Cytosplore Simian Viewer

Visual Exploration for Multi-Species Single-Cell RNA Sequencing Data

Basu, Soumyadeep; Eggermont, Jeroen; Kroes, Thomas; Jorstad, Nikolas ; Bakken, Trygve ; Lein, Ed; Lelieveldt, Boudewijn; Höllt, Thomas

DOI
10.2312/vcbm.20231219

Publication date
2023

Document Version
Final published version

Published in
Eurographics Workshop on Visual Computing for Biology and Medicine

Eurographics Workshop on Visual Computing for Biology and Medicine (2023)
C. Hansen, J. Procter, and R. G. Raidou (Editors)

Cytosplore Simian Viewer: Visual Exploration for Multi-Species Single-Cell RNA Sequencing Data

Soumyadeep Basu[1,2] Jeroen Eggermont[1] Thomas Kroes[1] Nikolas Jorstad[3]
Trygve Bakken[3] Ed Lein[3] Boudewijn Lelieveldt[1] Thomas Höllt[2]

[1]Leiden University Medical Center, Leiden, NL [2]Delft University of Technology, Delft, NL [3]Allen Institute for Brain Science, Seattle, WA, US

Figure 1: ***Screenshot of Cytosplore Simian Viewer*** *showing data on cell composition within the middle temporal gyrus of humans and chimpanzees with four different views: (a) heatmap, (b,c) scatterplots, (d) differential expression table, and (e) gene expression bar charts.*

Abstract

With the rapid advances in single-cell sequencing technologies, novel types of studies into the cell-type makeup of the brain have become possible. Biologists often analyze large and complex single-cell transcriptomic datasets to enhance knowledge of the intricate features of cellular and molecular tissue organization. A particular area of interest is the study of whether cell types and their gene regulation are conserved across species during evolution. However, in-depth comparisons across species of such high-dimensional, multi-modal single-cell data pose considerable visualization challenges. This paper introduces Cytosplore Simian Viewer, a visualization system that combines various views and linked interaction methods for comparative analysis of single-cell transcriptomic datasets across multiple species. Cytosplore Simian Viewer enables biologists to help gain insights into the cell type and gene expression differences and similarities among different species, particularly focusing on comparing human data to other species. The system validation in discovery research on real-world datasets demonstrates its utility in visualizing valuable results related to the evolutionary development of the middle temporal gyrus.

CCS Concepts

• ***Human-centered computing*** → ***Visualization systems and tools;***

1. Introduction

Primates have evolved unique features that have led to the emergence of several species across different groups. Among the great apes, humans possess distinctive genetic and intellectual traits that differentiate them from other great apes. Recent research [ASL*19; BJH*21; JSE*22] has compared human cortical features with those of other great apes to gain insight into the complex molecular, cellular, and circuit substrates that underlie their cognitive abili-

DOI: 10.2312/vcbm.20231219

ties. One method for exploring these features is single-cell transcriptomics, which generates large and complex datasets studied extensively by medical experts. Single-cell transcriptomics offers insights into gene functionality and the synthesis of proteins in individual cells, where the proteins generated by the multifunctional RNA molecules are essential for carrying out various cellular processes [DC13]. Comparative single-cell transcriptomics is a method to examine differences, e.g., between species, by comparing the RNA transcripts of different samples [Sha19; AKJA21]. Biologists analyze such single-cell transcriptomics data to figure out how individuals are related to each other and identify trends in how transcriptomes have evolved.

Interactive visualization tools allow researchers to identify patterns and relationships within large and complex transcriptomics data that may not be apparent through traditional statistical analysis alone. They also enable effective hypothesis generation and communication of findings to others through visual representations of data, such as heatmaps, scatterplots, and network diagrams, which are commonly used in biology [OGG*10].

In this paper, we present Cytosplore Simian Viewer: a visual analytics application designed to provide a comprehensive and intuitive way for explorative visualization of single-cell transcriptomic data with a specific focus on comparisons of cell-type differences and similarities between multiple species. The system combines different linked views and interactive tools that allow researchers to compare transcriptomic data from other species and identify patterns and relationships that may be difficult to discern using traditional scripted analysis methods. We visualized independently established neurobiology findings to validate the system using real-world datasets. Cytosplore Simian Viewer was developed to accompany the study conducted by Jorstad et al. on transcriptomic analysis of the human middle temporal gyrus, a specialized brain region, in comparison to other great apes [JSE*22]. Our results demonstrate the utility of our tool in generating valuable insights about the middle temporal gyrus.

2. Biological Background

The middle temporal gyrus (MTG) is a highly specialized region of the brain involved in higher cognitive functions [BTD*21], including language processing and visual recognition [DTH*16]. It is also linked to neurological disorders [CCL*22], making it an active field of study in neuroscience. Recent large-scale research efforts used single-cell RNA sequencing data to identify the cell types present in the MTG and the genes that regulate these cell types to understand the molecular mechanisms underlying MTG function [HBM*19; LFG21]. Another important aspect of studying the MTG is gaining an understanding of the spatial organization of cell types in different layers of the cortex [QLL*23]. The primate cortex has distinct layers, each with a mix of cell types, neuronal connections, and functions. This layered architecture is instrumental in processing information and performing complex cognitive tasks. In addition to understanding cell function and organization in individual species, scientists are interested in differences across various species. Here, both the preservation of cell types between species and their spatial organization in the layer structure of the brain are of interest. Such insights will improve the understanding of how information is processed and transmitted throughout the brain circuitry and ultimately enable insight into the shared biological mechanisms that underlie certain traits and behaviors. Apart from cell type characterization, the most recent studies [LNH*22; XMM*23; KN23] link human-specific cell types to genomic regions that are conserved during evolution, but also to regions that underwent accelerated evolutionary changes in humans. Jorstad et al. [JSE*22] compare the distinctive characteristics of the human brain's MTG with four other non-human primate species at a cellular and molecular level to understand specific changes in gene activity within certain types of brain cells during human brain evolution.

Human Accelerated Regions (*HAR*s) are regions of the genome that are conserved in vertebrates but have undergone accelerated changes in humans. They are thought to play a role in the development and function of the human brain [PSK*06]. Several genes have been identified that are expressed in the cell types specific to the MTG, and that can be associated with these HAR- regions, suggesting that they may play a role in human traits [KBW*09; JSE*22].

Human Ancestor Quickly Evolved Regions (*HAQER*s) are also fast-evolving genomic regions. HAQERs are heavily mutated in humans compared to the great apes and have undergone more genetic changes in a given time frame than other genomic regions [KN23]. Both HAQERs and HARs show enrichment for the brain and gastrointestinal tract [MAM*22].

Human-specific conserved deletions (*hCONDEL*s) are specific regions within the human genome that are conserved in vertebrates but have undergone human-specific deletions for certain genomic regions. They possess the potential to exert a profound influence on unique characteristics that are exclusive to the human species. hCONDELs exhibit abundant enrichment in transcriptomic datasets, pointing towards their notable correlation with human brain functions [XMM*23]. Altogether, these insights can ultimately be used to develop more effective treatments for neurological disorders and improve our overall understanding of the human brain [HBM*19].

