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, and lose expression information given by non-orthologs. To address this, we devise a novel analytical framework that redefines the analysis paradigm. This framework robustly classifies genes by expression specificity and distribution across cell types, allowing for a dataset-specific reassessment of the ortholog conjecture by evaluating the degree of ortholog class conservation. We utilise the 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, allowing hierarchical depiction of cell types and orthologous groups, and continuous integration of 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 transcriptomics data can shed light on the evolutionary links between cell types [1,2]. Currently, computational methods for cross-species comparisons often rely on expression correlation analysis or data integration, each with its own set of assumptions about genes or cells.

To measure expression similarity between cells across species, one needs to match features, i.e. genes, by gene orthology. This approach assumes that orthologs are equivalent in their ability to distinguish cell types [3–5]. To be specific, in correlation analysis, one-to-one (O2O) orthologs are often treated as identical features for calculating the degree of cell type similarity [6–10]. In data integration, algorithms require orthology mapped data as input to align transcriptomically similar cell types [6,11–13]. Both methods rely on the assumption that similar relative expression levels of orthologs make cell types from different species alike. This assumption is rooted in the ortholog conjecture, which is a foundational hypothesis in gene function prediction [14,15]. Here, the hypothesis is implicitly extended to cell type gene expression, as it can be considered as a functional readout of the gene. Although the validity of ortholog conjecture has been intensely debated [14–21], it is agreed that having an empirical, quantitative estimate of the functional similarity of orthologs, respective to their evolutionary relationship, is a more careful approach [14]. Therefore, examining the conservation of ortholog cell type expression patterns, by criteria such as the Human Protein Atlas (HPA) gene classification [22], is helpful for reassessing the validity of the ortholog conjecture in cross-species cell type comparisons.

Recently, several cross-species integration tools had been developed to tackle the restriction by gene orthology. For instance, SATURN [23] and TACTiCS [24] utilise protein sequence to refine gene homology annotations; SAMap [25] adopts an iterative process to update gene-gene mapping and cell-cell mapping; SATL [26] 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, or dataset comparability or some cross-species cell type mappings are given.

Cell type is inherently a hierarchical concept: the exact type of cells depends on the annotation granularity [27]. The knowledge of key cell type marker genes at low granularity 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 for a high cell type granularity comparison. One popular approach is to define a consensus cell type hierarchy between species by transferring cell type annotation from a well-studied reference species to others based on marker genes and gene orthology (e.g. with mapping algorithms such as XGBoost [28,29] or MetaNeighbor [30,31]). However, we reckon that potential circularity can emerge as a major challenge when the cross-species consensus cell types are subsequently used to evaluate the degree of gene expression conservation. This is because the consensus cell types are themselves defined by sharing gene expression features. This challenge is difficult to overcome, because cell type annotations are extensively *ad hoc* [27], meaning that there lacks a principled and unified system to define and name cell types between species. Nevertheless, annotation transfer of high granularity cell types is only appropriate among very closely related species.

Importantly, the assumptions about genes or cells become relatively questionable among evolutionarily remote species. Restricting the analysis to orthologs will become a prominent issue as expression signals given by species-specific genes become significant. The number of cell types and the cell type granularity that is appropriate for different species is also inherently divergent. In particular, when comparing distantly related species, it is important to check whether independently annotated data are truly comparable in terms of cell populations and annotation granularity.

On observing the current analysis paradigm, our foremost priority is to encourage a more holistic approach for cross-species cell type comparison tailored to the system of interest. It is essential to empirically calibrate the dependence on ortholog conjecture, and include the expression signals from species-specific genes during comparison. Furthermore, it is necessary to consider several contextual factors that influence the perceived similarity between cell types, such as the cell type granularity and gene orthology resolution. These considerations become increasingly pertinent as we enter the era of building cross-species cell atlases at scale [32].

In this study, we devise a novel analytical framework for cross-species comparison of single cell data, focusing on a dataset-specific examination of ortholog conjecture and a

contextualised analysis of cell type similarity 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) 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 GeneSpectra/GeneSpectraKG framework on well-studied datasets with aligned cell types. By performing gene classification, we show different conservation patterns of orthologs with different specificity and distribution classes. We utilise the gene classes to explain species effect [3] 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-O2O orthologs and species-unique cell type specific genes.

