GeneSpectra: a method for context-aware comparison of cell type gene expression across species

Yuyao Song[*1], Irene Papatheodorou[*1,2,3], Alvis Brazma[1]
1 European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, CB10 1SA, UK
2 Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK
3 Medical School, University of East Anglia, Norwich Research Park, Norwich, NR4 7UA, UK
* Corresponding authors

Abstract

Computational comparison of single cell expression profiles cross-species uncovers functional similarities and differences between cell types. Importantly, it offers the potential to refine evolutionary relationships based on gene expression. Current analysis strategies are limited by the strong hypothesis of ortholog conjecture, which implies that orthologs have similar cell type expression patterns. They also lose expression information from non-orthologs, making them inapplicable in practice for large evolutionary distances. To address these limitations, we devised a novel analytical framework, GeneSpectra, to robustly classify genes by their expression specificity and distribution across cell types. This framework allows for the generalization of the ortholog conjecture by evaluating the degree of ortholog class conservation. We utilise different gene classes to decode species effects on cross-species transcriptomics space and compare sequence conservation with expression specificity similarity across different types of orthologs. We develop contextualised cell type similarity measurements while considering species-unique genes and non-one-to-one orthologs. Finally, we consolidate gene classification results into a knowledge graph, GeneSpectraKG, allowing a hierarchical depiction of cell types and orthologous groups, while continuously integrating new data.

Introduction

Through millions of years, evolution shaped the building blocks of life, resulting in the plethora of species we observe today. Detecting conserved gene expression profiles by cross-species comparison of single-cell RNA-seq (scRNA-seq) data can shed light on the evolutionary links between cell types (Arendt et al., 2016; Marioni & Arendt, 2017). Currently, computational methods for cross-species comparisons of scRNA-seq data often rely on expression correlation analysis or data integration (Shafer, 2019; Song et al., 2023; Tanay & Sebé-Pedrós, 2021). For both methods, to measure expression similarity between cells across species, one first needs to match genes by orthology. To be specific, in correlation analysis, one-to-one (1-2-1) orthologs are often used as the only shared features to calculate the degree of cell type similarity (Geirsdottir et al., 2019; Jorstad et al., 2023; Najle et al., 2023; Sebé-Pedrós et al., 2018; Shafer et al., 2021). In data integration, algorithms require orthology mapped data as input to align transcriptomically similar cell populations (Franjic et al., 2022; Jorstad et al., 2023; Liu et al., 2021; Tritschler et al., 2022). Both strategies assume that cell types from different species are alike if they have similar relative expression levels of orthologs.

This assumption is rooted in the ortholog conjecture hypothesis (Koonin, 2005; Nehrt et al., 2011; Sonnhammer et al., 2014), which is a foundational hypothesis for the function prediction of genes. In short, the hypothesis proposes that orthologs typically perform very similar functions in the respective organisms (Koonin, 2005). As cell type gene expression pattern can be considered as a functional readout of the gene, the hypothesis is implicitly extended to the analysis of scRNA-seq data cross-species. There the ortholog conjecture implies that we can expect orthologs to have similar gene expression patterns across the respective organisms, which would further imply that the expression of orthologs should be correlated across the corresponding cell types between the organisms. It is crucial to point out that the validity of the ortholog conjecture has been intensely debated (Altenhoff et al., 2012; X. Chen & Zhang, 2012; Gabaldón & Koonin, 2013; Kryuchkova-Mostacci & Robinson-Rechavi, 2016; Nehrt et al., 2011; Rogozin et al., 2014; Sonnhammer et al., 2014; Stamboulian et al., 2020). It is generally agreed that having an empirical, quantitative estimate of the functional similarity of orthologs, respective to their evolutionary relationship, is a more careful approach (Sonnhammer et al., 2014). The application of the ortholog conjecture becomes even more problematic for more granularly defined cell types, which often are defined using gene expression – thus potentially leading to circularity.

For comparing evolutionary distant species, the reliance on ortholog conjecture hypothesis becomes somewhat questionable. First, there are very few 1-2-1 orthologs between distant species genomes, and the gene orthology relationship can be obscure (Natsidis et al., 2021). Secondly, the genome context has diverged between the species, and tissues experience concerted expression evolution (Arendt et al., 2016; Liang et al., 2018). The gene function and thus the expression, may not be well preserved even between orthologs. Lastly, new genes that emerged in a species can play a major role in the species' cell type expression landscape and contribute to species-unique cell populations (Florio et al., 2018; Santos et al., 2017). Examining the conservation of ortholog cell type expression patterns, with criteria such as the Human Protein Atlas (HPA) gene classification (Karlsson et al., 2021), can help reassess the degree of validity of the ortholog conjecture in cross-species comparison at different evolutionary distances. The potential circularity of defining cell types by gene expression and then using gene expression to map these cell types between species, can be broken by expanding the gene list beyond one-to-one orthologues.

Recently, several cross-species integration tools had been developed to tackle the restriction by gene orthology. For instance, SATURN (Rosen et al., 2024) and TACTiCS (Biharie et al., 2023) utilise protein sequence to refine gene homology annotations; SAMap (Tarashansky et al., 2021) adopts an iterative process to update both gene-gene mapping and cell-cell mapping; SATL (Park et al., 2024) performs species-agonistic transfer learning using shared labelled cell types without gene matching. Nevertheless, these tools take up other assumptions, such as considering sequence conservation as indicative of gene functional equivalency (SATURN, TACTiCS), or require population balance between species datasets (SAMap), or some cross-species cell type mappings are given (SATL).

It is increasingly acknowledged that cell type is inherently a hierarchical concept: the exact type of cells depends on the annotation granularity (Domcke & Shendure, 2023). This has important implications for the mapping of cell types cross-species via gene expression. At low granularity, the knowledge of key cell type marker genes makes it relatively straightforward to match related cell types between species, such as mapping cardiomyocytes between human, mouse and zebrafish. However, it is substantially more challenging when a high cell type granularity is aimed for, such as different immune cell subtypes or finer groups of neurons.The number of cell types and the cell type granularity that is appropriate can be inherently divergent between species, particularly for evolutionary

distant species, or species with highly specific adaptive features. Practically, it is also important to check whether independently annotated data are truly comparable in terms of cell populations coverage and annotation granularity, based on current available data.

On observing the existing analysis paradigms, it is essential to empirically calibrate the dependence on ortholog conjecture, and include the expression signals from non-1-2-1 orthologs and species-specific genes during comparison. Furthermore, in analysing distant species of highly granular cell types, it is necessary to consider contextual factors that influence the perceived similarity between cell types, such as the cell type hierarchy and gene orthology resolution. These considerations become increasingly pertinent as we enter the era of building cross-species cell atlases at scale ("Evolution at the Cellular Level," 2023).

Here, we propose a more holistic approach for cross-species cell type comparison tailored to the system of interest. Specifically, we devise a novel analytical framework for cross-species comparison of single cell transcriptomics data, focusing on a dataset-specific examination of ortholog conjecture and a contextualised cell type similarity analysis cross-species. We implemented the framework as freely available GeneSpectra/GeneSpectraKG Python modules (https://github.com/Papatheodorou-Group/GeneSpectra, https://github.com/Papatheodorou-Group/GeneSpectraKG). GeneSpectra provides functions to perform gene classification based on cell type expression using modified HPA classes and compare these classifications between species. GeneSpectraKG takes outputs from GeneSpectra as inputs, and leverages knowledge graphs (KGs)(Lobentanzer et al., 2023) to manage the varying resolutions and the complex relationship between genes, ortholog groups, cell types and species for a contextualised similarity measurement of cell types.

We demonstrate the utility of GeneSpectra/GeneSpectraKG framework by applying it on existing datasets with aligned cell types. We classify genes by their expression specificity and distribution across the cell types. By using this classification, we show different conservation patterns of orthologs with different specificity and distribution classes. We also utilise the gene classes to explain species effect (Song et al., 2023) on multi-species scRNA-seq data, and delineate the relationship between sequence conservation and expression specificity similarity in different types of orthologs. We then apply the framework on datasets with independently annotated cell types to show it is informative of dataset comparability. Using information in the KG, we propose and analyse two cell type similarity metrics from different perspectives, taking into account non-1-2-1 orthologs and

species-unique cell type specific genes. These metrics show that different cell types accumulate varying degrees of cross-species divergence of cell type-specific genes, likely due to concerted evolution.

Results

By robustly classifying genes according to their cell type expression specificity and distribution, we observed that orthologs with low specificity or broad expression mostly remained low specificity or broad expression across species. On the other hand, orthologs with some degree of specificity or selectivity can switch class, though this happens less with orthologues of highest degree of expression specificity. Moreover, genes that are specifically expressed in particular cell types in one species, can have orthologs that are specifically expressed in different cell types in another species. The orthologous genes that switched from cell type specific to nonspecific expression across species, contribute to the global expression shift observed on cross-species joint scRNA-seq embedding, when cross species data integration is being attempted. Comparing expression specificity similarity with sequence conservation suggests that closely related species have higher consistency between sequence conservation and expression similarity. Moreover, non-1-2-1 orthologs didn't show any less expression similarity than 1-2-1 orthologs at the same sequence conservation level. Finally, our cross-species KG-derived cell type similarity measurement, which takes into account all genes, delineates divergence in the accumulation of species-unique expression between different cell types. Below, we give details of our results.