3. Related Work

With the ongoing influx of computational methods into life sciences, visualization, and visual analysis have become important tools for researchers. O'Donoghue et al. [OBC*18] surveyed the use of visualization in emerging biological research areas. Within this space, so-called *omics*-data, an umbrella term for genomics, transcriptomics, proteomics, and others, has attracted significant attention. In an early overview of the field, Nielsen et al. [NCD*10] surveyed visualization tools for genome data, and Gehlenborg et al. [GOB*10] present visualization methods for general omics data in the scope of systems biology. Here, we review some of the existing omics-related visual analytic works.

Genomic data visualization typically deals with the visualization of sequence data. Several tools and techniques exist to visualize genomic data as surveyed by Nusrat et al. [NHG19]. The genome contains the entire genetic information of an organism in a long sequence of base pairs. As such genomic visualization typically deals with the visualization of these sequences which poses significant scalability issues, e.g., the human genome con-

ID	Description	Species	Type	Items	Attributes
DG1	pairwise cluster distances	all	graph	nodes/links × species pair	nodes: in-species cluster (id) links: distance (quantitative)
DT1	cell information	all	tabular	cells × species	gene expressions (quantitative) cluster assignments (categorical)
DT2	cluster hierarchy	all	tabular	in-species clusters	cluster assignments (categorical) e.g., cross-species cluster, neighborhood, etc.
DT3	gene attributes	human	tabular	genes × cross-species cluster	HARs/hCONDELs/HAQERs (binary)
DM1	various metadata	mixed	tabular	cells × species	layer assignment, color (categorical) UMAP coordinates (quantitative)

Table 1: ***Data Overview*** *including cluster distance graph and several cell and gene information tables detailed in Section 4.1. Dataset identifiers: DGx - Graph, DTx - Table, DMx - Metadata.*

sists of approximately three billion base pairs. A common way to deal with these issues is by aggregating repeating or similar sequence snippets [NJBJ09; ORRL10]. Given a single genome per individual, genomic visualization often deals with comparative tasks, e.g., comparing genome sequences to identify mutations, or differences between species, e.g., by aligning blocks with the sequences [ARJL14; HJW*16]. MizBee [MMP09] compares genomic data across multiple species by visualizing synteny data relationships at the genome, chromosome, and block levels and Pathline [MWS*10] facilitates the visualization of pathway relationships for genomic datasets.

Quantitative analysis of gene expression, i.e., which and how many of the genes of the genome are active in a specific tissue, and protein expression, i.e., which and how many proteins are formed, are important areas of life science research. Visual analysis of such data has been an active field of research in recent years. MulteeSum [MMDP10] facilitates the visual analysis of gene expression of cells in Drosophila embryos. It provides temporal gene expression profiles for individual cells, accompanied by their respective spatial positions. Cytosplore [HPvU*16] allows the analysis of single-cell protein expression data. It uses a combination of dimensionality reduction and clustering approaches to identify functionally similar groups of cells and label them. Single Cell Explorer [FWS*19], iS-CellR [Pat18], and ASAP [GDS*17] implement similar workflows, combining dimensionality reduction with secondary detail views, but for single-cell gene expression data. Brainscope [HvM*17] introduces a novel dual dimensionality reduction. It links an embedding of the original data, i.e., samples according to their expressed gene, with an embedding of the transposed data matrix, i.e., genes according to their expression in different samples. ImaCytE [SVK*19], Facetto [KBJ*20], or Vitessce [KGM*21] are visual analytics systems extending the previously described concepts for visualizing abstract single-cell data to spatially resolved imaging data. Somarakis et al. later also extended ImaCytE with functionality for cohort comparing [SIL*21].

While comparative analysis is common in genome visualization, it has not been widely employed in the quantitative omics space. Cytosplore Simian Viewer targets this area, with a focus on single-cell transcriptomics data and the similarities and differences in cellular and genetic composition between multiple species.

4. Domain Abstraction

This work is part of a longstanding collaboration between visualization researchers and neurobiologists within the Cytosplore Viewer project. During development, we embrace a participatory design approach [JKKS20] with the domain experts. Over the course of a year, we held regular Zoom meetings as well as an on-site workshop. In the first phase, which was conducted completely remotely, we identified the design goals and requirements for organizing and simplifying the datasets to make them more manageable for interactive visualization. In the following, we iteratively discussed prototypes until the final design, which was presented to and discussed with a wider audience of experts at an on-site workshop. Throughout, prototypes were deployed with and tested by our partners and the final version will be made available publicly with the publication of this study.

4.1. Data

A comparative study of various primate species was conducted by our collaborators, as reported by Jorstad et al. [JSE*22]. Single-cell RNA sequencing data were generated from five species; humans, chimpanzees, gorillas, macaques, and marmosets. All cells in the data were classified into three major classes: excitatory neurons, inhibitory neurons, and non-neuronal cells. They were subsequently further categorized into subclasses based on marker gene expression. A total of 24 cellular subclasses were defined (18 neuronal, 6 non-neuronal). To establish a consensus cell type taxonomy across species, cross-species clusters are defined, which represent groups of clusters demonstrating similarity between species. As a result, 151 in-species and 86 cross-species cluster categories were defined for the different cells, providing a comprehensive classification of cell types across species [JSE*22, Figure 1D]. The complete data comprises a variety of parts, listed in Table 1. Below, we will describe the individual components of the overall data in more detail.

DG1: Cluster Distances between the in-species clusters across all five species (pairwise) are stored as a graph for each pair. Nodes in the graphs represent in-species clusters of the compared species and the links represent the distances between nodes. A detailed description of how these inter-species distances are derived can be found in the original study [JSE*22, Taxonomy comparisons].