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. The GeneSpectra module conducts a robust cross-species gene classification based on cell type expression, by reducing the sparsity of scRNA-seq data and applying gene classification after normalisation. The GeneSpectraKG module utilises GeneSpectra's gene classes, integrates orthology annotations from the EggNOG database [33], and generates a KG to effectively represent the results and enable further analysis. Figure 1 summarises the key steps involved in this pipeline.

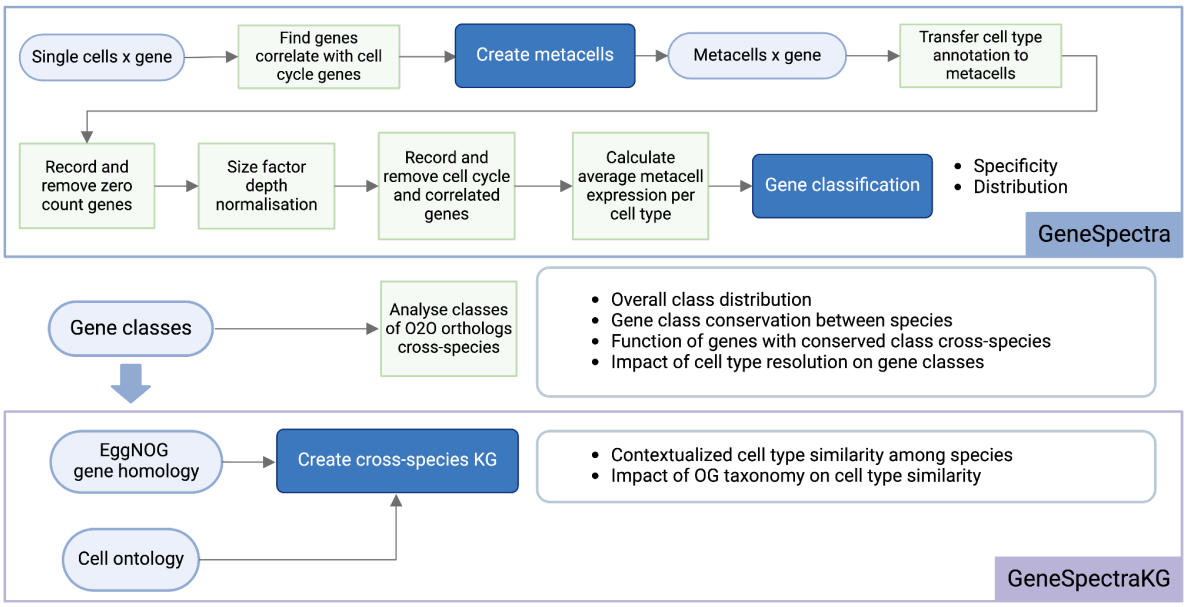

Figure 1 Schematic of the GeneSpectra and GeneSpectraKG framework. O2O, one-to-one; OG, ortholog groups; KG, knowledge graph.

Since scRNA-seq experiments only capture a small percentage of transcripts per cell [34], 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 [35,36] (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 Human Protein Atlas classification [22] 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 5 mammals [6] for closely related species and an embryogenesis dataset from zebrafish and xenopus [37] representing distantly related species (see Methods).

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).