Robust gene classification based on cell type expression

We developed the GeneSpectra/GeneSpectraKG framework to provide methods to characterise ortholog expression conservation and perform contextualised analysis of cell type similarity across different species. The GeneSpectra module first reduces the sparsity of scRNA-seq data, then, after depth normalisation, applies gene classification. It assigns different classes to each gene based on their cell type expression specificity and distribution patterns, referencing the HPA gene classes (Karlsson et al., 2021). The GeneSpectraKG module utilises GeneSpectra's gene classes, integrates orthology annotations from the EggNOG database (Hernández-Plaza et al., 2023), and generates a KG to effectively represent the results and enable further analysis. Figure 1 summarises the key steps involved in this pipeline. Taken together, for the first time this framework allows one to

explore cell type correspondence between the species, in a manner that is not entirely based on 1-2-1 gene orthologs.

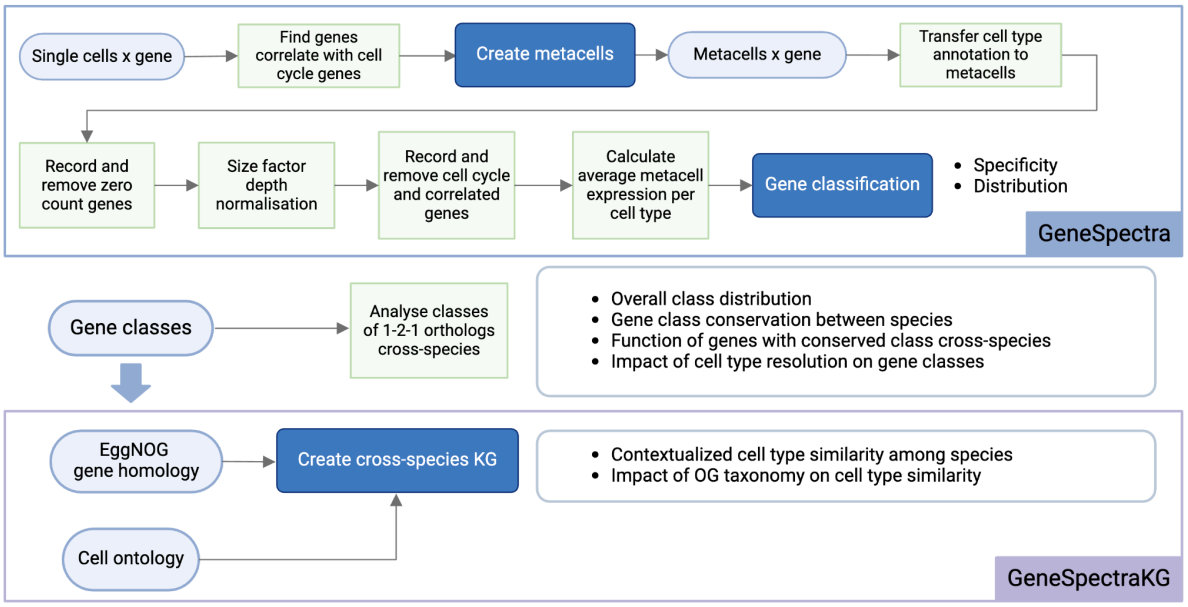

Figure 1 Schematic of the GeneSpectra and GeneSpectraKG framework. 1-2-1, one-to-one; OG, ortholog groups; KG, knowledge graph.

Since scRNA-seq experiments only capture a small percentage of transcripts per cell (Zheng et al., 2017), raw scRNA-seq data represents repeated sparse sampling of biological single cells. To reduce technical sparsity and maintain single cell granularity of the transcriptome profile, we employed the metacell approach (Baran et al., 2019; Ben-Kiki et al., 2022) (Supplementary figure 1, see details in Methods). We calculated the averaged expression data among metacells of the same cell type as the cell type expression profile input to GeneSpectra.

Following size factor normalisation, we performed the gene classification using all cell types in each species. We modified the HPA classification (Karlsson et al., 2021) classes and classification rules to implement in this study (Table 1). Genes were grouped by their expression specificity, a relative criteria, and expression distribution, an absolute criteria (see details in Methods). We applied the GeneSpectra classification to two datasets: a sn-RNA-seq data of the middle temporal gyrus (MTG) from five primates (Jorstad et al., 2023) for closely related species and an embryogenesis dataset from zebrafish and xenopus (Briggs et al., 2018) representing distantly related species (see Methods). It is important to

note that the primate data were analysed and annotated separately, based on marker genes of neuron and non-neuronal cell types found in the human MTG and M1 cortex (Jorstad et al., 2023). The cell types at this granularity are expected to be one-to-one corresponding among the 5 primates. The embryogenesis data were annotated independently with knowledge of developmental lineages.

Applying GeneSpectra to the MTG dataset shown that the specificity and distribution class proportions for all genes were highly similar among species, with old World monkeys (marmoset, macaque) and great apes (chimpanzees, gorilla) showing a notable resemblance within their respective groups (Figure 2a, b). Comparing the percentage of genes showing enrichment/enhancement in different cell types suggested that non-neuronal cell types had a higher number of specific genes, and the percentages stayed consistent among species (Figure 2c, d). We further observed that transcriptomically similar cell types shared more group specific genes, with non-neuronal cells again showing a high number of intersections (Supplementary Figure 2).

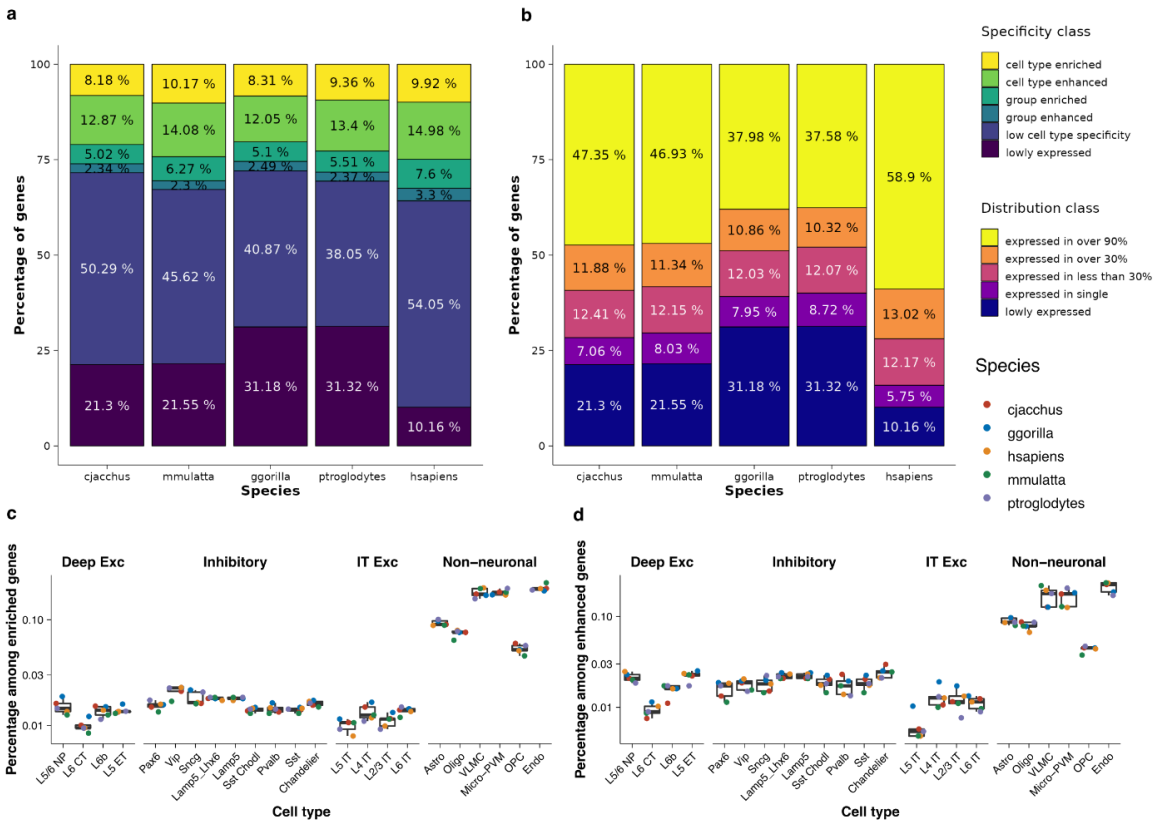

Figure 2 Gene class distribution in the MTG datasets. (a) Gene specificity class and (b) gene distribution class percentage of all genes in each species. (c) percentage of enriched genes and (d) enhanced genes in each cell type. Data points were shown as dots. The bar in the boxplot shows the arithmetic mean, lower and upper hinges correspond to the first and third quartiles, whiskers extend from the hinge to the largest value no further than 1.5 * interquartile range.

Gene class divergence is prominent among 1-2-1 orthologs

Since 1-2-1 orthologs were considered to be matching features in cross-species expression analysis, we asked to what degree they align in terms of cell type expression specificity and distribution. Therefore, we examined if 1-2-1 orthologous genes stayed in the same class between pairs of species in the MTG dataset and the embryogenesis dataset. These two datasets have small technical batch effects, therefore providing a reliable baseline scenario for different divergence time: ~40 million years ago (MYA) for the MTG data and ~420 MYA for the embryogenesis data. Note that the gene classes in each species were calculated completely independently.

We observed a notable frequency of specificity class switching among enriched or enhanced genes across species, with lowly specific genes predominantly maintaining their low specificity. Figure 3a shows the average gene specificity class conservation matrix of 1-2-1 orthologs in pairs of species for the MTG dataset (see Methods for matrix calculation, see Supplementary figure 3 for each species pair results). For genes enriched in a specific cell type or group, an average of 46.87% or 36.47% remained in the same category in the other species. The numbers were lower for cell type or group enhanced genes, with 30.35% or 17.42% maintaining their specificity category, respectively. In contrast, genes with low specificity had a remarkable degree of consistency (averagely 84.30% overlap between species). The embryogenesis dataset led to a similar conclusion (Supplementary figure 4), with even fewer genes remaining in the same specificity category, possibly due to larger divergence time.

