Phylogenetic Tree

For **T5** – compare species according to prior knowledge, such as traits or phylogeny according to differences in marker gene properties – we integrate findings from previous tasks within the developmental relationship of species. The core idea is to use a phylogenetic tree encompassing the species in the dataset. Phylogenetic trees represent evolutionary relationships and can be constructed from various properties such as gene frequencies or traits. We received such a tree from our collaborators, where distances between nodes are inscribed as edge attributes. The tree is visualized as a node-link diagram with the root on the left, leaves on the right, and distance encoded by horizontal edge lengths (Figure 7).

To compare information from cell-type abundance and gene expression with the phylogeny, we considered different methods. In our initial prototype, we used an explicit tree comparison approach, calculating dendrograms for each property of interest (e.g., cell type abundance) and used Bremm et al.'s [5] approach to visualize difference trees. However, the changing layouts were confusing for our collaborators, and the encoding of differences on the nodes was hard to interpret.

Subsequently, we adopted a simpler approach, preserving the phylogenetic tree structure while encoding a single attribute (e.g., cell type abundance, appearance rank, and gene expression, Figures 7a, 7b, and 7c, respectively) using color-mapping on the nodes. For gene expression, we used the Viridis colormap as for other views (Sec. 5.4). We use distinct colormaps selected for other attributes to visually indicate the different attributes. We chose the Magma colormap, for differential expression. For rank and abundance, we use Colorbrewer's OrRd and BuPu sequential colormaps [17], respectively. For rank, we inverted and log-transformed the colormap to emphasize lower-ranked values (Sec. 5.3.2). For leaf nodes, we directly used the species values and averaged those for inner nodes in the tree.

The tree also enables species selection. We tested various methods, ultimately choosing an approach with a single selection for one or multiple species. In early prototypes, we tested various methods for two selections, which enabled differential gene expression calculations between the selections. One user study participant expressed interest in comparing species using differential gene expression (see also Sec. 6.3). However, during prototyping, this functionality was removed to maintain a more focused approach.

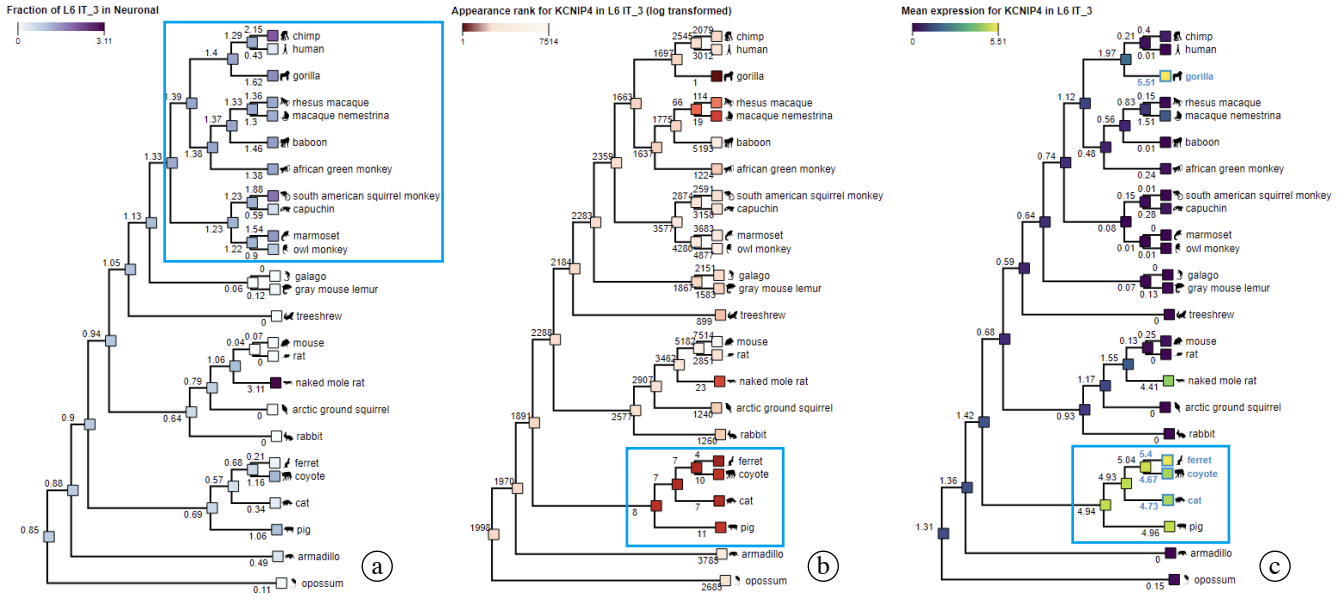

Figure 7: **The tree view** displays the phylogeny of various species, with nodes colored by species abundance (a), appearance rank (b), and mean expression (c) of a selected gene in a cell type. These values are averaged from the bottom up to highlight where in the hierarchy expression changes occur, helping to understand what makes certain groups of species unique from one another. The highlighted branches of the trees show high abundance in the top part of the tree (a), lower rank (b), and high mean expression (c) in the bottom branches of the tree.