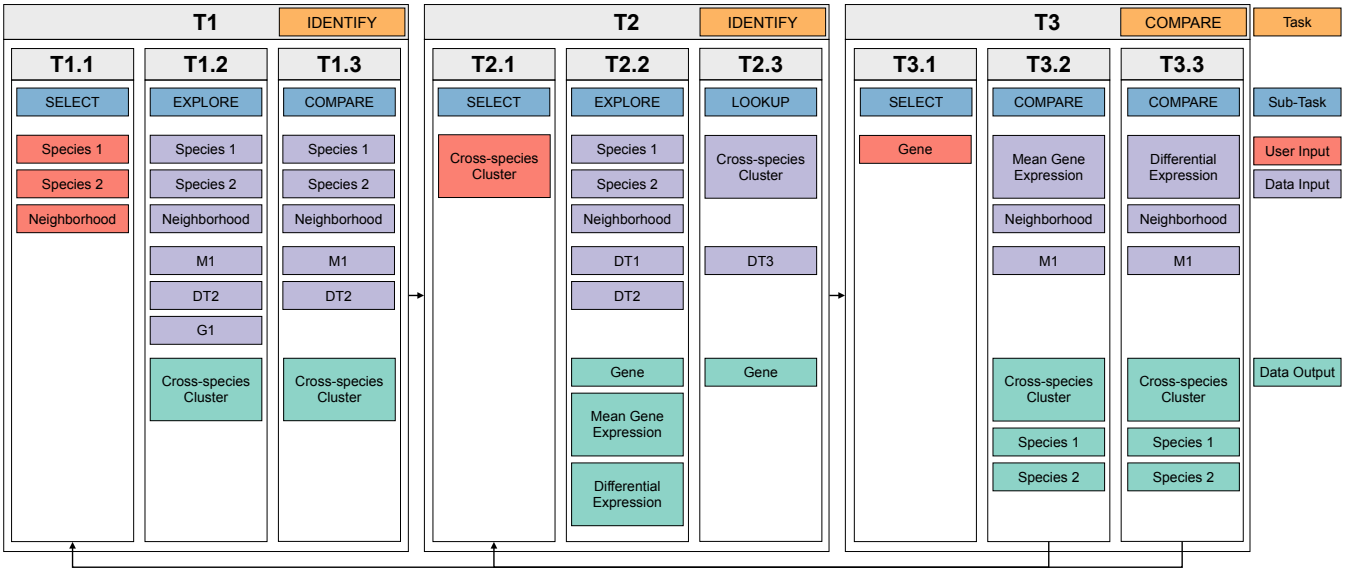

Figure 2: *Overview of the **Task Abstraction** showing three main tasks: **T1 identify** cross-species clusters of interest, **T2 identify** genes of interest, and **T3 compare** expression values across species.*

DT1: Gene Expression values are stored as tabular data. The table provides information about the expression levels of genes within each cell. Each row in the data table corresponds to a cell, while the columns represent different genes. Depending on the species, the data contains approximately 3000 to 9000 genes for neuronal cells and 1500 to 3000 genes for non-neuronal cells. Additional columns contain information on the different cluster assignments of the cells following the classification by Jorstad et al. [JSE*22, Within-species cell type taxonomies].

DT2: Cluster Mapping consists of information presented in a tabular format linking in-species clusters to various categories of the cells such as cross-species clusters, subclasses, neighborhoods, and classes. This mapping allows access to the higher-level category information of cell groups corresponding to the in-species clusters.

DT3: Gene Attributes provide information into genes linked to HARs, hCONDELs, and HAQERs. This data is stored as a table, associating the evolving genomic region values with specific genes found in cross-species clusters for humans. Additional information regarding this can be found in the publication by our collaborators [JSE*22, Enrichment of HARs and hCONDELs].

DM1: Various Metadata. Various tabular datasets consist of information on the cell distribution across different layers per species. Additionally, it consists of information on the total number of cells considered for this distribution analysis, pre-calculated dimensionality reduction of **DT1** to two dimensions, using Uniform Manifold Approximation and Projection [MHM18], and predefined colors for each cluster on all levels of the cluster hierarchy, consistent with domain conventions.

In our discussions with the biologists, we realized that they were mainly focused on the comparison at a cluster level instead of a cellular level for the different primate species. This means that we can aggregate **DT1** to mean gene expression per cluster, instead of keeping the complete cell $\times$ gene matrix which can be prohibitively large (e.g., $\sim$ 140.000 $\times$ 12.000 values for human, resulting in over 6GB at 32bit precision). While we keep the meta information (**DM1**) per cell (e.g., for the UMAP visualization), this reduction allowed us to design the system around clusters, e.g., to support on-the-fly computation of differential gene expression between clusters.

4.2. Analysis Goals

Based on our discussions we have formulated the following high-level analysis questions together with our collaborators:

Q1 How do cell type characteristics, such as layer distribution or gene expression, differ across species?

Q2 How are cell-type specific genes in human cross-species clusters associated with HARs, hCONDELs, and HAQERs, and how many genes are involved?

Q3 Are there genes that have comparable expression patterns across multiple species?

Q4 Are there genes that have highly variable expression patterns across multiple species?

While these questions are general when it comes to comparison between two or more species, a main driver for the analysis of our collaborators was the identification of human-specific traits. Thus analysis is often started by pairwise comparison between human and one of the other species. Only after that more general comparison, i.e., of gene expression within a cross-species cluster across all species is of interest.

4.3. Tasks

Based on the high-level goals, presented in Section 4.2, we identified the following analysis tasks based on Brehmer and Munzner's model [BM13] that we aim to support with our solution:

T1 identify cross-species clusters of interest by exploring in-species cell cluster attributes (**Q1**)

T2 identify genes of interest (**Q2**) for the identified cross-species clusters according to their attached meta information

T3 compare gene expression values between species (**Q3**, **Q4**) based on cross-species clusters

In the typical workflow, the user starts on the cluster level (**T1**), then moves to the gene level (**T2**, **T3**), and finally iterates between the gene and cluster levels for different species (**T1**, **T2**, **T3**). This iterative process helps determine the underlying similarities and differences between the species.

The individual tasks in this workflow can be further refined to provide a clearer picture of the individual steps (Figure 2). In Task **T1** the user **identifies** a cross-species cluster of interest. The trivial way to do this is to simply **lookup** a cluster the user was already interested in beforehand. In the explorative setting, the user would **identify** the cluster in the following way. First, they would **select** a pair of species to compare and a so-called neighborhood. The neighborhood here stands for a group of similar cell types rather than a spatial region. For this selection, the user **explores** the distances of the cross-species clusters and their contained in-species clusters. Finally, the user **compares** the additional information, like the count, distribution, and layer assignment of the contained cells.

After the cluster identification stage, the user transitions from the cluster level to gene-level exploration. Task **T2**, is to **identify** genes of interest either by **lookup** if a gene of interest is known beforehand, or by **selecting** a cross-species cluster of interest and then **exploring** genes with large differential expression values. Additionally, for human comparisons, the users may also want to **lookup** if the gene is associated with additional metadata such as HARs, hCONDELs, and HAQERs.

Finally, in task **T3**, the user aims to **compare** gene expression values among species by **selecting** a gene of interest and **comparing** the expression values across species to **identify** similarities and differences. Furthermore, they may also **compare** the differential expression values of humans with the other species.

Following the indicated iterative process, the results from the previous steps may lead the user to **identify** and **select** a new species to navigate to its cluster level and **explore** the cluster-level attributes and the gene-level attributes.

5. Cytosplore Simian Viewer

We designed and implemented Cytosplore Simian Viewer in an iterative process together with our domain expert collaborators. The implementation is done as a plugin for Manivault [VKT*24] in C++, Qt, and D3.js [BOH11]. Source code is available on Github. Executables are distributed through the Cytosplore Viewer Project.