We further examined whether genes retained specificity within the same cell type (Figure 3b, c, see Supplementary figure 3 for each species pair results). In the MTG dataset, for genes with cell type level specificity, over 92.22% of enriched genes that remained enriched were found in the same cell type. In contrast, approximately 38.38% of enhanced genes that

remained enhanced switched to another cell type (Figure 3b). The cell type changes for genes that were group enriched or enhanced were more pronounced, with around 28.01% or 40.00% of events occurring in a different cell type, respectively (Figure 3c). Moreover, almost half of the events in which the gene transitioned between group enriched and group enhanced also changed the corresponding cell types. Inspection of the frequency of switching between cell type pairs suggested that most genes switch their specificity between cell types known to be transcriptomically similar, such as between cerebral cortex endothelial cell and vascular leptomeningeal cell (see Supplementary figure 5a for frequency, Supplementary figure 5b for examples). The embryogenesis dataset showed a larger percentage of genes which switched cell type in different specificity categories (Supplementary figure 4).

Gene Ontology (GO) enrichment analysis on the most frequently conserved genes within each broad cell type among all species elucidated the roles of conserved specific genes between species in the MTG data (Supplementary figure 6). Genes involved in foundational neuronal activity, such as synaptic signalling, neural and glial development and extracellular matrix and cytoskeleton function mainly maintained their cell type specificity. The results suggest that the basic mechanism of neurotransmitter signalling across mammalian cells is highly conserved at expression level. Furthermore, the enriched terms are highly consistent with a recent study of the ancestral neural-specific gene modules using bulk RNA-seq data across vertebrates and insects (Figure 2i by (Mantica et al., 2024)).

In terms of expression distribution, widely expressed genes maintained their broad expression while selectively expressed genes exhibited divergence. Figure 3d illustrates the averaged gene distribution class conservation between 1-2-1 orthologs in pairs of species (see Supplementary figure 3 for each species pair results). For genes expressed in over 90% of cell types, the frequency of expression in a different cell type was only 1% (Figure 3e). Yet as high as 67.9% of genes changed cell types between species, if they were expressed in a single cell type (Figure 3f). Genes expressed in less than 30% and 30%-90% of cell types demonstrated more frequent cell type switching, particularly when they also changed between these distribution classes (Figure 3g). The distribution class is generally less conserved, and the cell type switching is more frequent in the embryogenesis data.

In summary, our analysis highlights substantial divergence in the gene expression specificity and distribution profiles of 1-2-1 orthologs with varying levels of cell type selectivity across

species. The divergence is already prominent among five mammals and is only stronger for evolutionarily distant species such as zebrafish and xenopus.

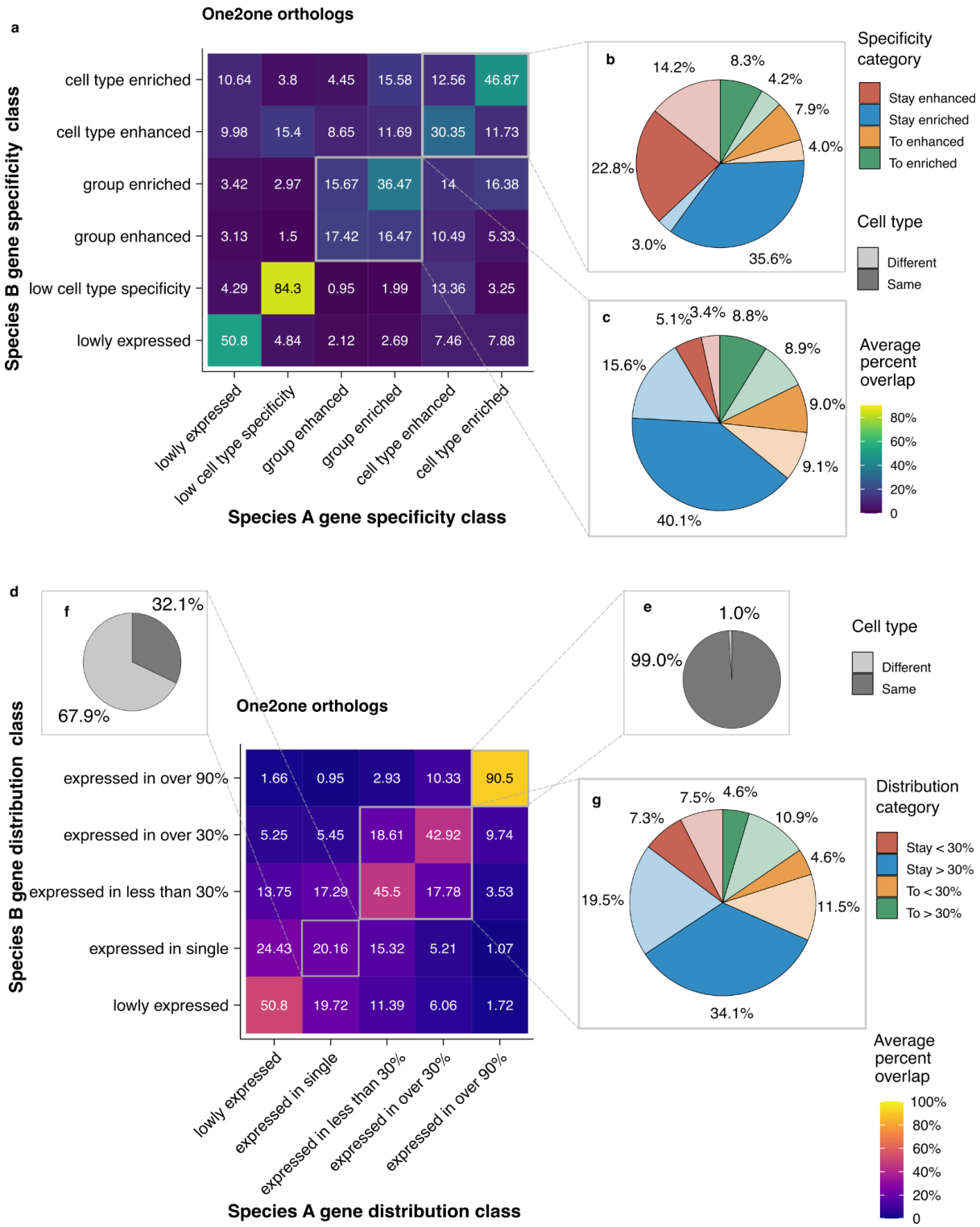

Figure 3 Gene class conservation among 1-2-1 orthologs in MTG dataset. To give an overview of the class conservation of 1-2-1 genes, we averaged the gene class conservation matrices among 10 pairs of species. The species were paired in a directional way, from species A to species B, with species B closer to humans (see Supplementary figure 3a for species tree and Supplementary figure 3 for results for each species pair). Therefore, the class matrices were interpreted as gene stays or changed from class in species A to class in species B along the direction of time. (a) Average specificity class conservation heatmap showing the percent overlap among one2one orthologs in various specificity class combinations (calculation process see methods). (b)-(c) Pie chart showing the aggregated percentage of cell type conservation and class conservation for genes with (b) cell type level specificity or (c) group level specificity. (d) Average distribution class conservation heatmap (calculation process see methods). (e)-(g) Pie chart showing the aggregated percentage of cell type conservation of genes stayed expressed in (e) over 90% of cell types; genes stayed expressed in (f) a single cell type and genes expressed in (g) some cell types.

Impact of cell type granularity and dataset comparability on gene classes

By varying the resolutions, one can classify cells into cell types, adding an extra layer of complexity for cross-species comparison. This is not only an ontological question of the annotation granularity, but also a biological question as the cell type evolves and specifies in more complex species. Here, demonstrate the flexibility of the GeneSpectra framework to allow for analysis across different cell type resolutions. We also show how gene classification results can inform the comparability of populations and granularity between datasets.

We first used a reduced granularity of the MTG dataset to obtain a gene class conservation matrix and compared it with the original results. After using a more coerced classification of cell types, we found that the gene class conservation specificity and distribution matrix remained largely consistent (Supplementary Figure 8 c, e). Moreover, a comparable percentage of specific genes switched cell type (Supplementary Figure 8). These results demonstrate that the gene classification and conservation analysis is stable with respect to the cell type granularity when the granularity used is comparable among species datasets, supporting the observation of frequent class switching between species. It is worth noting that changing the cell type granularity only requires rerunning the same script with a different

parameter (anno_col) in GeneSpectra. Such convenience allows for easy exploration of different granularities.

The MTG dataset is unique in the sense that it has the same cell populations and matched annotation among the 5 species due to conscious sampling and data processing (Suresh et al., 2023). When the analysis concerns independently annotated data from different species, it is challenging to determine the true comparability in terms of cell population and annotation granularity. We found that the gene specificity class conservation matrix can be an effective tool for assessing such comparability. Using GeneSpectra, we generated gene class conservation matrices from independently annotated human and mouse bone marrow datasets (Stumpf et al., 2020). The mouse dataset includes three additional cell types that were not profiled in the human dataset. When considering all cell types, the gene specificity class conservation matrices displayed a stripe of genes with specificity in the mouse but not in the human dataset (Supplementary Figure 9a). In contrast, when only shared cell types were used, the matrices exhibited a clear diagonal pattern (Supplementary Figure 9b). We conclude that the stripe pattern of genes with specificity on the gene specificity class conservation matrices can inform us that one species data has additional population. Following a similar intuition, if granularity is clearly higher in one species data, the gene class specificity and distribution conservation matrices will show a triangle pattern by having more specific genes, or more widely expressed genes in that species, respectively.