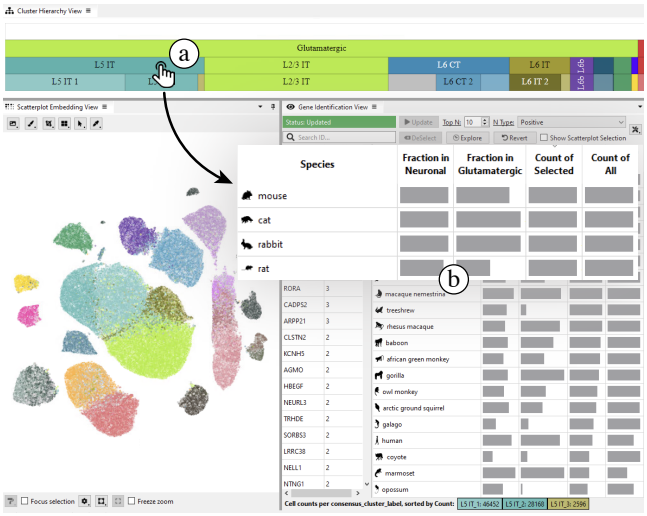

Figure 8: **The first step in the case study** is to identify the L5IT cells in the cell type hierarchy (a), select it, and explore the corresponding abundances in the detail table view (b).

6 Validation

To validate Cytosplore EvoViewer, we set up a small case study covering the three goals and their respective tasks (Sec. 4), as laid out in Sec. 6.1. For this study, we developed a set of user tasks to determine whether users could successfully recreate the intended outcomes. Additionally, we included a short questionnaire to gather feedback on the ease of use. The setup of this study is described in detail in Sec. 6.2 and the results in Sec. 6.3.

6.1 Case Study

The study focuses on the Layer 5 Intratelencephalic (L5IT) neuronal cell type in the Primary Motor Cortex (M1) and the RAR Related Orphan Receptor Beta (RORB) gene within this cell type.

We will explore the abundance of L5IT cells across different species in the dataset presented in Sec. 4.2. By selecting the species list in the table view (Figure 8a), we can see species ranked by absolute abundance, with options to display the fraction of all neuronal cells or the direct parent in the hierarchy (Figure 8b). This allows us to identify the mouse as having the highest absolute abundance while highlighting the impact of varying sampling across species.

Next, we will analyze the RORB gene, which is typically a marker for Layer 4 neurons. Recent findings suggest that RORB expression may extend to neurons near the L4/L5 boundary [16], indicating its potential role as a marker for a subset of L5IT neurons. Selecting the RORB gene (Figure 9a), reveals it as the top-ranked differentially expressed gene for several species (Figure 9b). The highlighting and sorting features enable easy identification in the table, while the UMAP embedding (Figure 9c) further supports higher expression in L5IT cells for these species.

The table indicates a steep increase in rank further down, suggesting that RORB lacks relevant marker functionality for those species. For pig, the first species where RORB is not in the top ten genes, it ranks 22*nd* with a mean differential expression value of 0.79 (Figure 10a), deserving further investigation. The UMAP filtered for pig cells shows a split into purple (left) and green (right) compartments (Figure 10b), indicating low and high expression, respectively. This suggests a division in cell functionality. One participant in our user study noted this and created further investigations, concluding that RORB remains a good marker for L5IT1 but not for L5IT2 (see also Sec. 6.3).

In the final step, we will explore whether these findings align with phylogeny or can be attributed to specific traits. Figure 7c presents a phylogenetic tree of all species in the dataset, overlaid with the mean differential expression of the RORB gene in L5IT cells. The gene shows high differential expression in a sub-tree comprising ferret, coyote, cat, and pig. Notably, this sub-tree includes pig, which would not have been identified using a strict top ten gene threshold, reinforcing the relevance of RORB for L5IT in pig. Conversely, the gorilla, which exhibits the highest absolute differential expression, appears as a clear outlier in the tree, suggesting the need for further investigation.