Cytosplore Simian Viewer supports the tasks defined above with five linked views shown in Figure 1. The heat map view (Figure 1a) provides a display of the in-species cluster distance values (**DG1**) between two species, supporting the identification of clusters of interest (**T1**). From here, the user can query and compare meta information on cluster-related attributes such as layer distribution and cell count information (**DM1**) (**T1**). The next two views include the scatterplot views (Figure 1b and Figure 1c), presenting the low-dimensional embeddings (**DM1**) of the cells which help to compare the cluster overlap between species (**T1**). The differential expression view (Figure 1d) allows the computation of and displays differential gene expression values for each gene in the cross-species cluster (**DT1**) (**T2**). It also displays the HAR, hCONDEL, and HAQER (**DT3**) information (**T2**, **T3**). The expression comparison view (Figure 1e) presents data related to gene expression (**DT1**) and differential expression per neighborhood for a gene selection to help compare expression values between all species in the selected neighborhood (**T3**). All views are linked to support smooth interaction and easy transitioning from one state to another in the visual exploration process.

5.1. Heatmap View

One critical challenge with respect to the visualization design was supporting the cluster distance comparison of multiple species in terms of screen space availability. Given the focus of our users on pairwise comparison, usually with a focus on human vs. one of the other species, we decided to show the cluster distance graph (**DG1**) for a selected pair of species in a heatmap representation (Figure 3). The heatmap supports reading the distance values provided as attributes on the edges well, given the large enough size of individual cells. Additional information such as the cluster mapping (**DT2**), and cell layer distribution (**DM1**) to support task **T1** are shown in a detail view on hovering over the corresponding cell (Figure 4). To make sure cells are large enough for readability, but also to function well as a click target we split the complete graph into the individual cell type neighborhoods, resulting in heatmaps with tens of items on each axis compared to over a hundred for the complete graph. The neighborhoods are also usually inspected separately, meaning no information is lost in the process. The organization of cell types within these neighborhoods is based on the cluster hierarchy (**DT2**), provided by our collaborators. The same holds for the order within the neighborhoods, which we can use directly to order the cell types on the heatmap axes.

The cell type hierarchy also contains the organization of in-

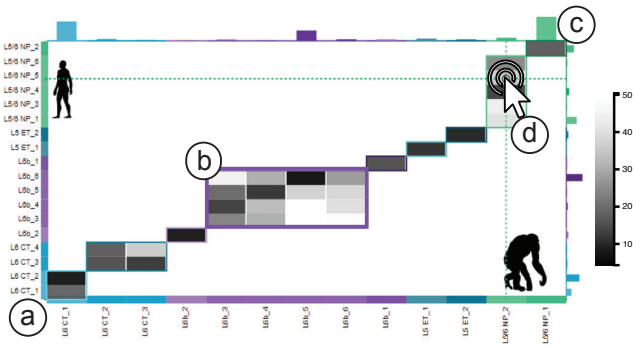

Figure 3: ***Heatmap view*** *with the in-species cluster of the two species on the two axes (a), cross-species clusters indicated by larger block (b), cell counts per in-species cluster (c), hovering (d) reveals detail information in the detail view (Figure 4).*

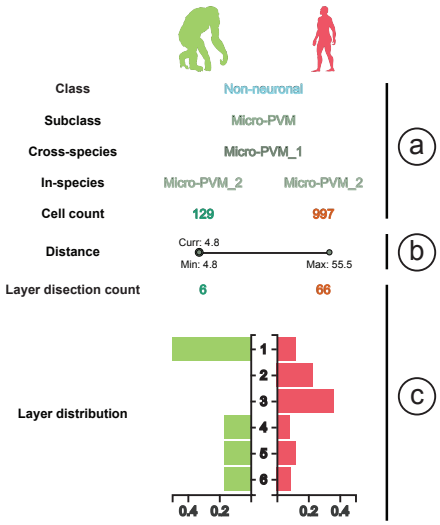

Figure 4: ***Heatmap detail view*** *consisting of three main sections, cluster affiliation (a), cluster distance (b), and layer information (c).*

species clusters (Figure 3a) into cross-species cluster groups. We highlight these in the heatmap as large blocks (Figure 3b) with a colored outline, using their respective cross-species cluster colors. The axis labels are also marked with the provided colors according to domain conventions (**DM1**) and should not be changed. To ensure minimal interference caused by the predefined cluster colors, we opted for a monochrome grayscale color map to depict the distance values between cluster centroids. Our collaborators are interested in identifying low-distance values in the heatmap. Therefore, we map the minimal distance to black, and maximum to white, such that low values are highlighted on the white background.

Besides the similarity between species, clusters of interest are also identified based on other attached data (**T1**) in the tackled workflow. Therefore, we augment the heatmap with two integrated views. The first is a bar chart on each axis representing the size (i.e., the number of cells within) of each cluster (Figure 3c). Hovering over or clicking a cell in the heatmap (Figure 3d) shows or sticks an extended detail view (Figure 4) to the side of the heatmap. The tooltip is split into three main sections, the topmost rows (Figure 4a) show information on the cluster hierarchy, i.e., the class, subclass, and cross-species label for the selected cluster combination, as well as the cell count. Below that we show the cluster distance between the two species as a dot plot including the minimum and maximum distances within the corresponding cross-species cluster and the two selected species (Figure 4b). In case the cross-species cluster only contains a single in-species cluster, we show the distance value as a number without the dot plot. Finally, the bottom-most block (Figure 4c) shows information on the layer distribution of the cells in the selected cluster. The layer dissection count is shown by the corresponding numbers. The distribution of cells across the six layers is shown as two juxtaposed bar charts to allow easy comparison of the differences between the two species.

5.2. Scatterplot View

To further support task **T1**, we provide two scatterplot views showing the UMAP embeddings (**DM1**) of the selected cell neighborhood of the two selected species (Figure 1b, Figure 1c). Each point in the scatterplot embedding represents a cell and cells are laid out according to their similarity and colored using the provided cross-species cluster colors. The UMAP embeddings are further calculated using a cross-species layout, meaning clusters are positioned at the same location for the different species. In summary, this allows to identify cell similarities within species, and also to find cross-species differences, such as the varying size of clusters, or differences in structure, such as a cluster being split in two for one species, but only one for another species.

5.3. Differential Gene Expression View

The Differential Gene Expression View (Figure 1d) provides insight into gene expression for a selected pair of clusters (**DT1**) and differences between gene expression between species, supporting tasks **T2** and **T3**. The view is a standard table view, listing the mean expression for the two selected species and the differential expression for all genes. Depending on the current task, the user can sort by any of the columns. Sorting by mean expression targets **T2**, while **T3** is supported by sorting by differential expression. The differential expression values are here calculated on the fly for the selected pair of clusters. Further, the table view includes columns for HARs, hCONDELs, and HAQERs (**DT3**) when one of the selected species is human (**T2**).