Gene classes explain species effect on the transcriptomic space

It is known that concatenated scRNA-seq data from different species exhibit strong global expression shift (Shafer, 2019) known as species effects (Song et al., 2023): cells show higher expression similarity within species, rather than within matching cell types, on the shared transcriptomic space built with 1-2-1 orthologs. We previously compared different computational strategies to correct this species effect and found it challenging to distinguish the source of such divergence (Song et al., 2023). Here, we demonstrate that gene classification results can explain the species effect using the MTG dataset.

For each pair of species, we obtained a list of 1-2-1 orthologs that stayed specific in the same specificity class (i.e. cell type or group enriched or enhanced, referred to as conserved specific genes hereafter). If the gene was also specific to the exact same cell type or groups, it was considered to be strictly conserved specific (examples of genes stayed enriched but

switched cell type is shown in supplementary figure 5b). We reason that the expression of these genes represent the evolutionarily conserved cell type transcriptomic profiles between species. Conversely, we also obtained a list of 1-2-1 orthologs that switched from specific to non-specific or lowly expressed in the other species (referred to as diverged to unspecific genes hereafter). We infer that these genes mainly contribute to the global expression divergence between the two species. Using the full list of 1-2-1 orthologs, strictly conserved specific genes, conserved specific genes and diverged to unspecific genes, we applied dimensional reduction to the single cell expression profile of data from each pair of species, and measured the strength of the species effect.

Figure 4a shows an example of the UMAP visualisation of data from human and chimpanzee (other species pair results available in Supplementary figure 10). It is clear that species effects present on the full set of 1-2-1 orthologs data. When using conserved specific genes, cell types such as Sst; Pavalb; Vip neurons and Oligo; OPC; Astro and Chandelier showed partial overlap on the expression space, suggesting a much reduced species effect. On the other hand, when using diverged to unspecific genes, we observed a stronger species effect which led to the two species data being linearly separable on the first 2 dimensions on the PC space (Figure 4b). It appeared that the global species effect was extracted by restricting the features to diverge to unspecific genes. The strictly conserved specific genes data was highly similar with the conserved specific genes both in UMAP space and in PC space, possibly due to similar highly variable features being selected for dimension reduction. We quantified the degree of species effect by using PC regression score (Luecken et al., 2021) (Figure 4c) and observed aligning trends with the dimensional reduction analysis.

We conclude that ortholog gene class switching from specific to non-specific mainly contributes to the species effect, alongside changes of the expression value of other genes. It is worth noting that since the MTG data have small technical batch effects, using GeneSpectra classes enables explaining the biological species effect.

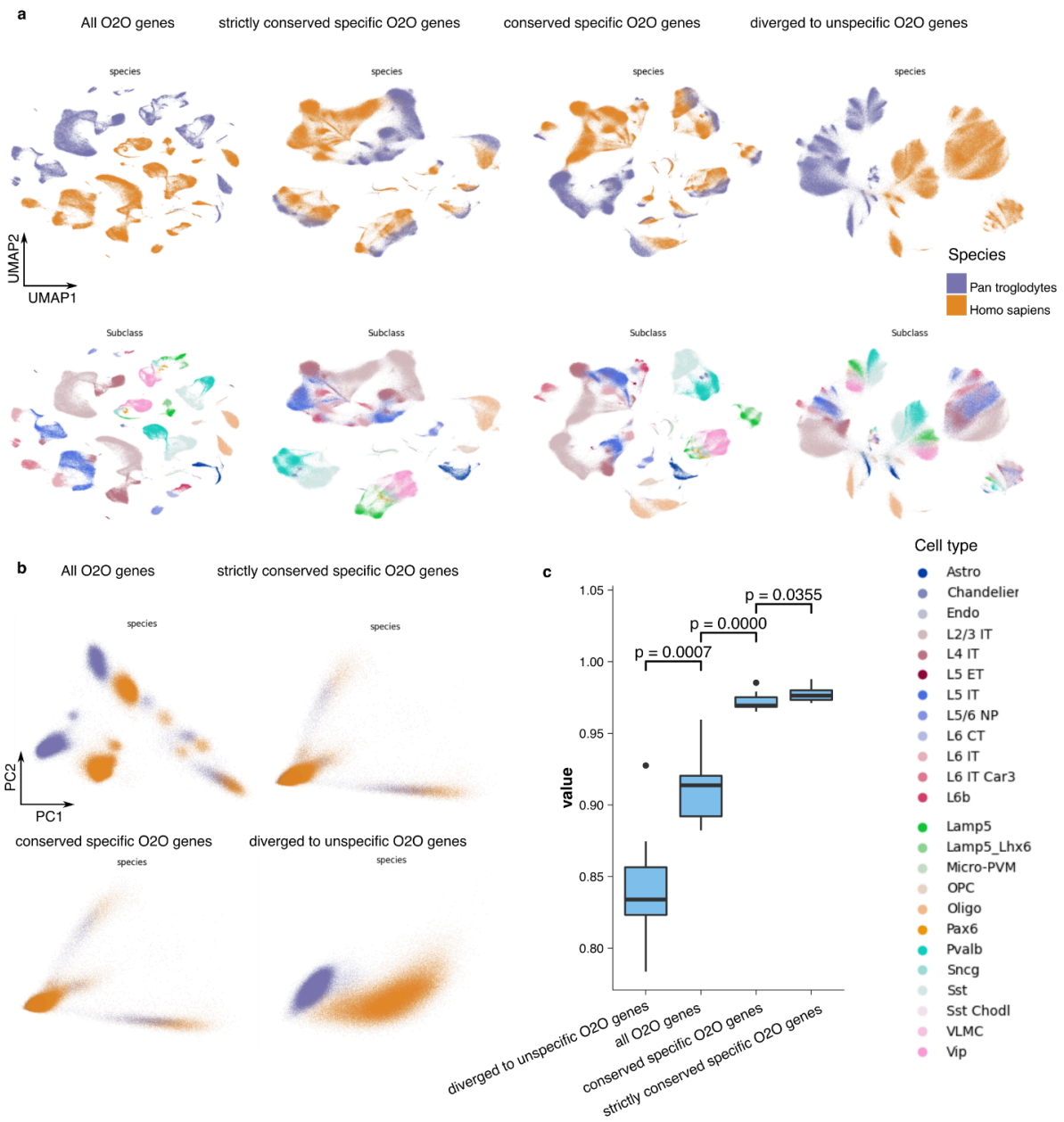

Figure 4 Gene class switching explains species effect on transcriptomic space. (a) UMAP visualisation using different sets of 1-2-1 orthologs for the human and chimpanzee data. Showing the strong species effect using all genes, the reduced species effect using (strictly) conserved specific genes, as well as a purified specie effect shown in diverged to unspecific genes (b) The first 2 PCs coloured by species using different sets of 1-2-1 orthologs for the human and chimpanzee data. (c) PCR score for data with different sets of genes. The bar in the boxplot shows the arithmetic mean, lower and upper hinges correspond to the first and third quartiles, whiskers extend from the hinge to the largest value no further than 1.5 * interquartile range and outliers beyond this range are plotted as individual data points.

Wilcoxon signed-rank test was used to determine statistical significance (p value shown). 1-2-1, one-to-one; UMAP, Uniform Manifold Approximation and Projection; PC, principal component.

Comparing sequence conservation and expression specificity similarity

Leveraging the gene classification results, we investigated the relationship between sequence conservation and the expression specificity similarity across cell types among different types of orthologs, with respect to the evolutionary proximity between species using the MTG dataset.

For orthologs of different types, we computed the gene peptide product sequence alignment using BLASTp (Altschul et al., 1990) and used the bit scores as the sequence conservation score. The specificity classification results of each gene can be represented as a vector, with values corresponding to the specificity scores of that gene in different cell types (Supplementary figure 11). By applying distance measurements such as the cosine distance, we quantified the expression specificity similarity between genes from different species. We plotted the cosine similarity of expression specificity against the bit score for cell type specific orthologs for each pair of species in the MTG data (Figure 5 for cell type enriched genes, see supplementary figure 12 for cell type enhanced genes).

We found a group of genes situated on the top right corner of the sequence-expression conservation landscape, representing those having a very high sequence conservation and a conserved cell type specificity (top left triangle in Figure 5). Simultaneously, the plot depicted a significant group of genes with diverse sequence conservation but no expression similarity, positioned at the lower edge. There are relatively few genes in the middle area. Interestingly, with the increase of species divergence time, we observed a spread out pattern of genes towards having a medium sequence conservation and expression specificity similarity, as well as an increase of the percentage of genes with high sequence conservation but zero expression specificity similarity. This suggests that distant species have a more diverged genome both in terms of sequence and gene expression. Considering different types of orthologs, we found that non-1-2-1 orthologs in general have lower sequence conservation than 1-2-1 orthologs (bottom right triangle in Figure 5), but we saw no clear bias in expression specificity similarity. Many non-1-2-1 orthologs with a range of sequence conservation showed high expression conservation. Our results agree with the observations

by Stamboulian, M. et al. (Stamboulian et al., 2020) that the types of homologs are irrelevant to function prediction tasks, highlighting the importance of including non-1-2-1 orthologs into functional analysis.

In a nutshell, our results suggested an overall positive relationship between sequence conservation and expression similarity in the MTG dataset, particularly between more closely related species. Nevertheless, a noteworthy amount of genes did not follow this pattern, and non-1-2-1 orthologs didn't show differences of expression specificity similarity compared with 1-2-1 orthologs, despite lower sequence conservation.