6.2 Participants and Setup

We recruited three expert participants from our collaborating institute, which is currently analyzing the data presented in Sec. 4.2 for the user study. We will refer to them as P1, P2, and P3. P1 is a male scientist with nine years of experience; P2 is a female post-baccalaureate with one year of experience; and P3 is a male scientist with eight years of experience. A summary is shown in Table 1. None of the participants were familiar with the software before this study or involved in its development.

For the evaluation, we conducted separate Zoom calls with each participant, scheduling each session for one hour, including 15 minutes for a brief introduction and setup if needed. Participants could choose to run the software on their computers or through a remote session on one of our workstations to minimize setup. Ultimately, all participants opted for the remote setup. All sessions were recorded, including the shared screen with the application.

Figure 9: **Inspecting table view** for the selected RORB gene (a) reveals six species for which the gene ranks as a potential marker gene for L5IT (b). The UMAP embedding (c) shows high expression for those species in a confined area, comprising L5IT cells.

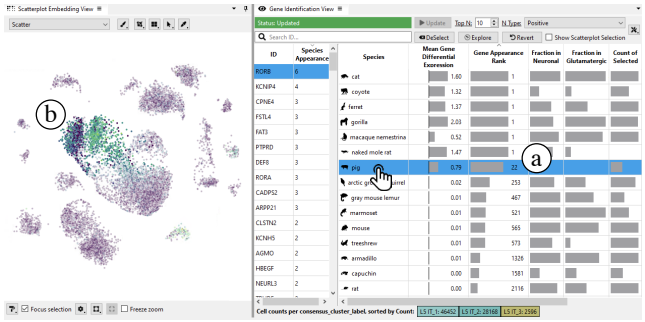

Figure 10: **Further analysis** shows that pig also ranks high but not in the top 10 (a). Inspecting the UMAP embedding (b) reveals a split of the cluster into a mainly low and a mainly high expression area indicating that the Cytosplore EvoViewer can reveal detailed differences in the cell type composition between species.

Table 1: **User Study Participants.** The participants' research area (**Research**), **Role**, years of experience (**Exp.**), and **Gender** were queried in the questionnaire. All participants used the software through screen-sharing one of our workstations (**Setup**). **SUS** scores were calculated based on the questionnaire.

	Research	Role	Exp.	Gender	Setup	SUS
P1	Neuroscience	Scientist	9y	male	remote	70.0
P2	Neuroscience	PostBacc	1y	female	remote	82.5
P3	CompBio	Scientist	8y	male	remote	87.5

Before the session, we provided participants with a seven-minute video presenting the software features (see Supplemental). We also reviewed the functionality during the introduction. For testing the software, we derived three steps from the case study, presented in Sec. 6.1 which we asked participants to complete. After each step, participants filled out a brief questionnaire that included one question assessing their success and understanding of the step, along with three statements about the software rated on a five-point Likert scale from Strongly Agree to Strongly Disagree (Table 2). We also asked for open-ended comments for each step. After completing all steps, participants responded to the standard System Usability Scale (SUS) [6] and provided additional open feedback. The completed questionnaires are included in the supplemental material.

6.3 User Study

Following the case study presented in Sec. 6.1, we have defined three short exercises, each covering one of the questions presented in Sec. 4.3 and their related abstracted tasks (Sec. 4.4).

In the first exercise, we asked participants to identify the species with the highest abundance of L5IT cells. This exercise is targeted to cover **Q1** and the corresponding tasks **T1** (lookup cell types at different levels of the taxonomy) and **T2** (compare cell type abundance across species). It is intentionally kept simple to provide a smooth introduction to the tool. As described in the case study, users can directly select the desired cell type in the cluster hierarchy view. The detail table view in the center updates accordingly, allowing participants to see both absolute and relative cell abundances, with the option to sort the table by either metric. To test success, we asked the participants to indicate the species they found on the questionnaire. P1 and P2 answered this with the correct species for absolute abundance and P3 answered correctly for relative abundance. As shown in Table 2, all participants strongly agreed that *It was easy to select the L5IT cells* (statement S1 of the questionnaire) and *It was easy to identify the species with the highest abundance of L5IT cells* (S2). P1 indicated that he was confident and P2 and P3 were very confident in the outcome (S3).

In the open feedback, P2 and P3 both indicate the helpfulness of seeing both absolute and relative abundances side-by-side to account for sampling differences. P2: *"[...] it's easy to see differences in cell type proportions across species without bias towards species with more donors."* and *"Columns are helpful to see total counts of cells alongside fraction for species such as naked mole rat with limited sampling but similar abundances"* (P3). P1 indicated that the separation of the table can be a bit confusing, as it was hard to see the split between gene selection and detail panel on a small screen *"I was initially confused about the gene pane being aligned with the species and didn't see that they weren't a single pane on my small screen"*. This may be partly due to the scaling of the remote workstation's higher resolution to the participant's laptop screen.