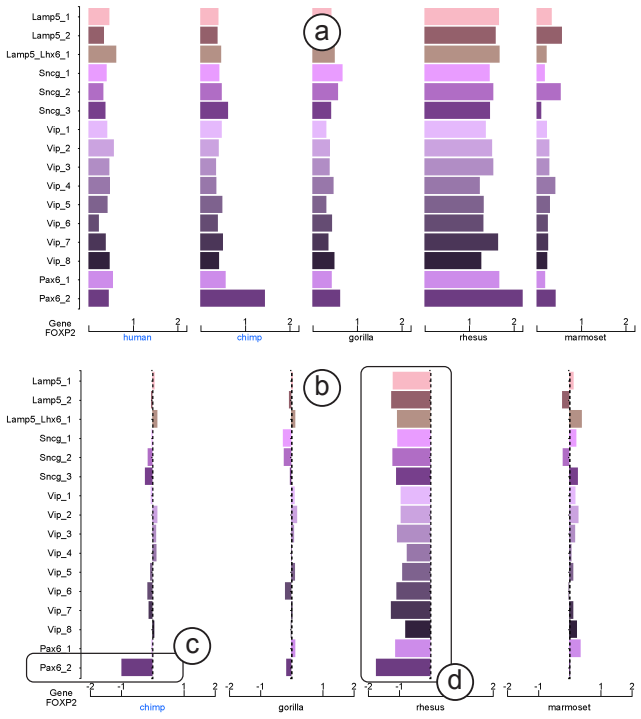

Figure 5: *Analysis of the FOXP2 gene using the different* ***gene expression comparison view*** *modes: (a) Multi-species mean and (b) Differential expression with human*

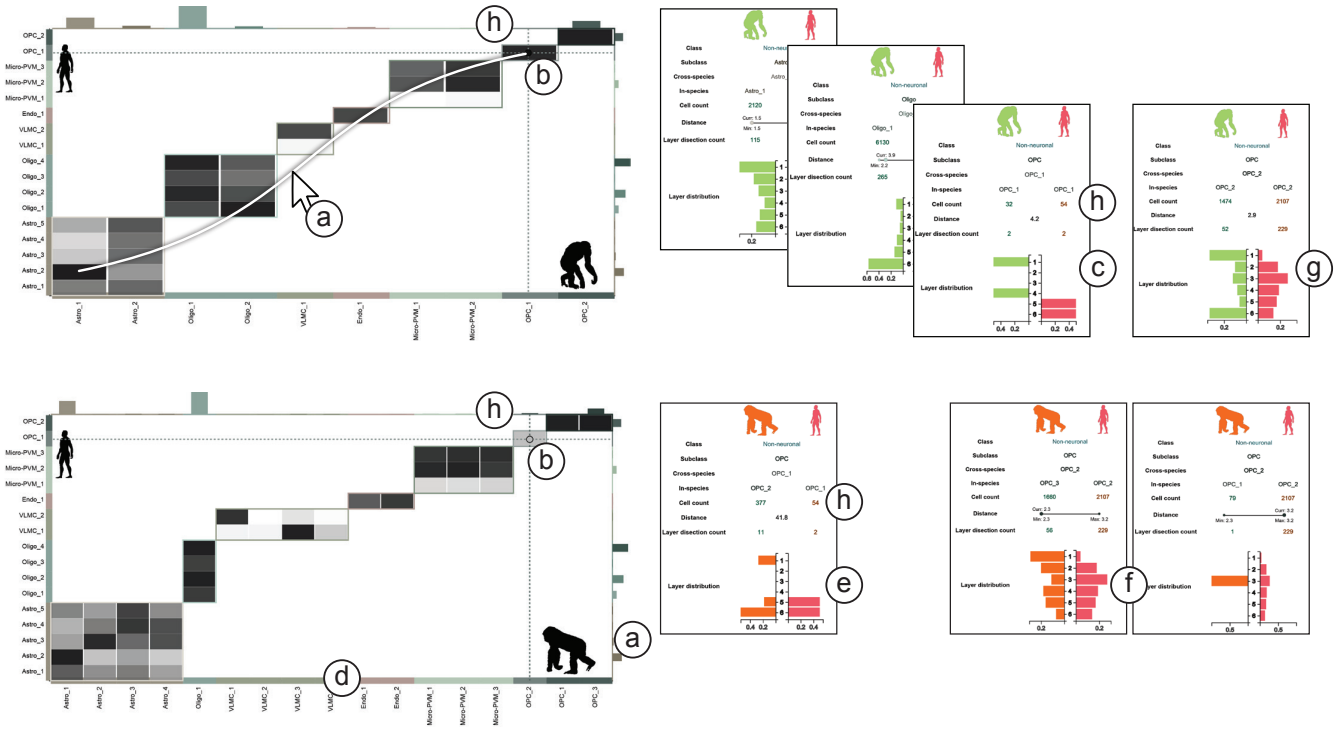

Figure 6: ***Analysis of laminar cell distribution:*** *hovering over different cells in the heatmap (a) reveals the detail view. The OPC_1 cluster (b) shows strong differences in the layer distribution for human vs. chimpanzee (c). Switching the heatmap to human vs. gorilla (d) and selecting the same cluster reveals similar differences in human compared to gorilla (e). Comparing the OPC_2 cross-species clusters reveals a much more even distribution in human, chimpanzee, and gorilla (f, g). The small number of cells (h) indicates results need to be taken with care.*

5.4. Gene Expression Comparison View

Similar to the Differential Gene Expression View, the Gene Expression Comparison View (Figure 5) supports tasks **T2** and **T3**. Here the focus is shifted from the cluster of interest to a gene of interest. Instead of showing all genes for a cluster, we show the gene expression of all clusters (of a selected neighborhood). In the default setting (Figure 5a), the view displays the mean expression data (**DT1**) for a selected gene of interest as multiple bar charts. The items in the bar charts are the different cell clusters in the selected neighborhood. Going beyond the pairwise comparison of the previous views. Here, we show bar charts for each of the five species juxtaposed. This allows a quick overview of all species and provides pointers for further investigation of other species, i.e., the iterative back and forth between tasks, laid out in Section 4.3.

Given the focus on comparison to human, we also provide a setting with an explicit comparison of any species to human. Here, we calculate the difference between human and the other four species and show four bar charts, one for each human-other species pair. Just as before, we then show these difference values directly in a set of bar charts as shown in Figure 5b. The explicit encoding now highlights outlying clusters, e.g., in Figure 5c where the *FOXP2* gene is similar between human and chimpanzee in most clusters, except for *Pax6_2*, or even completely different species such as in Figure 5d, showing the rhesus having much higher expression across all clusters for the same gene.

6. Use Cases

We developed Cytosplore Simian Viewer to accompany the study by Jorstad et al. on the transcriptomic analysis of the human MTG compared to other big apes [JSE*22]. To demonstrate the effectiveness of our system we present two real-world use cases, recreating findings described in that study.

6.1. Analysis of Laminar Cell Distribution

Jorstad et al. investigated the layer structure of the MTG and found that *"laminar distributions of types were remarkably conserved across the great apes"* [JSE*22, Consensus cell type conservation and divergence] with the exception of two cell clusters, *Sst Chodl_1*, a subset of GABAergic neurons, and *OPC_1* a set of oligodendrocyte precursor cells. The latter *"was present in layer 1 of chimpanzee and gorilla but not human MTG"*.