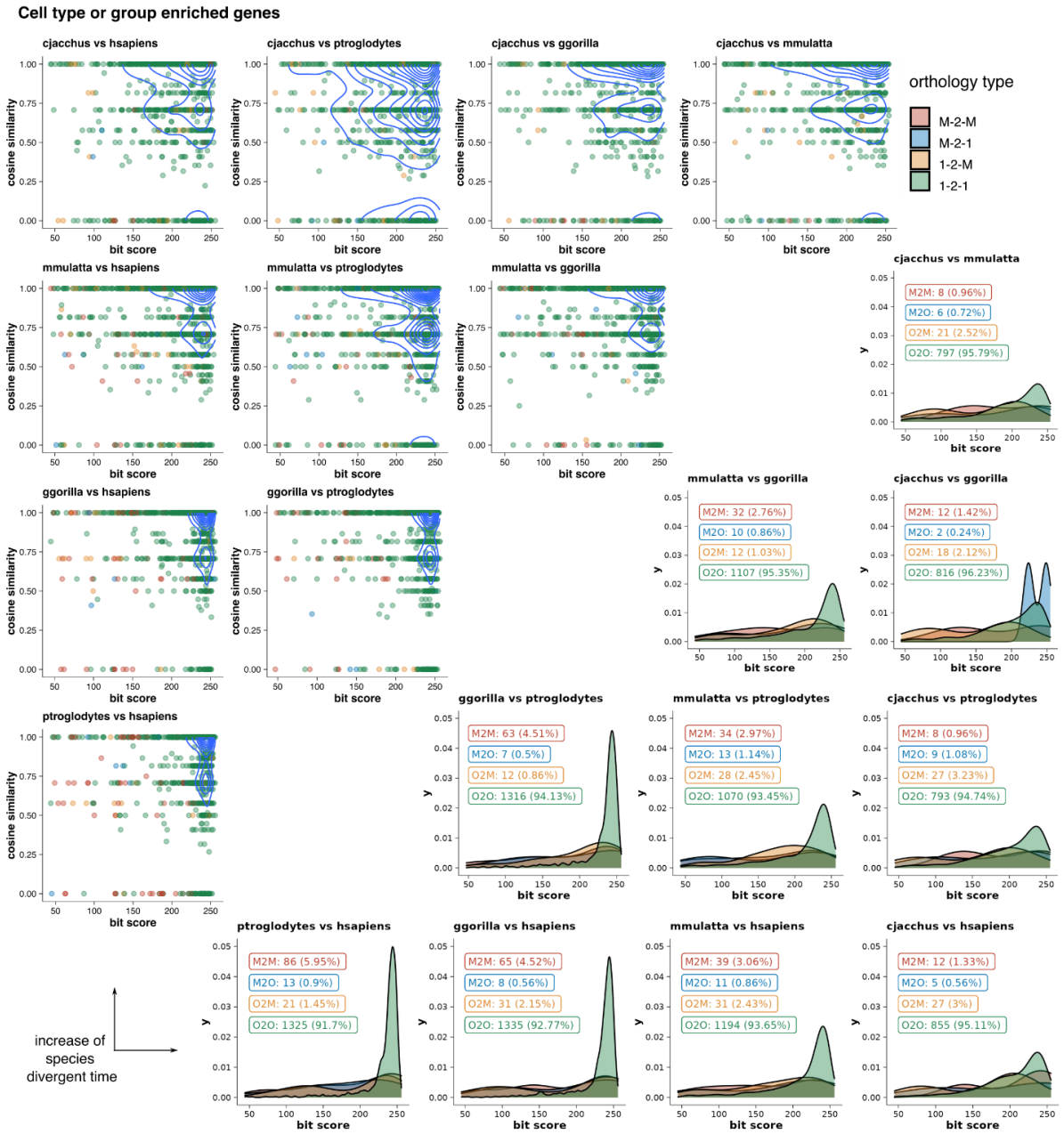

Figure 5 Comparing sequence conservation with cell type expression specificity similarity across species for enriched genes. The top left triangle plots the sequence conservation (bit score) against the expression similarity (cosine similarity) for each pair of cell type enriched orthologous in each pair of species. The bottom right triangle shows the distribution of bit scores for different types of orthologs for the same set of genes. 1-2-1: one-to-one, 1-2-M: one-to-many, M-2-1: many-to-one, M-2-M: many-to-many.

Constructing a knowledge graph of the cell type pattern of gene expression

To capture the intricacy of cross-species gene classification results, we constructed a knowledge graph which we named GeneSpectraKG. Using a graph structure to represent the results added several important pieces of information, providing the previously overlooked contexts. It enabled us to include genes unmapped by homology into the cross-species comparison, effectively represent non-1-2-1 orthologs, encompass an unlimited number of species in a single analysis and consider the taxonomy resolution on which orthologs were computed. Furthermore, the efficient database system to store and analyse data provides flexibility and scalability that is crucial for atlas-level multi-species comparisons.

An example of the GeneSpectraKG schema is shown in Figure 6a. There were four types of nodes: gene; OG; cell type and species. Edges connected the nodes to represent the species of origin of genes and cell types, the specificity of genes in cell types and the membership of genes in OGs. We collected the enriched events, either group or cell type, into genes enriched in cell type edges. Similar process was applied to enhanced genes. Finally, we also include edges for genes that were not specific in the respective species.

Contextualized cell type similarity measurements

We proceeded to use the GeneSpectraKG to perform a contextualised analysis of cell type similarity between species. We proposed two progressing layers to compare: the similarity by sharing cell type specific genes, and the similarity by having matching degrees of specificity of shared specific genes. Using information in the KG, we were able to measure both at the same time to achieve a holistic view of the cross-species cell type similarity.

To measure a contextualised similarity by sharing cell type specific genes, we employed a Jaccard-derived similarity score. The Jaccard-derived similarity score calculates the number

of specific genes in shared OGs divided by the total number of specific genes between two cell type nodes in different species (Figure 6b, can be extended to two cell type nodes in the same species by simply calculating the proportion of shared specific genes). Note that this calculation naturally includes species-specific genes that were not mapped to OGs, and considers the varying number of in-paralogs that are in the same OG with genes in other species. We calculated the Jaccard-derived similarity using enriched genes across all cell type pairs among the 5 species, showing in Figure 6c. Applying hierarchical clustering to the Jaccard-derived similarity matrix, we observed distinct clustering patterns depending on the cell type. Inhibitory neurons grouped by species, excitatory neurons grouped by species and broad type, and non-neuronal cells grouped by cell type. We also found that non-neuronal cells showed an overall higher Jaccard-derived similarity than neurons, in consensus with the author's original report that non-neuronal cells had less human-specific markers (Jorstad et al., 2023). Performing similar analysis on the reduced granularity data showed similar patterns, only with Endo, Micro_PVM and VLMC clustered mostly by species instead (Supplementary Figure 13). Finally, with the increase of species divergence time, there was a general decrease in Jaccard-derived similarity for each cell type (Supplementary Figure 14).

Focusing on shared specific genes between species, we examined the degree of similarity based on the relative strength of specificity. This was measured by calculating the cosine similarity of enrichment score, that is the expression level of a gene in the specific cell type divided by the second highest expression level, using all pairs of orthologs for each cell type (Figure 6b). The cosine similarity scores using enriched genes were calculated per cell type (Figure 6d). Interestingly, some cell types showed a very high cosine similarity score even between remote species, such as L5 ET; L6 IT Car3; L2/3 IT and L6 IT. However, non-neuronal cells had a lower value of cosine similarity, suggesting that the expression level of shared specific genes underwent divergence. There was also no longer a clear difference between species pairs of different divergent time, suggesting that cross-species differences at this layer were more driven by individual cell types.

In summary, our results highlight the importance of analysing cell type similarity cross-species at two different layers. With the increase of species evolutionary distance, cell types share less cell type-specific genes, but the relative degree of expression similarity of shared cell type-specific genes is more dependent on the cell type identity.