Table 2: **Summary of participants' responses** to statements (S) in our targeted questionnaire. Responses were measured on a 5-point Likert scale ranging from *very positive* (++) to *very negative* (−−). Color hues represent different participants (P1 ●, P2 ●, P3 ●). Note: Rows with no responses (–, −−) are omitted for clarity.

S	1	2	3	4	5	6	7	8	9
++	●●●	●●●	●●	●●	●●	●	●●●	●	
+			●	●	●	●		●●	●●●
○						●			

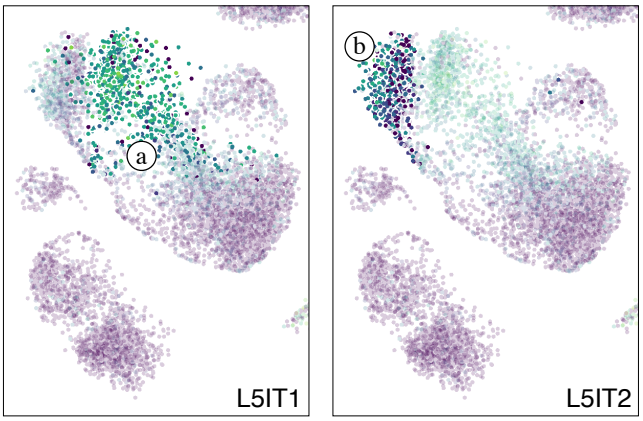

Figure 11: **Detailed inspection in the UMAP** reveals a split of the L5IT cells for RORB in pig into the high expression L5IT1 (a) and low expression L5IT2 (b), explaining, why the RORB gene is not among the top-ranked differential expressed genes for L5IT in pig.

In the second exercise, we enhance flexibility by focusing on potential marker genes and their variation across species (**Q2**). This aligns with tasks **T3** (identify relevant genes) and **T4** (explore which species express them). We first asked participants to determine which species could consider the RORB gene as a potential marker for the previously selected L5IT cells and to visually verify this. We then opened the exercise and instructed participants to examine additional candidates in more detail. We calculated the differential expression for each gene in the selected cells compared to the rest and ranked all genes by species (Sec. 5.3.2). In the default setting, we highlight all species where the gene is ranked among the top ten differentially expressed genes. This approach is relatively conservative, meaning that species with slightly lower ranks may still be of interest. Since we display the rank and sort the table accordingly, these species can be easily explored. In the first part, all participants identified the highest-ranked species and inspected the expression of the RORB gene using the single-cell embedding in the scatterplot for verification. P2 and P3 strongly agreed to *It was easy to confirm whether the RORB gene showed marker gene properties for a chosen species* (S4) and *I am confident about the marker gene-related findings of the RORB gene in the L5IT cells* (S5), while P1 agreed on both.

During the session we observed all participants use the embedding view to inspect it for further details and already made some observations on the distribution of cells with high RORB expression; *"The subsetted umap for each species makes it easy to see the expression of the gene and compare it across species"*. In particular, P3 indicated that *"RORB is a good marker for the 6 species although only for L5IT1 after closer inspection."* L5IT1 is one of two subsets of L5IT in the dataset and indeed for several species, the expression of RORB splits the L5IT here, which can support the tentative splitting of L5IT into the two not finalized sub-clusters (Sec. 4.2). P3 spent some time to further explore this, doing individual selections of the subgroups to verify the observation.

When further exploring the data the participants' reactions were

more diverse on *It was clear to me where I can find additional details about the species* (S6) ranging from netural (P1), agree (P2) to strongly agree (P3). This is also reflected in our observations from the session. While P1 indicated that he was already using the phylogenetic tree, which might have been a bit overloading at this time, P2 and P3 were deeply drilling into different aspects of the data. They were mainly using the embedding view to explore other species. P2 noted, *"In the UMAPs it looks like for some species it is highly expressed in L5IT neurons but also in other subclasses which is why it isn't a good maker"*. *P3 added, "It could be argued that RORB is also a good marker in Pig which didn't meet the threshold to be included"*. Additionally, P3 remarked on cell types, stating *"Astrocyte expression of RORB is clear on the UMAP"* (Figure 11). Nonetheless, P2 and P3 express a desire for additional non-visual verification, specifically through the inclusion of p-values and fold changes in the table. They also suggest the option to export selections for further analysis. As one participant noted, "I think it would be beneficial to include the p-value and log fold change value for the genes to express the significance of a gene as a marker."