Finding differences in the layer distribution between different species is part of task **T1** in our system. Initially, an analyst focuses on pairwise comparisons, usually involving human. We start by selecting two species, here human and chimpanzee, along with a specific cell neighborhood (non-neuronal cells). The system then displays corresponding comparisons. We mainly utilize the heatmap with the attached detail view for this use case and exclude the gene expression. Notable clusters like *OPC_1* and *OPC_2*, which share a one-to-one correspondence between human and chimpanzee, show extremely low cross-species distance. We hover over dark cells to

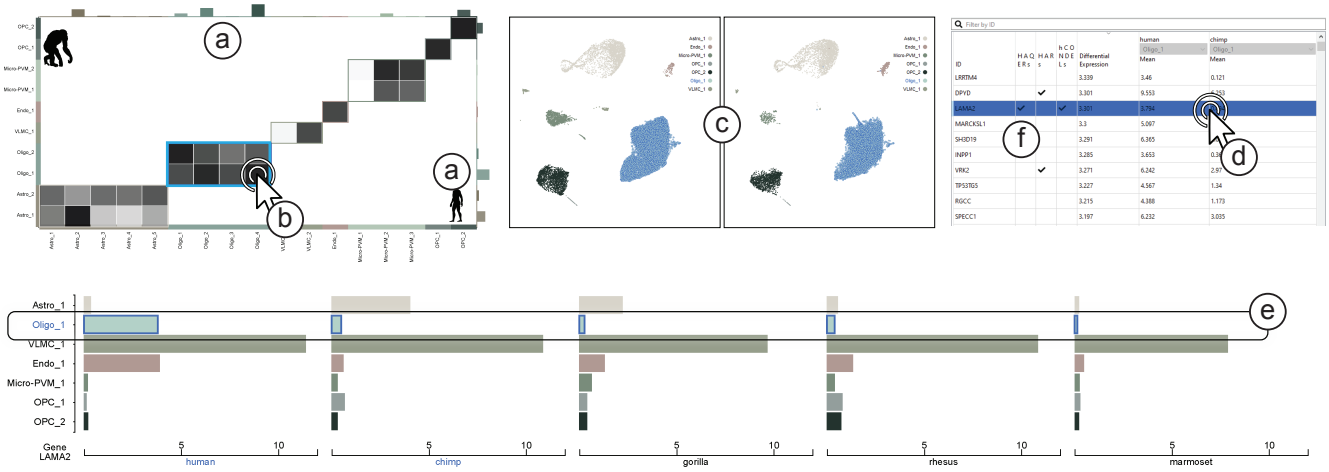

Figure 7: ***Analysis of gene expression patterns****: The cell count chart indicates the large abundance of oligodendrocytes in the non-neuronal cell neighborhood comparing human and chimpanzee (a). Selecting the cross-species cluster (b) highlights the data in the UMAP scatterplots (c) and calculates the differential expression. Selecting genes in the differential expression table (d) sets the gene for comparison in the gene comparison view, indicating higher expression in human compared to any other compared species (e). Human accelerated regions are also indicated in the differential expression table (f).*

explore these clusters using the detail view (Figure 6a). The linked views allow quick glances at several clusters in succession. Once we identify a cluster of interest we pin the detail view by clicking the hovered cell. In the example, when exploring the *OPC_1* cross-species cluster (Figure 6b), a strong difference in the layer distribution between human and chimpanzee becomes evident in the detail view (Figure 6c).

Extending the analysis by replacing chimpanzee with gorilla in the main interface, the system adjusts accordingly, showing cross-species cell cluster distances between human and gorilla (Figure 6d). It is necessary to re-select the *OPC_1* cluster in the heatmap, as the corresponding in-species cluster in gorilla changed compared to chimpanzee. Here, cross-species *OPC_1* corresponds to *OPC_1* in human but *OPC_2* in gorilla. Clicking on the *OPC_1* cluster pins the corresponding detail view. The pinned detail view now compares human vs. gorilla, revealing that both share cells in layers 5 and 6, along with layer 1, which was also observed with chimpanzee (Figure 6e).

To validate the results, we examine the related *OPC_2* cross-species cluster. In gorilla, this splits into two in-species clusters, *OPC_1* and *OPC_3*. Both show a wide distribution across all layers for human and *OPC_3* in gorilla. However, *OPC_1* in gorilla only appears in a single layer (Figure 6f). Going back to chimpanzee, a similarly wide distribution is observed for *OPC_2*, confirming the different nature of *OPC_1* (Figure 6g). Finally, considering the significance of this finding based on the cell counts, the bar chart and the detail view indicate a small number of cells (Figure 6h), indicating the need for further verification, as also stated by Jorstad et al.; *"although more sampling of these rare types is needed for validation"* [JSE*22, Consensus cell type conservation and divergence].

6.2. Analysis of Gene Expression Patterns

In this second use case, we want to explore genes and compare them between species in order to identify genes that are related to accelerated cell changes in humans. As an example, we follow Jorstad et al., who found that the expression of *LAMA2* in human oligodendrocytes differs from its expression in chimpanzee oligodendrocytes indicating accelerated changes of glial cell gene expression in humans [JSE*22, Human specializations of glial cells]. Our framework supports the exploration of gene expression patterns, such as the identification of genes with strong differential expression between humans and other species, primarily through tasks **T2** and **T3**.

We begin by comparing human with chimpanzee in the non-neuronal cells neighborhood. The *Oligo_1* cross-species cluster contains the most cells in this neighborhood (Figure 7a). Selecting the cross-species cluster (blue box, Figure 7b) highlights it in the other views, and differential expressions for all genes between human and chimpanzee is automatically calculated on the fly. The cross-species cluster is split into four in-species clusters in human and two in chimpanzee. However, all cells in the heatmap in the cross-species block (blue box, Figure 7b) are dark, indicating low pairwise distances between any combination of in-species clusters. We, therefore, expect rather strong similarities in gene expression across the in-species clusters. Inspecting the single-cell embeddings in the scatterplots (Figure 7c) reveals rather compact clusters (blue highlights) that are of comparable shape and structure between human and chimpanzee confirming the impression gained from the heatmap.

Next, we further analyze the gene expression patterns for the selected cluster. Upon selecting the cluster, differential gene expression is calculated and displayed in the differential expression view (Figure 7d). This view allows sorting genes by mean or signed

differential expression. While sorting with absolute differential expression values would be possible, we use the sign to indicate species with higher expression. We sort from large to small differential expression, such that genes with a significantly higher expression in human than chimpanzee will show up on the top of the list. By selecting genes from the top and checking the gene comparison view, we identify the *LAMA2* gene as strongly differentially expressed in oligodendrocytes between humans and chimpanzees. The gene expression view not only displays expression in other clusters but also across all five species, helping us assess if this finding extends to other primates. Notably, *LAMA2* is seen to have high expression only in humans for the *Oligo_1* cluster (Figure 7e). Finally, we also note that *LAMA2* is associated with two types of genomic regions that underwent human-specific changes (hCONDEL and HAQER) (Figure 7f).

The highlighted findings demonstrate how species and cell-type-specific gene expression can be linked to species-specific changes in the genome. Such insights contribute to a better understanding of evolutionary biology and can be visually explored within our system. They emphasize the significance of cross-species comparative studies in unraveling the complex nature of biological systems and advancing our knowledge in the field.