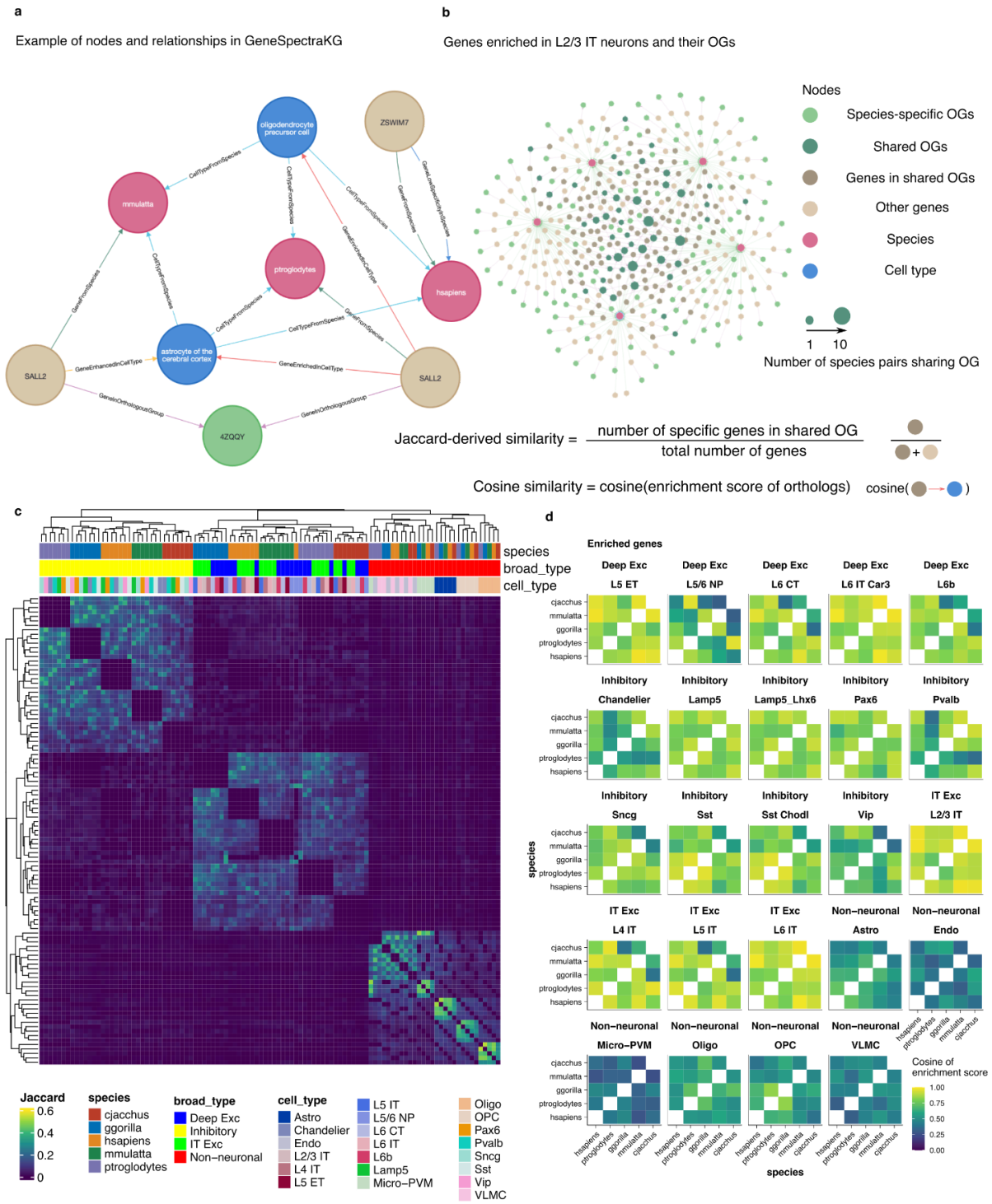

Figure 6 Contextualised analysis of cell type similarity between species using GeneSpectraKG. (a) An example showing all types of nodes and relationships in the GeneSpectraKG. (b) An example query showing all the genes that enrich in the L2/3 IT neurons from all species. Taken from a screenshot of the Neo4j Bloom. Formulas show how to calculate the Jaccard similarity and the cosine similarity between cell types based on the

query result. (c) Jaccard-derived similarity matrix of all cell types in all species. Heatmap with hierarchical clustering represents the Jaccard-derived similarity score with OGs calculated at Primates level. (d) Cosine similarity of each cell type in all pairs of species using orthologs at Primates level.

Impact of orthologous group taxon on cell type similarity

The orthology relationships between genes depends on the reference speciation event (Hernández-Plaza et al., 2023; Huerta-Cepas et al., 2019). Leveraging the flexibility of GeneSpectraKG to include OGs at various taxonomy levels, we analysed the cell type Jaccard-derived similarity of the same cell type across these levels. It is clear that when considering higher granularity OGs, the overall Jaccard-derived similarity decreases. We attribute this to the fact that calculating OGs on more recent taxonomy levels will therefore lead to smaller OGs with less in-paralogs.

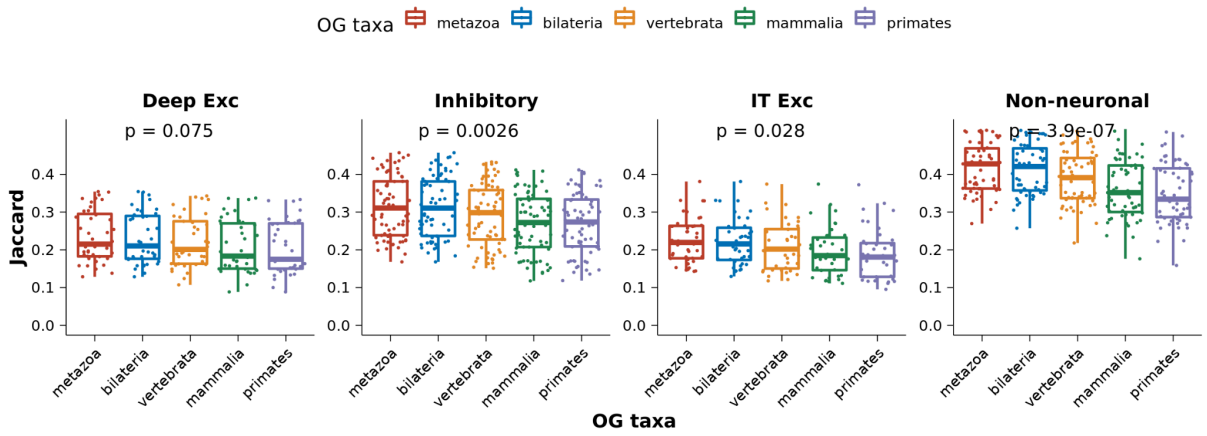

Figure 7 Impact of orthologous group taxonomy on cell type Jaccard-derived similarity. Each dot corresponds to the Jaccard-derived similarity score of the same cell type in one pair of species using OGs calculated at a certain taxonomy level, grouped by their broad type. The box plot shows the Jaccard-derived similarity of the same cell type using different orthologous group taxons. The bar in the boxplot shows the arithmetic mean, lower and upper hinges correspond to the first and third quartiles, whiskers extend from the hinge to the largest value no further than 1.5 * interquartile range. Kruskal–Wallis test was used to determine statistical significance (p value shown).

Discussion

We present GeneSpectra/GeneSpectraKG, a computational framework for contextualised cross-species comparison of cell type gene expression using scRNA-seq data. This work represents a conceptual advancement by reassessing the assumption of ortholog conjecture and including in-paralogs and species-specific genes to measure cell type expression similarity. Our reassessment of the ortholog conjecture showed that only a subset of 1-2-1 orthologs retains cell type specificity across species, while most lowly specific genes maintained their low specificity. This captures the important distinction of the specificity conservation level between species, for genes whose expression is regulated per cell type, and genes that have broad and consistent expression across cell types. The high frequency of specificity class switching between orthologs also suggests that it is essential to include non-1-2-1 orthologs and species-specific genes into calculating the expression similarity of cell types. Additionally, it is worth noting that many cross-species expression differences detected computationally could be evolutionarily neutral (Jorstad et al., 2023). The addition of functional data will help to disentangle the signals that are relevant to species adaptation.

Understanding the expression specificity conservation of orthologs provides a basis for developing new approaches for evolutionary mapping of cell types cross-species. The evolutionary definition of cell types emphasises the distinction between the regulatory mechanism that establishes cell type identity, and gene modules that produce specific cell type phenotypes in a species (Arendt et al., 2016). The terminal selector transcription factors (Hobert, 2008) and their effector genes maintain their cell type-specific expression between species and is indicative of evolutionary cell type homology. Other genes that contribute to cellular function but are not deterministic of cell identity can have more variable expression. Our analysis to obtain strictly conserved cell type-specific genes essentially detangles the two signals in scRNA-seq data. Strictly conserved cell type-specific orthologs can serve as candidates of terminal selectors and effector genes for developing methods for aligning evolutionarily related cell types across species, a process that can be further refined with additional data modalities for gene regulation.

Focusing on the evolutionary mapping of cells based on a few strictly conserved specific genes can also help resolve the challenge of circularity. The issue of circularity arises when

cell types are defined via aligning the full scRNA-seq atlas between species, and then these cell type classifications are used to study cross-species expression conservation of genes. For instance, one common approach involves creating a consensus cell type hierarchy between species by transferring cell type annotation from a well-studied reference species to others based on a set of genes and gene orthology (e.g. with mapping algorithms such as XGBoost (T. Chen & Guestrin, 2016; van Zyl et al., 2020) or MetaNeighbor (Crow et al., 2018; Suresh et al., 2023)). While this method can yield detailed consensus cell type clusters between closely related species, it may inadvertently lead to circular reasoning when the cross-species consensus cell types are subsequently used to evaluate the degree of gene expression conservation. Therefore, it's important to only use the gene expression signals that define cell type identity to match cell types between species, or annotate cell types independently between species, then compare at the full transcriptome level for studying species-specific expression.

Our KG-based cell type similarity analysis reveals that certain cell types have accumulated more species-unique cell type group enriched genes, which makes them exhibit greater transcriptomic similarity to other cell types within the same species than to their corresponding cell types in different species. Importantly, this is observed in some cell types (e.g. neurons in MTG data) but not in others (e.g. non-neuronal cells in MTG data). In short, the analysis highlighted concerted expression evolution (Arendt et al., 2016) among the neuronal cell types. This finding also echoes the analysis by Liang et al., (Liang et al., 2018) regarding the pervasive correlated expression evolution between cell types and tissues. They attribute such observation to the accumulation of species-specific expression, likely caused by the pleiotropic effect of mutations (Liang et al., 2018). While they found that different tissues or cell types differ in correlated evolution signals in bulk RNA-seq data, we were the first to observe agreeing patterns in scRNA-seq data. We further hypothesise that the varying degree of species-specific expression accumulated in different cell types provides the opportunity to align the timeline of cell type divergence respective to speciation.

We envision that the GeneSpectra/GeneSpectraKG paradigm will be increasingly beneficial as we begin to explore the evolutionary relationships of cell types between distantly related species. As the genome and its context evolve, new genes specific to a species can significantly contribute to the identity-defining regulatory mechanisms of species-specific cell types. It is known that after gene duplication, if both copies are retained in the genome, they tend to evolve faster than their orthologs that have only single copies (Kondrashov et

al., 2002). It will be interesting to study how the gene expression of such amplified genes evolve compared with 1-2-1 orthologs.