Finally, we aimed to test **Q3** and open exploration, combining all defined tasks **T1**–**T5** in the third exercise that aims to explore how genes function across different phylogenies and evolutionary contexts. Participants were asked to use the phylogenetic tree view to explore whether relations between the previously identified cell-gene-species combination can be verified and if further information can be derived from the hierarchy. We explicitly asked the participants to test other cell types and genes of their interest to see if the tool supports free exploration. There is generally very positive feedback again, with all participants strongly agreeing that *The phylogenetic tree helped me identify interesting species for specific cell/gene combinations* (S7). P1 and P3 agree and P2 strongly agrees that *It was easy to compare expression differences between species groups in the phylogenetic tree* (S8). Additionally, all participants agree that they are confident about their findings (S9).

The comments suggest that the participants found the phylogenetic tree in combination with the other views very helpful. While P1 had already used it in the previous exercise, P2 indicates that *"The tree worked well to show relationships within and between clades* [a grouping of organisms that are composed of a common ancestor and all its lineal descendants] *as well as outliers in gene expression"*. P2 provided a concrete example regarding carnivores, highlighting the preservation of the RORB gene *"It's clear that the carnivora clade has RORB as a top marker common between all species within it which it great to see it conserved in that clade!"*. P3 states further *"The phylogenetic tree was very helpful to see which clades had expression of conserved marker genes, really great way to visualize the expression across"*. P1 liked the possibility of selecting multiple species in the tree view but would like to run differential gene expression analysis on the data (*"Selecting multiple species in the phylo tree was very useful, and would be even better if you could run D[ifferential]E[expression] between them"*. This functionality was available in an earlier version of the software but was removed for simplicity, as it was not part of the identified goals. Both P2 and P3 were at times confused by our presentation of the different attributes on the phylogenetic tree, P2 indicated that selecting the attribute directly through the table would be easier while P3 suggested using different default color maps for the different types of attributes, especially the latter being a clear oversight in our initial design.

In summary, the feedback was positive as shown in Table 2 as well as the SUS score (Table 1). All participants successfully completed the tasks and expressed high confidence in their findings, particularly regarding the properties of marker genes. The phylogenetic tree, which illustrates derived properties across species and clades, received particular praise from all three participants (e.g., *"I really enjoyed the phylogenetic tree tab. It was so convenient to quickly be able to see what clades share d[ifferential] e[expressed] genes, proportions, etc."* (open feedback P3)). We also identified room for further improvement that we are integrating in future updates. These include features such as exporting selections and enhancing visualizations, as well as adding properties like statistical tests and fold changes for identifying marker genes.

7 Design Reflections

Reflecting on the design process, we learned the importance of deploying prototypes with collaborators as early as possible to gather actionable feedback. One critical issue identified post-deployment was the discoverability of features. Despite providing a written manual and a tutorial video, the large number of features was overwhelming, leading to some being overlooked. To address this, we implemented integrated shortcut menus in every view. This enhancement was well-received by collaborators, complementing traditional onboarding materials. Future plans include adding more onboarding features, as suggested by Stoiber et al. [37].

A key limitation of our implementation is its reliance on system RAM for data storage. The current 25-species dataset exceeds 70GB, with expected growth. To address this, we are exploring out-of-core approaches that load only actively processed data into memory and lower-precision data types like bfloat16. Given the sparsity of RNA sequencing data, even binary storage is feasible [4], though precision reduction requires careful error analysis. Calculations like differential gene expression are performed on the fly using parallel processing, but larger datasets may necessitate smart caching strategies. Our collaborators provided the data as monolithic files. However, the field is transitioning to out-of-memory, partial I/O data formats like Zarr. We are currently exploring ways to support these formats in our system to reduce the reliance on large system memory further.

For visual scalability, we focused on managing species numbers in cell-focused views, implementing zoom and pan features to navigate cluster hierarchies and examine subclasses or child clusters.

The application was developed as a desktop application to meet the project's specific needs. While web-based solutions offer easier deployment, the desktop approach ensures real-time computations, and parallel processing, which are hard to achieve in client-side web-processing and secure handling of sensitive data—critical for performance and confidentiality, which are harder to achieve in server-based web environments.

Our pilot user study involved a small group of participants due to the highly specialized nature of the problem. The three expert participants provided invaluable feedback that complemented earlier input from collaborators, resulting in targeted improvements. The specificity of the problem complicates comparisons with existing tools, as no off-the-shelf solution addresses the same challenges.