7. Discussion and Conclusion

In this paper, we introduced Cytosplore Simian Viewer, an interactive system designed for the exploration and comparison of multi-species transcriptomic datasets. Cytosplore Simian Viewer is used to communicate and recreate findings from a large biological study and allows free exploration of the corresponding data to support the discovery of new findings. The system empowers researchers to gain insights into gene expression patterns, identify differentially expressed genes, and uncover potential biological relationships and pathways for cross-species comparative exploration.

The presented use cases are a first step to show that our system effectively facilitates the exploration and comparison of the given multi-species transcriptomic datasets. A comprehensive user study, with participants and data from different institutions, would be a logical next step to strengthen these findings. While the system was designed and implemented based on a provided dataset, new data that adhere to the same format can be added. For both, species and clusters, we used color hue as a visual encoding. We acknowledge that some of the chosen colors pose issues (e.g., readability for people with color vision deficiencies), they were chosen following established domain conventions to minimize adjustment for the users. Further, we encode all variables encoded with these colors with a second channel, i.e., position, in most views. In the future, we want to expand the tool's capabilities beyond pairwise analyses as well as scaling to a larger number of species.

Acknowledgements

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

[AKJA21] ADIL, ASIF, KUMAR, VIJAY, JAN, ARIF TASLEEM, and ASGER, MOHAMMED. "Single-cell transcriptomics: current methods and challenges in data acquisition and analysis". *Frontiers in Neuroscience* 15 (2021), 591122. DOI: 10.3389/fnins.2021.591122 2.

[ARJL14] AURISANO, JILLIAN, REDA, KHAIRI, JOHNSON, ANDREW, and LEIGH, JASON. "Bacterial gene neighborhood investigation environment: A large-scale genome visualization for big displays". *2014 IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV)*. 2014, 103–104. DOI: 10.1109/LDAV.2014.7013210 3.

[ASL*19] ARDESCH, DIRK JAN, SCHOLTENS, LIANNE H, LI, LONGCHUAN, et al. "Evolutionary expansion of connectivity between multimodal association areas in the human brain compared with chimpanzees". *Proceedings of the National Academy of Sciences* 116.14 (2019), 7101–7106. DOI: 10.1073/pnas.1818512116 1.

[BJH*21] BAKKEN, TRYGVE E, JORSTAD, NIKOLAS L, HU, QIWEN, et al. "Comparative cellular analysis of motor cortex in human, marmoset and mouse". *Nature* 598.7879 (2021), 111–119. DOI: 10.1038/s41586-021-03465-8 1.

[BM13] BREHMER, MATTHEW and MUNZNER, TAMARA. "A multi-level typology of abstract visualization tasks". *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), 2376–2385. DOI: 10.1109/TVCG.2013.124 4.

[BOH11] BOSTOCK, MICHAEL, OGIEVETSKY, VADIM, and HEER, JEFFREY. "D^3 data-driven documents". *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), 2301–2309. DOI: 10.1109/TVCG.2011.185 5.

[BTD*21] BRIGGS, ROBERT G, TANGLAY, ONUR, DADARIO, NICHOLAS B, et al. "The unique fiber anatomy of middle temporal gyrus default mode connectivity". *Operative Neurosurgery* 21.1 (2021), E8. DOI: 10.1093/ons/opab109 2.

[CCL*22] CHEN, SHUO, CHANG, YUZHOU, LI, LIANGPING, et al. "Spatially resolved transcriptomics reveals genes associated with the vulnerability of middle temporal gyrus in Alzheimer's disease". *Acta Neuropathologica Communications* 10.1 (2022), 1–24. DOI: 10.1186/s40478-022-01494-6 2.

[DC13] DONG, ZHICHENG and CHEN, YAN. "Transcriptomics: advances and approaches". *Science China Life Sciences* 56 (2013), 960–967. DOI: 10.1007/s11427-013-4557-2 2.

[DTH*16] DAVEY, JAMES, THOMPSON, HANNAH E, HALLAM, GLYN, et al. "Exploring the role of the posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes". *NeuroImage* 137 (2016), 165–177. DOI: 10.1016/j.neuroimage.2016.05.051 2.

[FWS*19] FENG, DI, WHITEHURST, CHARLES E., SHAN, DECHAO, et al. "Single Cell Explorer, collaboration-driven tools to leverage large-scale single cell RNA-seq data". *BMC Genomics* 20.1 (2019). DOI: 10.1186/s12864-019-6053-y 3.

[GDS*17] GARDEUX, VINCENT, DAVID, FABRICE P A, SHAJKOFCI, ADRIAN, et al. "ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data". *Bioinformatics* 33.19 (2017), 3123–3125. DOI: 10.1093/bioinformatics/btx337 3.

[GOB*10] GEHLENBORG, NILS, O'DONOGHUE, SEÁN I, BALIGA, NITIN S, et al. "Visualization of omics data for systems biology". *Nature Methods* 7.Suppl 3 (2010), S56–S68. DOI: 10.1038/nmeth.1436 2.

[HBM*19] HODGE, REBECCA D, BAKKEN, TRYGVE E, MILLER, JEREMY A, et al. "Conserved cell types with divergent features in human versus mouse cortex". *Nature* 573.7772 (2019), 61–68. DOI: 10.1038/s41586-019-1506-7 2.

[HJW*16] HESS, MARTIN, JENTE, DANIEL, WIEMEYER, JOSEF, et al. "Visual analysis and comparison of multiple sequence alignments". *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine*. 2016, 31–40. DOI: 10.2312/vcbm.20161268 3.

[HPvU*16] HÖLLT, THOMAS, PEZZOTTI, NICOLA, van UNEN, VINCENT, et al. "Cytosplore: interactive immune cell phenotyping for large single-cell datasets". *Computer Graphics Forum*. Vol. 35. 3. 2016, 171–180. DOI: 10.1111/cgf.12893 3.

[HvM*17] HUISMAN, SJOERD M.H., van LEW, BALDUR, MAHFOUZ, AHMED, et al. "BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome". *Nucleic Acids Research* (2017), gkx046. DOI: 10.1093/nar/gkx046 3.

[JKKS20] JÄNICKE, STEFAN, KAUR, PAWANDEEP, KUZMICKI, PAWEL, and SCHMIDT, JOHANNA. "Participatory Visualization Design as an Approach to Minimize the Gap between Research and Application." *VisGap - The Gap between Visualization Research and Visualization Software*. 2020, 35–42. DOI: 10.2312/visgap.20201108 3.

[JSE*22] JORSTAD, NIKOLAS L, SONG, JANET HT, EXPOSITO-ALONSO, DAVID, et al. "Comparative transcriptomics reveals human-specific cortical features". *bioRxiv* (2022). DOI: 10.1101/2022.09.19.508480 1–4, 7, 8.