There are still many ongoing challenges in the field. Ensuring dataset comparability is essential for informative comparisons, so it is important to generate datasets with future cross-study compatibility in mind. It is also challenging to find the correct granularity for cell type comparison that is biologically meaningful and technically feasible. We show that gene class conservation matrices can serve as an initial check for the comparability of cell populations and annotation granularity, yet biological knowledge about the species biology and evolution is crucial to interpret these data and perform cell type mapping between species.

Methods

Single cell RNA-seq data used in this study

All datasets analysed in this study are publicly available. Sn-RNA-seq count matrices from humans, chimpanzees, gorillas, rhesus macaques, and marmosets were downloaded from the publication release “Great_Ape_MTG_Analysis" from the NeMO archive (*Index of /publication_release/Great_Ape_MTG_Analysis*, n.d.). Author cell ontology annotations were obtained from Cellxgene of the same dataset (*Cellxgene Data for Comparative Transcriptomics Reveals Human-Specific Cortical Features*, n.d.). Only 10x V3 3’ data and cells passed the original QC of the published dataset were analysed in this study to minimise technical effects. InDrops data of zebrafish and xenopus embryos are available via the GEO database under accession code GSE11229434 and GSE11307435, respectively. Human and mouse bone marrow data are available from ArrayExpress under accession number E-MTAB-8629 and E-MTAB-8630.

EggNOG homology data

EggNOG homology data were downloaded from EggNOG V6 (Hernández-Plaza et al., 2023). The e6.og2seqs_and_species.tsv data were splitted by taxa and only the records containing at least one gene from the snRNA-seq data of the species analysed were included. Note that ortholog groups are computationally calculated results therefore other sources can also be used for GeneSpectraKG.

Constructing metacells

Metacells partition the raw single cell data into disjoint, homogeneous groups ideally representing re-sampling from the same cellular state (Ben-Kiki et al., 2022). We downloaded metacells (V0.9.4) from GitHub (https://github.com/tanaylab/metacells). Metacells were constructed with default hyperparameters following the tutorial https://tanaylab.github.io/metacells-vignettes/iterative.html. Lateral genes related to cell cycle and stress, and genes highly correlate with these genes were masked during metacell construction. Metacell cell type annotations were conveyed from single cell annotations by maximum percentage per metacell. In the MTG data, approximately 15-25 single cells, nearly 100% of the same cell type, constituted one metacell (Supplementary figure 1).

Cell type pooling

Due to the continuity of cell type expression in embryonic settings, we employed an alternative approach for sparsity reduction which is cell type pooling. We summed the count values of cells from the same type as input for GeneSpectra, which was then subjected to depth normalisation.

The GeneSpectra module for gene classification

The GeneSpectra module was developed under Python 3.8.16. It takes an AnnData (v0.9.1) object containing metacells and calculates the average normalised expression for each group (in this context, cell type). After transforming the results into a long data format, it applies a classification function to each gene when using multithreading to speed up the computation. Finally, it outputs a dataframe containing specificity and distribution classification results. Various plotting and analysis functions were also implemented to aid comprehension and downstream analysis.

The gene expression specificity class is determined as follows: if the normalised expression level of the gene in a certain cell type, or the average of a group of cell types, was 4 folds higher than the maximum of all the other cell types, it was considered enriched. On the other hand, if the elevation was only 4 folds higher than the average of all the other cell types, it was considered enhanced. Cell type or group enriched or enhanced genes were collectively

referred to as specific genes in this paper. The gene expression distribution class was determined by the percentage of cell types where the normalised expression level was above 1 CPM. In both criteria, genes that had an expression < 1 CPM in all cell types are considered lowly expressed.

Gene class conservation matrix

Gene specificity and distribution class conservation matrix of 1-2-1 orthologs for each pair of species was calculated as follows: we first matched the 1-2-1 orthologs using ENSEMBL and counted the number of 1-2-1 orthologs in each class for both species. We then counted the intersection of 1-2-1 orthologs in all class combinations between the pair of species. For each species, we then calculated the percentages of intersection respective to the total number of genes in that class. The final percentages shown in the heatmap is the harmonic mean of the percentages from the two species.

To show an average class conservation among 10 pairs of species, the species were paired in a directional way, from species A to species B, with species B more closely related to humans. As such, the averaged heatmap can be comprehended as genes move from class in species A to species B along the direction of time.

ScRNA-seq data analysis and visualisation

To perform analysis and visualisation of scRNA-seq data using different sets of 1-2-1 orthologs from pairs of species, we used Scanpy (v1.4.6) with the following parameters: sc.pp.highly_variable_genes(adata, min_mean = 0.0125, max_mean = 3, min_disp = 0.5); sc.pp.scale(adata, max_value = 10); sc.tl.pca(adata, svd_solver = 'arpack'); sc.pp.neighbours(adata, n_neighbors = 15, n_pcs = 40); sc.tl.umap(adata, min_dist = 0.3, spread = 1).

BLASTp

The following protein translation of ENSEMBL genes were downloaded from ENSEMBL: human (Homo_sapiens.GRCh38.pep.all.fa); chimpanzee(Pan_troglodytes.Pan_tro_3.0.pep.all.fa); gorilla (Gorilla_gorilla.gorGor4.pep.all.fa); rhesus macaque (Macaca_mulatta.Mmul_10.pep.all.fa);

marmoset (Callithrix_jacchus.mCalJac1.pat.X.pep.all.fa). We ran BLASTp (V2.15.0) using default parameters.

KG construction

Using outputs from the GeneSpectra module and orthology data from EggNOG, we constructed GeneSpectraKG, a cross-species knowledge graph for the cell type pattern of gene expression. We leveraged BioCypher (Lobentanzer et al., 2023) (V0.5.19) to parse data into a KG and export an object to be hosted in Neo4j (V4.4.26) graph database management system. We used the Neo4j browser and Neo4j Bloom from Neo4j desktop (V1.5.8) to visualise and interact with the graph.

Software availability

The GeneSpectra module is available at https://github.com/Papatheodorou-Group/GeneSpectra and the version of code used in this study is available via Zenodo with DOI: 10.5281/zenodo.13303986. The GeneSpectraKG module is available at https://github.com/Papatheodorou-Group/GeneSpectraKG and the version of code used in this study is available via Zenodo with DOI: 10.5281/zenodo.12209581.

Declarations

The authors declare no competing interests.

Author contributions

All authors conceived this study. Y.S. developed the softwares, performed analysis and wrote the draft manuscript with supervision from I.P. and A.B.. All authors reviewed and edited the manuscript.

Acknowledgements

This work was supported by: the European Molecular Biology Laboratory (Y.S., A.B., I.P.); the EMBL international PhD program (Y.S.); and the Biotechnology and Biological Sciences

Research Council (BBSRC) grant 'Fly Cell Atlas' [BB/T014563/1] (I.P.). Y.S. thanks Prof Detlev Arendt, Prof Jake Musser and Dr Sebastian Lobentanzer for discussions.

Tables

Table 1 Gene expression specificity and distribution classes criteria. The gene classes and criteria are modified from the Human Protein Atlas (Karlsson et al., 2021) gene classification on single cell RNA-seq data. CPM, counts per million.

Gene expression specificity	Class	Criteria
	Lowly expressed	Average expression in all cell type < 1 CPM
	Low cell type specificity	All genes which do not satisfy the criteria of being specific
	Group Enhanced	Expression level (CPM) in some cell types >= 4-fold average expression across all cell types
	Cell type Enhanced	Expression level (CPM) in one cell type >= 4-fold average expression across all cell types
	Group enriched	Average expression (CPM) in a group of cell types >= 4-fold maximum expression in all other cell types
	Cell type enriched	Expression (CPM) in one cell type >= 4-fold maximum expression in all other cell types
Gene expression distribution	Class	Criteria
	Lowly expressed	Average expression in all cell type < 1 CPM
	Expressed in single	Expression level over 1 CPM in one cell type
	Expressed in few	Expression level over 1 CPM in less than 30% of all cell types
	Expressed in many	Expression level over 1 CPM in over 30% but less than 90% of all cell types

	Expressed in most	Expression level over 1 CPM in over 90% of all cell types

References

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., & Dessimoz, C. (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology*, *8*(5), e1002514.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D., & Wagner, G. P. (2016). The origin and evolution of cell types. *Nature Reviews. Genetics*, *17*(12), 744–757.

Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., & Tanay, A. (2019). MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biology*, *20*(1), 206.

Ben-Kiki, O., Bercovich, A., Lifshitz, A., & Tanay, A. (2022). Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. *Genome Biology*, *23*(1), 100.

Biharie, K., Michielsen, L., Reinders, M. J. T., & Mahfouz, A. (2023). Cell type matching across species using protein embeddings and transfer learning. *Bioinformatics* , *39*(39 Suppl 1), i404–i412.

Briggs, J. A., Weinreb, C., Wagner, D. E., Megason, S., Peshkin, L., Kirschner, M. W., & Klein, A. M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, *360*(6392). https://doi.org/10.1126/science.aar5780

Cellxgene data for Comparative transcriptomics reveals human-specific cortical features. (n.d.).

https://cellxgene.cziscience.com/collections/4dca242c-d302-4dba-a68f-4c61e7bad553

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, X., & Zhang, J. (2012). The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Computational Biology*, *8*(11), e1002784.

Crow, M., Paul, A., Ballouz, S., Huang, Z. J., & Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nature Communications*, *9*(1), 884.