8 Conclusion

In this paper, we have presented Cytosplore EvoViewer, a tool to support the exploration of multi-species single-cell transcriptomics data, supporting the identification of conserved cell and gene properties across species. Cytosplore EvoViewer was developed using a participatory design approach in close collaboration with our expert collaborators. Our expert evaluation indicates that the tool can support their workflow and lead to useful discoveries. Our current workflow is focused on identifying cell type properties and genes of interest and exploring their relation to species' development. A future step in our current workflow could be to find further related marker genes or cell types by working backward from the selected species. For example, assuming the analyst has found a group of species that seems correlated with previously identified cell type abundances or genes of interest, they could now query the data to identify other cell types and genes that are conserved across the identified species.

Acknowledgments

The authors thank Jessie Puckett, Matthew Schmitz, and Nelson Johansen for graciously providing their time to evaluate Cytosplore EvoViewer. We thank our Allen Institute partners for delivering the cross-species transcriptomic datasets, taxonomies, annotations, and feedback. Further, we would like to thank the anonymous reviewers for their valuable feedback.

This work received financial support from the NWO TTW project 3DOMICS (NWO: 17126) and the NWO Gravitation project BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (NWO: 024.004.012).

References

[1] T. E. Bakken, N. L. Jorstad, Q. Hu, B. B. Lake, W. Tian, B. E. Kalmbach, M. Crow, R. D. Hodge, F. M. Krienen, S. A. Sorensen, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature*, 598(7879):111–119, 2021. doi: 10.1038/s41586-021-03465-8 1, 2, 4

[2] S. Basu, J. Eggermont, T. Kroes, N. Jorstad, T. Bakken, E. Lein, B. Lelieveldt, and T. Höllt. Cytosplore Simian Viewer: Visual exploration for multi-species single-cell RNA sequencing data. In *Proc. VCBM*, pp. 111–120. EUROGRAPHICS, Eindhoven, NL, 2023. doi: 10.2312/VCBM.20231219 3

[3] N. Bhattacharya, C. C. Nelson, G. Ahuja, and D. Sengupta. Big data analytics in single-cell transcriptomics: Five grand opportunities. *WIREs Data Min. Knowl. Discovery*, 11(4):article no. e1414, 2021. doi: 10.1002/widm.1414 3

[4] G. A. Bouland, A. Mahfouz, and M. J. T. Reinders. Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. *Genome Biol.*, 24(1):article no. 86, 2023. doi: 10.1186/s13059-023-02933-w 9

[5] S. Bremm, T. von Landesberger, M. Hess, T. Schreck, P. Weil, and K. Hamacherk. Interactive visual comparison of multiple trees. In *Proc. VAST*, pp. 31–40. IEEE, New York, 2011. doi: 10.1109/vast.2011.6102439 6

[6] J. Brooke. SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, eds., *Usability Evaluation in Industry*, pp. 189–194. Taylor and Francis, 1996. 8

[7] J. Costa-Silva, D. Domingues, and F. M. Lopes. RNA-seq differential expression analysis: An extended review and a software tool. *PLoS One*, 12(12):article no. e0190152, 2017. doi: 10.1371/journal.pone.0190152 2

[8] C. Darwin. *The origin of species: By means of natural selection, or the preservation of favoured races in the struggle for life*. Cambridge University Press, 1859. doi: 10.1017/cbo9780511694295 2

[9] K. Davie, J. Janssens, D. Koldere, M. De Waegeneer, U. Pech, L. Kreft, S. Aibar, S. Makhzami, V. Christiaens, C. Bravo González-Blas, et al. A single-cell transcriptome atlas of the aging Drosophila brain. *Cell*, 174(4):982–998, 2018. doi: 10.1016/j.cell.2018.05.057 3

[10] J. Ding, A. Alavi, M. R. Ebrahimkhani, and Z. Bar-Joseph. Computational tools for analyzing single-cell data in pluripotent cell differentiation studies. *Cell Rep. Methods*, 1(6):article no. 100087, 2021. doi: 10.1016/j.crmeth.2021.100087 3

[11] L. Duhan, D. Kumari, M. Naime, V. S. Parmar, A. K. Chhillar, M. Dangi, and R. Pasrija. Single-cell transcriptomics: background, technologies, applications, and challenges. *Mol. Biol. Rep.*, 51(1):article no. 600, 2024. doi: 10.1007/s11033-024-09553-y 2