[KBJ*20] KRUEGER, ROBERT, BEYER, JOHANNA, JANG, WON-DONG, et al. "Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 227–237. DOI: 10.1109/tvcg.2019.2934547 3.

[KBW*09] KONOPKA, GENEVIEVE, BOMAR, JAMEE M, WINDEN, KELLEN, et al. "Human-specific transcriptional regulation of CNS development genes by FOXP2". *Nature* 462.7270 (2009), 213–217. DOI: 10.1038/nature08549 2.

[KGM*21] KELLER, MARK S, GOLD, ILAN, MCCALLUM, CHUCK, et al. "Vitessce: a framework for integrative visualization of multi-modal and spatially-resolved single-cell data". *OSF Preprints* (2021). DOI: 10.31219/osf.io/y8thv 3.

[KN23] KUN, EUCHARIST and NARASIMHAN, VAGHEESH M. "Fast-evolving genomic regions underlie human brain development". *Nature* 613 (2023), 37–38. DOI: 10.1038/d41586-023-00069-2 2.

[LFG21] LU, SHAINA, FÜRTH, DANIEL, and GILLIS, JESSE. "Integrative analysis methods for spatial transcriptomics". *Nature Methods* 18.11 (2021), 1282–1283. DOI: 10.1038/s41592-021-01272-7 2.

[LNH*22] LINKER, SARA B, NARVAIZA, IÑIGO, HSU, JONATHAN Y, et al. "Human-specific regulation of neural maturation identified by cross-primate transcriptomics". *Current Biology* 32.22 (2022), 4797–4807. DOI: 10.1016/j.cub.2022.09.028 2.

[MAM*22] MANGAN, RILEY J, ALSINA, FERNANDO C, MOSTI, FEDERICA, et al. "Adaptive sequence divergence forged new neurodevelopmental enhancers in humans". *Cell* 185.24 (2022), 4587–4603. DOI: 10.1016/j.cell.2022.10.016 2.

[MHM18] MCINNES, LELAND, HEALY, JOHN, and MELVILLE, JAMES. "Umap: Uniform manifold approximation and projection for dimension reduction". *arXiv* (2018). DOI: 10.48550/arXiv.1802.03426 4.

[MMDP10] MEYER, MIRIAH, MUNZNER, TAMARA, DEPACE, ANGELA, and PFISTER, HANSPETER. "MulteeSum: a tool for comparative spatial and temporal gene expression data". *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), 908–917. DOI: 10.1109/TVCG.2010.137 3.

[MMP09] MEYER, MIRIAH, MUNZNER, TAMARA, and PFISTER, HANSPETER. "MizBee: a multiscale synteny browser". *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), 897–904. DOI: 10.1109/TVCG.2009.167 3.

[MWS*10] MEYER, MIRIAH, WONG, BANG, STYCZYNSKI, MARK, et al. "Pathline: A tool for comparative functional genomics". *Computer Graphics Forum*. Vol. 29. 3. 2010, 1043–1052. DOI: 10.1111/j.1467-8659.2009.01710.x 3.

[NCD*10] NIELSEN, CYDNEY B, CANTOR, MICHAEL, DUBCHAK, INNA, et al. "Visualizing genomes: techniques and challenges". *Nature Methods* 7.S3 (2010), S5–S15. DOI: 10.1038/nmeth.1422 2.

[NHG19] NUSRAT, SABRINA, HARBIG, THERESA, and GEHLENBORG, NILS. "Tasks, techniques, and tools for genomic data visualization". *Computer Graphics Forum*. Vol. 38. 3. 2019, 781–805. DOI: 10.1111/cgf.13727 2.

[NJBJ09] NIELSEN, CYDNEY B, JACKMAN, SHAUN D, BIROL, INANÇ, and JONES, STEVEN JM. "ABySS-Explorer: visualizing genome sequence assemblies". *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), 881–888. DOI: 10.1109/TVCG.2009.116 3.

[OBC*18] O'DONOGHUE, SEÁN I, BALDI, BENEDETTA FRIDA, CLARK, SUSAN J, et al. "Visualization of biomedical data". *Annual Review of Biomedical Data Science* 1 (2018), 275–304. DOI: 10.1146/annurev-biodatasci-080917-013424 2.

[OGG*10] O'DONOGHUE, SEÁN I, GAVIN, ANNE-CLAUDE, GEHLENBORG, NILS, et al. "Visualizing biological data—now and in the future". *Nature Methods* 7.Suppl 3 (2010), S2–S4. DOI: 10.1038/nmeth.f.301 2.

[ORRL10] O'BRIEN, TREVOR, RITZ, ANNA, RAPHAEL, BENJAMIN, and LAIDLAW, DAVID. "Gremlin: an interactive visualization model for analyzing genomic rearrangements". *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), 918–926. DOI: 10.1109/TVCG.2010.163 3.

[Pat18] PATEL, MITULKUMAR V. "iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data". *Bioinformatics* 34.24 (2018), 4305–4306. DOI: 10.1093/bioinformatics/bty517 3.

[PSK*06] POLLARD, KATHERINE S, SALAMA, SOFIE R, KING, BRYAN, et al. "Forces shaping the fastest evolving regions in the human genome". *PLOS Genetics* 2.10 (2006), e168. DOI: 10.1371/journal.pgen.0020168 2.

[QLL*23] QIAN, JINGYANG, LIAO, JIE, LIU, ZIQI, et al. "Reconstruction of the cell pseudo-space from single-cell RNA sequencing data with scSpace". *Nature Communications* 14.1 (2023), 2484. DOI: 10.1038/s41467-023-38121-4 2.

[Sha19] SHAFER, MAXWELL ER. "Cross-species analysis of single-cell transcriptomic data". *Frontiers in Cell and Developmental Biology* 7 (2019), 175. DOI: 10.3389/fcell.2019.00175 2.

[SIL*21] SOMARAKIS, ANTONIOS, IJSSELSTEIJN, MARIEKE E., LUK, SIETSE J., et al. "Visual cohort comparison for spatial single-cell omics-data". *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2021), 733–743. DOI: 10.1109/TVCG.2020.3030336 3.

[SVK*19] SOMARAKIS, ANTONIOS, VAN UNEN, VINCENT, KONING, FRITS, et al. "ImaCytE: Visual Exploration of Cellular Micro-Environments for Imaging Mass Cytometry Data". *IEEE Transactions on Visualization and Computer Graphics* 27.1 (2019), 98–110. DOI: 10.1109/TVCG.2019.2931299 3.

[VKT*24] VIETH, ALEXANDER, KROES, THOMAS, THIJSSEN, JULIAN, et al. "ManiVault: A Flexible and Extensible Visual Analytics Framework for High-Dimensional Data". *IEEE Transactions on Visualization and Computer Graphics* 30.2 (2024). DOI: 10.48550/arXiv.2308.01751 5.

[XMM*23] XUE, JAMES R, MACKAY-SMITH, AVA, MOURI, KOUSUKE, et al. "The functional and evolutionary impacts of human-specific deletions in conserved elements". *Science* 380.6643 (2023), eabn2253. DOI: 10.1126/science.abn2253 2.