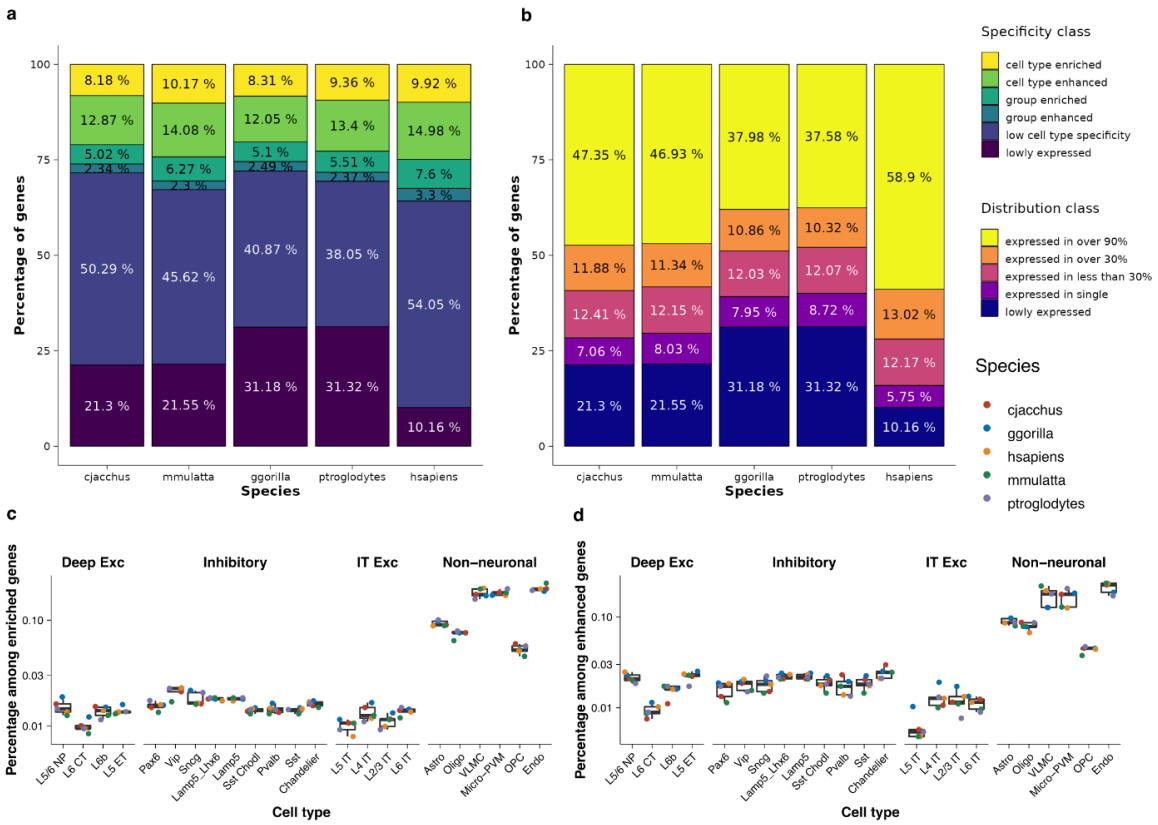

Figure 2 Gene class distribution in the MTG datasets. (a) Gene specificity class and (b) gene distribution class percentage of all genes in each species. (c) percentage of enriched genes and (d) enhanced genes in each cell type. Data points were shown as dots. The bar in the boxplot shows the arithmetic 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 O2O orthologs

Since O2O 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 O2O 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.

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 overlap for O2O orthologs in pairs of species for the MTG dataset (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 [38]).

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 O2O 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 O2O 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.