[12] R. Fang, C. Xia, J. L. Close, M. Zhang, J. He, Z. Huang, A. R. Halpern, B. Long, J. A. Miller, E. S. Lein, and X. Zhuang. Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH. *Science*, 377(6601):56–62, 2022. doi: 10.1126/science.abm1741 4

[13] D. Feng, C. E. Whitehurst, D. Shan, J. D. Hill, and Y. G. Yue. Single Cell Explorer, collaboration-driven tools to leverage large-scale single cell RNA-seq data. *BMC Genomics*, 20(1):article no. 676, 2019. doi: 10.1186/s12864-019-6053-y 3

[14] X. Gao, D. Hu, M. Gogol, and H. Li. Clustermap: Compare multiple single cell RNA-seq datasets across different experimental conditions. *Bioinformatics*, 35(17):3038–3045, 2019. doi: 10.1093/bioinformatics/btz024 3

[15] V. Gardeux, F. P. A. David, A. Shajkofci, P. C. Schwalie, and B. Deplancke. ASAP: A web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics*, 33(19):3123–3125, 2017. doi: 10.1093/bioinformatics/btx337 3

[16] T. A. Hage, A. Bosma-Moody, C. A. Baker, M. B. Kratz, L. Campagnola, T. Jarsky, H. Zeng, and G. J. Murphy. Synaptic connectivity to L2/3 of primary visual cortex measured by two-photon optogenetic stimulation. *eLife*, 11:article no. e71103, 2022. doi: 10.7554/elife.71103 7

[17] M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. doi: 10.1179/000870403235002042 6

[18] S. Herculano-Houzel. The human brain in numbers: A linearly scaled-up primate brain. *Front. Hum. Neurosci.*, 3:article no. 31, 2009. doi: 10.3389/neuro.09.031.2009 5

[19] R. D. Hodge, T. E. Bakken, J. A. Miller, K. A. Smith, E. R. Barkan, L. T. Graybuck, J. L. Close, B. Long, N. Johansen, O. Penn, et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772):61–68, 2019. doi: 10.1038/s41586-019-1506-7 4

[20] A. A. Hoffmann and P. A. Parsons. *Extreme environmental change and evolution*. Cambridge University Press, 1997. 2

[21] T. Höllt, N. Pezzotti, V. van Unen, F. Koning, E. Eisemann, B. Lelieveldt, and A. Vilanova. Cytosplore: Interactive immune cell phenotyping for large single-cell datasets. *Comput. Graphics Forum*, 35(3):171–180, 2016. doi: 10.1111/cgf.12893 3

[22] S. M. Huisman, B. van Lew, A. Mahfouz, N. Pezzotti, T. Höllt, L. Michielsen, A. Vilanova, M. J. Reinders, and B. P. Lelieveldt. BrainScope: Interactive visual exploration of the spatial and temporal human brain transcriptome. *Nucleic Acids Res.*, 45(10):article no. gkx046, e83, 2017. doi: 10.1093/nar/gkx046 3

[23] S. Jänicke, P. Kaur, P. Kuzmicki, and J. Schmidt. Participatory visualization design as an approach to minimize the gap between research and application. In *Proc. VisGap*, pp. 35–42. EUROGRAPHICS, Eindhoven, NL, 2020. doi: 10.2312/VISGAP.20201108 3

[24] Y. Jeong, J. Ronen, W. Kopp, P. Lutsik, and A. Akalin. ScMaui: A widely applicable deep learning framework for single-cell multiomics integration in the presence of batch effects and missing data. *BMC Bioinf.*, 25(1):article no. 257, 2024. doi: 10.1186/s12859-024-05880-w 3

[25] N. L. Jorstad, J. H. Song, D. Exposito-Alonso, H. Suresh, N. Castro-Pacheco, F. M. Krienen, A. M. Yanny, J. Close, E. Gelfand, B. Long, et al. Comparative transcriptomics reveals human-specific cortical features. *Science*, 382(6667):article no. ade9516, 2023. doi: 10.1126/science.ade9516 1, 3, 4

[26] M. S. Keller, I. Gold, C. McCallum, T. Manz, P. V. Kharchenko, and N. Gehlenborg. Vitessce: Integrative visualization of multimodal and spatially-resolved single-cell data. *OSF Preprints*, 2021. doi: 10.31219/osf.io/y8thv 3

[27] R. Krueger, J. Beyer, W.-D. Jang, N. W. Kim, A. Sokolov, P. K. Sorger, and H. Pfister. Facetto: Combining unsupervised and supervised learning for hierarchical phenotype analysis in multi-channel image data. *IEEE Trans. Visual Comput. Graphics*, 26(1):227–237, 2020. doi: 10.1109/tvcg.2019.2934547 3