Domcke, S., & Shendure, J. (2023). A reference cell tree will serve science better than a reference cell atlas. *Cell*, *186*(6), 1103–1114.

Evolution at the cellular level. (2023). *Nature Ecology & Evolution*, *7*(8), 1155–1156.

Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., Wimberger, P., Huttner, W. B., & Hiller, M. (2018). Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife*, *7*, e32332.

Franjic, D., Skarica, M., Ma, S., Arellano, J. I., Tebbenkamp, A. T. N., Choi, J., Xu, C., Li, Q., Morozov, Y. M., Andrijevic, D., Vrselja, Z., Spajic, A., Santpere, G., Li, M., Zhang, S., Liu, Y., Spurrier, J., Zhang, L., Gudelj, I., … Sestan, N. (2022). Transcriptomic taxonomy and neurogenic trajectories of adult human, macaque, and pig hippocampal and entorhinal cells. *Neuron*, *110*(3), 452–469.e14.

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews. Genetics*, *14*(5), 360–366.

Geirsdottir, L., David, E., Keren-Shaul, H., Weiner, A., Bohlen, S. C., Neuber, J., Balic, A., Giladi, A., Sheban, F., Dutertre, C.-A., Pfeifle, C., Peri, F., Raffo-Romero, A., Vizioli, J., Matiasek, K., Scheiwe, C., Meckel, S., Mätz-Rensing, K., van der Meer, F., … Prinz, M. (2019).

Cross-Species Single-Cell Analysis Reveals Divergence of the Primate Microglia Program. *Cell*, *179*(7), 1609–1622.e16.

Hernández-Plaza, A., Szklarczyk, D., Botas, J., Cantalapiedra, C. P., Giner-Lamia, J., Mende, D. R., Kirsch, R., Rattei, T., Letunic, I., Jensen, L. J., Bork, P., von Mering, C., & Huerta-Cepas, J. (2023). eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Research*, *51*(D1), D389–D394.

Hobert, O. (2008). Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(51), 20067–20071.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314.

Index of /publication_release/Great_Ape_MTG_Analysis. (n.d.). Retrieved March 4, 2024, from https://data.nemoarchive.org/publication_release/Great_Ape_MTG_Analysis/

Jorstad, N. L., Song, J. H. T., Exposito-Alonso, D., Suresh, H., Castro-Pacheco, N., Krienen, F. M., Yanny, A. M., Close, J., Gelfand, E., Long, B., Seeman, S. C., Travaglini, K. J., Basu, S., Beaudin, M., Bertagnolli, D., Crow, M., Ding, S.-L., Eggermont, J., Glandon, A., … Bakken, T. E. (2023). Comparative transcriptomics reveals human-specific cortical features. *Science*, *382*(6667), eade9516.

Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., Edfors, F., Oksvold, P., von Feilitzen, K., Zwahlen, M., Arif, M., Altay, O., Li, X., Ozcan, M., Mardinoglu, A., … Lindskog, C. (2021). A single–cell type transcriptomics map of human tissues. In *Science Advances* (Vol. 7, Issue 31). https://doi.org/10.1126/sciadv.abh2169

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biology*, *3*(2), RESEARCH0008.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, *39*, 309–338.

Kryuchkova-Mostacci, N., & Robinson-Rechavi, M. (2016). Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. *PLoS Computational Biology*, *12*(12), e1005274.

Liang, C., Musser, J. M., Cloutier, A., Prum, R. O., & Wagner, G. P. (2018). Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes. *Genome Biology and Evolution*, *10*(2), 538–552.

Liu, T., Li, J., Yu, L., Sun, H.-X., Li, J., Dong, G., Hu, Y., Li, Y., Shen, Y., Wu, J., & Gu, Y. (2021). Cross-species single-cell transcriptomic analysis reveals pre-gastrulation developmental differences among pigs, monkeys, and humans. *Cell Discovery*, *7*(1), 1–17.

Lobentanzer, S., Aloy, P., Baumbach, J., Bohar, B., Carey, V. J., Charoentong, P., Danhauser, K., Doğan, T., Dreo, J., Dunham, I., Farr, E., Fernandez-Torras, A., Gyori, B. M., Hartung, M., Hoyt, C. T., Klein, C., Korcsmaros, T., Maier, A., Mann, M., … Saez-Rodriguez, J. (2023). Democratizing knowledge representation with BioCypher. *Nature Biotechnology*, *41*(8), 1056–1059.

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., & Theis, F. J. (2021). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*. https://doi.org/10.1038/s41592-021-01336-8

Mantica, F., Iñiguez, L. P., Marquez, Y., Permanyer, J., Torres-Mendez, A., Cruz, J., Franch-Marro, X., Tulenko, F., Burguera, D., Bertrand, S., Doyle, T., Nouzova, M., Currie, P. D., Noriega, F. G., Escriva, H., Arnone, M. I., Albertin, C. B., Wotton, K. R., Almudi, I., … Irimia, M. (2024). Evolution of tissue-specific expression of ancestral genes across

vertebrates and insects. *Nature Ecology & Evolution*, 1–14.

Marioni, J. C., & Arendt, D. (2017). How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. *Annual Review of Cell and Developmental Biology*, *33*, 537–553.

Najle, S. R., Grau-Bové, X., Elek, A., Navarrete, C., Cianferoni, D., Chiva, C., Cañas-Armenteros, D., Mallabiabarrena, A., Kamm, K., Sabidó, E., Gruber-Vodicka, H., Schierwater, B., Serrano, L., & Sebé-Pedrós, A. (2023). Stepwise emergence of the neuronal gene expression program in early animal evolution. *Cell*, *186*(21), 4676–4693.e29.

Natsidis, P., Kapli, P., Schiffer, P. H., & Telford, M. J. (2021). Systematic errors in orthology inference and their effects on evolutionary analyses. *iScience*, *24*(2), 102110.

Nehrt, N. L., Clark, W. T., Radivojac, P., & Hahn, M. W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, *7*(6), e1002073.

Park, Y., Muttray, N. P., & Hauschild, A.-C. (2024). Species-agnostic transfer learning for cross-species transcriptomics data integration without gene orthology. *Briefings in Bioinformatics*, *25*(2). https://doi.org/10.1093/bib/bbae004

Rogozin, I. B., Managadze, D., Shabalina, S. A., & Koonin, E. V. (2014). Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biology and Evolution*, *6*(4), 754–762.

Rosen, Y., Brbić, M., Roohani, Y., Swanson, K., Li, Z., & Leskovec, J. (2024). Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nature Methods*. https://doi.org/10.1038/s41592-024-02191-z

Santos, M. E., Le Bouquin, A., Crumière, A. J. J., & Khila, A. (2017). Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. *Science*, *358*(6361), 386–390.

Sebé-Pedrós, A., Saudemont, B., Chomsky, E., Plessier, F., Mailhé, M.-P., Renno, J., Loe-Mie, Y., Lifshitz, A., Mukamel, Z., Schmutz, S., Novault, S., Steinmetz, P. R. H., Spitz, F., Tanay, A., &

Marlow, H. (2018). Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell*, *173*(6), 1520–1534.e20.

Shafer, M. E. R. (2019). Cross-Species Analysis of Single-Cell Transcriptomic Data. *Frontiers in Cell and Developmental Biology*, *7*. https://doi.org/10.3389/fcell.2019.00175

Shafer, M. E. R., Sawh, A. N., & Schier, A. F. (2021). Gene family evolution underlies cell-type diversification in the hypothalamus of teleosts. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-021-01580-3

Song, Y., Miao, Z., Brazma, A., & Papatheodorou, I. (2023). Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nature Communications*, *14*(1), 6495.

Sonnhammer, E. L. L., Gabaldón, T., Sousa da Silva, A. W., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P. D., Dessimoz, C., & Quest for Orthologs consortium. (2014). Big data and other challenges in the quest for orthologs. *Bioinformatics* , *30*(21), 2993–2998.

Stamboulian, M., Guerrero, R. F., Hahn, M. W., & Radivojac, P. (2020). The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics* , *36*(Suppl_1), i219–i226.

Stumpf, P. S., Du, X., Imanishi, H., Kunisaki, Y., Semba, Y., Noble, T., Smith, R. C. G., Rose-Zerili, M., West, J. J., Oreffo, R. O. C., Farrahi, K., Niranjan, M., Akashi, K., Arai, F., & MacArthur, B. D. (2020). Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell RNA sequencing. *Communications Biology*, *3*(1), 736.

Suresh, H., Crow, M., Jorstad, N., Hodge, R., Lein, E., Dobin, A., Bakken, T., & Gillis, J. (2023). Comparative single-cell transcriptomic analysis of primate brains highlights human-specific regulatory evolution. *Nature Ecology & Evolution*, *7*(11), 1930–1943.

Tanay, A., & Sebé-Pedrós, A. (2021). Evolutionary cell type mapping with single-cell genomics. *Trends in Genetics: TIG*, *37*(10), 919–932.

Tarashansky, A. J., Musser, J. M., Khariton, M., Li, P., Arendt, D., Quake, S. R., & Wang, B. (2021). Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *eLife*, *10*. https://doi.org/10.7554/eLife.66747

Tritschler, S., Thomas, M., Böttcher, A., Ludwig, B., Schmid, J., Schubert, U., Kemter, E., Wolf, E., Lickert, H., & Theis, F. J. (2022). A transcriptional cross species map of pancreatic islet cells. *Molecular Metabolism*, *66*, 101595.

van Zyl, T., Yan, W., McAdams, A., Peng, Y.-R., Shekhar, K., Regev, A., Juric, D., & Sanes, J. R. (2020). Cell atlas of aqueous humor outflow pathways in eyes of humans and four model species provides insight into glaucoma pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(19), 10339–10349.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049.