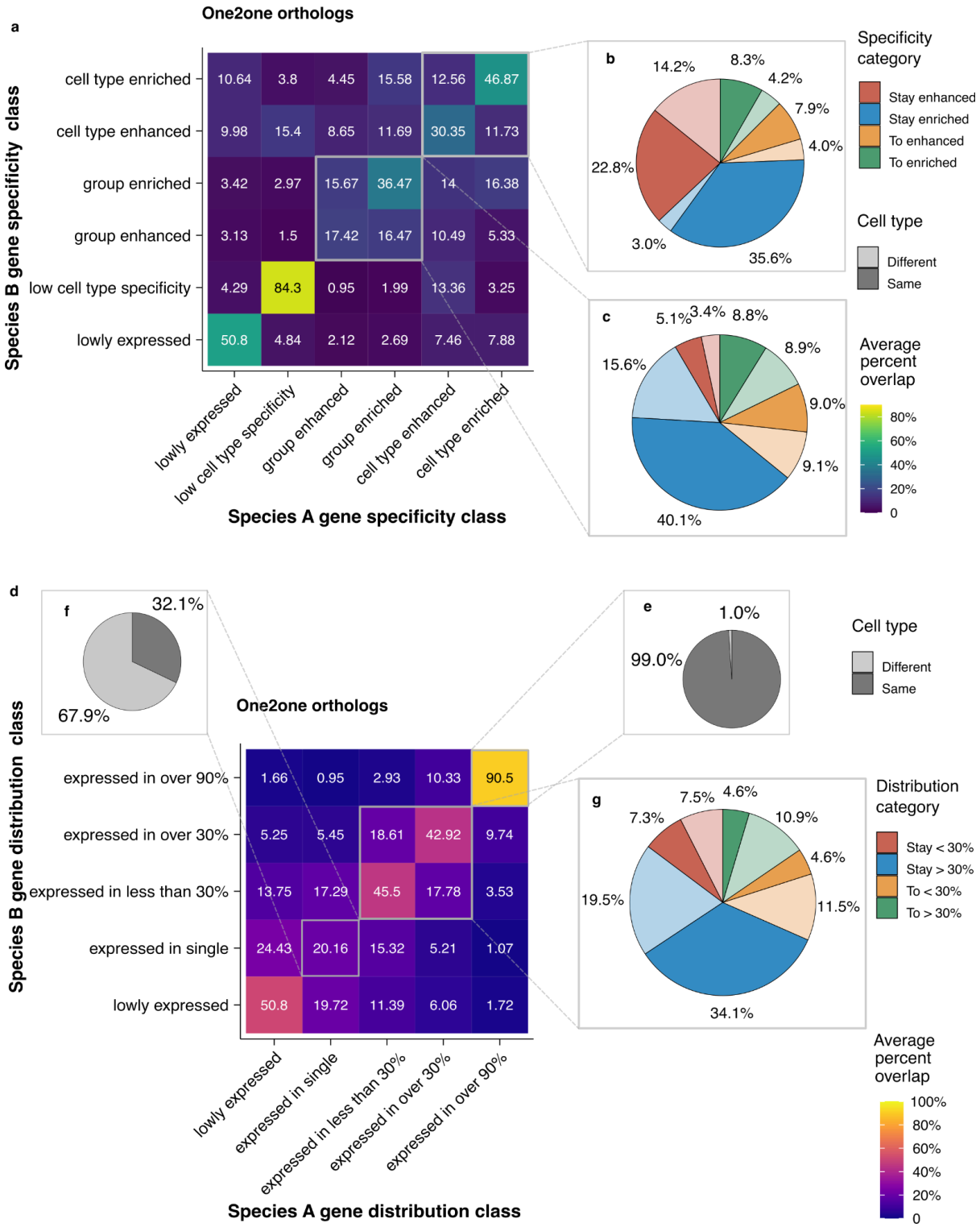

Figure 3 Gene class conservation among O2O orthologs in MTG dataset. To give an overview of the class conservation of O2O genes, we averaged the gene class conservation matrices among 10 pairs of species. The species were paired in a directional way, from species A to species B, with species B closer 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 well-matched annotation among the 5 species due to conscious sampling and data

processing [30]. We reckon that 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 is an effective tool for assessing such comparability. Using GeneSpectra, we generated gene class conservation matrices from independently annotated human and mouse bone marrow datasets [39]. 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 species effects [3]: cells show higher expression similarity within species, rather than within matching cell types, on the shared transcriptomic space built with O2O orthologs. We previously compared different computational strategies to correct this species effect and found it challenging to distinguish the source of such divergence [3]. 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 O2O 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 conserved element in transcriptomic profile between the species. Conversely, we also obtained a list of O2O 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 O2O 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. Note that we did not perform any integration of the raw counts data.

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 O2O 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 [40] (Figure 4c) and observed aligning trends with the dimensional reduction analysis.

We conclude that the orthologs 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.