[28] R. Machold and B. Rudy. Genetic approaches to elucidating cortical and hippocampal GABAergic interneuron diversity. *Front. Cell. Neurosci.*, 18:article no. 1414955, 2024. doi: 10.3389/fncel.2024.1414955 4

[29] M. Meyer, T. Munzner, A. DePace, and H. Pfister. MulteeSum: A tool for comparative spatial and temporal gene expression data. *IEEE Trans. Visual Comput. Graphics*, 16(6):908–917, 2010. doi: 10.1109/tvcg.2010.137 3

[30] S. I. O'Donoghue, A.-C. Gavin, N. Gehlenborg, D. S. Goodsell, J.-K. Héricheé, C. B. Nielsen, C. North, A. J. Olson, J. B. Procter, D. W. Shattuck, et al. Visualizing biological data-now and in the future. *Nat. Methods*, 7(S3):S2–S4, 2010. doi: 10.1038/nmeth.f.301 2

[31] M. V. Patel. iS-CellR: A user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. *Bioinformatics*, 34(24):4305–

4306, 2018. doi: 10.1093/bioinformatics/bty517 3

[32] Y. Qiu, J. Wang, J. Lei, and K. Roeder. Identification of cell-type-specific marker genes from co-expression patterns in tissue samples. *Bioinformatics*, 37(19):3228–3234, 2021. doi: 10.1093/bioinformatics/btab257 2

[33] D. Rosati, M. Palmieri, G. Brunelli, A. Morrione, F. Iannelli, E. Frullanti, and A. Giordano. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Comput. Struct. Biotechnol. J.*, 23:1154–1168, 2024. doi: 10.1016/j.csbj.2024.02.018 2

[34] C. Scharff and J. Petri. Evo-devo, deep homology and FoxP2: Implications for the evolution of speech and language. *Philos. Trans. R. Soc. B Biol. Sci.*, 366(1574):2124–2140, 2011. doi: 10.1098/rstb.2011.0001 4

[35] M. E. R. Shafer. Cross-species analysis of single-cell transcriptomic data. *Front. Cell Dev. Biol.*, 7:article no. 175, 2019. doi: 10.3389/fcell.2019.00175 1, 3

[36] A. Somarakis, V. Van Unen, F. Koning, B. Lelieveldt, and T. Hollt. ImaCytE: Visual exploration of cellular micro-environments for imaging mass cytometry data. *IEEE Trans. Visual Comput. Graphics*, 27(1):98–110, 2021. doi: 10.1109/tvcg.2019.2931299 3

[37] C. Stoiber, D. Ceneda, M. Wagner, V. Schetinger, T. Gschwandtner, M. Streit, S. Miksch, and W. Aigner. Perspectives of visualization onboarding and guidance in VA. *Vis. Inform.*, 6(1):68–83, 2022. doi: 10.1016/j.visinf.2022.02.005 9

[38] J. Tooby and L. Cosmides. On the universality of human nature and the uniqueness of the individual: The role of genetics and adaptation. *J. Pers.*, 58(1):17–67, 1990. doi: 10.1111/j.1467-6494.1990.tb00907.x 2

[39] C. F. Underwood and L. C. Parr-Brownlie. Primary motor cortex in Parkinson's disease: Functional changes and opportunities for neurostimulation. *Neurobiol. Dis.*, 147:article no. 105159, 2021. doi: 10.1016/j.nbd.2020.105159 2

[40] A. Vieth, T. Kroes, J. Thijssen, B. van Lew, J. Eggermont, S. Basu, E. Eisemann, A. Vilanova, T. Höllt, and B. Lelieveldt. ManiVault: A flexible and extensible visual analytics framework for high-dimensional data. *IEEE Trans. Visual Comput. Graphics*, 30(1):175–185, 2023. doi: 10.1109/tvcg.2023.3326582 4

[41] Y. Wang, L. Chen, J. Jo, and Y. Wang. Joint t-SNE for comparable projections of multiple high-dimensional datasets. *IEEE Trans. Visual Comput. Graphics*, 28(1):623–632, 2022. doi: 10.1109/tvcg.2021.3114765 3

[42] J. Zhao, A. Jaffe, H. Li, O. Lindenbaum, E. Sefik, R. Jackson, X. Cheng, R. A. Flavell, and Y. Kluger. Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proc. Natl. Acad. Sci.*, 118(22):article no. e2100293118, 2021. doi: 10.1073/pnas.2100293118 3