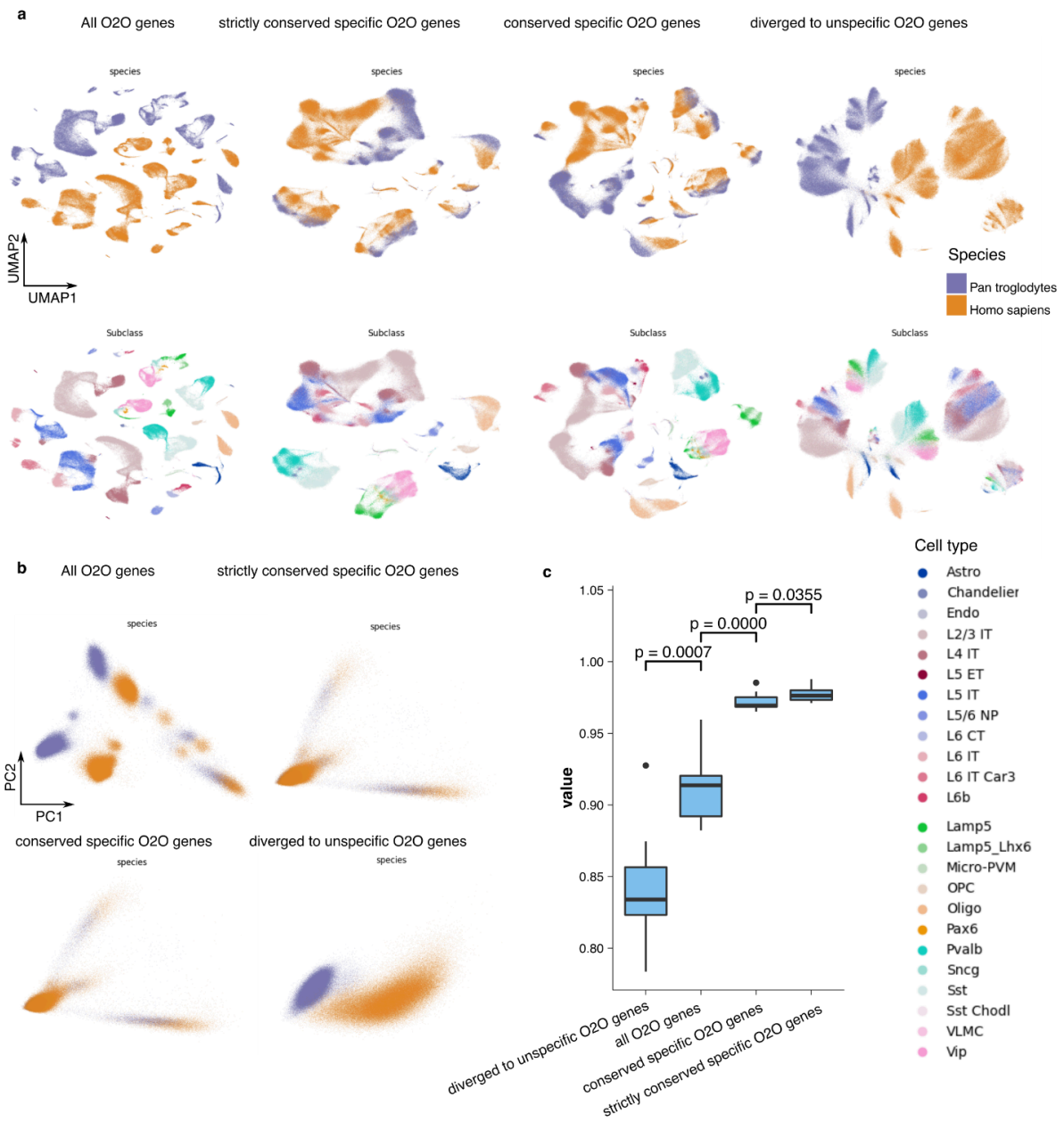

Figure 4 Gene class switching explains species effect on transcriptomic space. (a) UMAP visualisation using different sets of O2O orthologs for the human and chimpanzee data. Showing the strong species effect using all genes, the reduced species effect using (strictly) conserved specific genes, as well as a purified specie effect shown in diverged to unspecific genes (b) The first 2 PCs coloured by species using different sets of O2O 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). O2O, 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 [41] 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-O2O orthologs in general have lower sequence conservation than O2O orthologs (bottom right triangle in Figure 5), but we saw no clear bias in expression specificity similarity. Many non-O2O orthologs with a range of sequence conservation showed high expression conservation. Our results agree with the observations

by Stamboulian, M. et al. [16] that the types of homologs are irrelevant to function prediction tasks, highlighting the importance of including non-O2O orthologs into functional analysis.

In a nutshell, our results suggest that the ortholog conjecture mostly holds true in the MTG dataset, particularly between closely related species. Nevertheless, a noteworthy amount of genes did not follow this convention, thereby requiring caution in expression knowledge transfer.

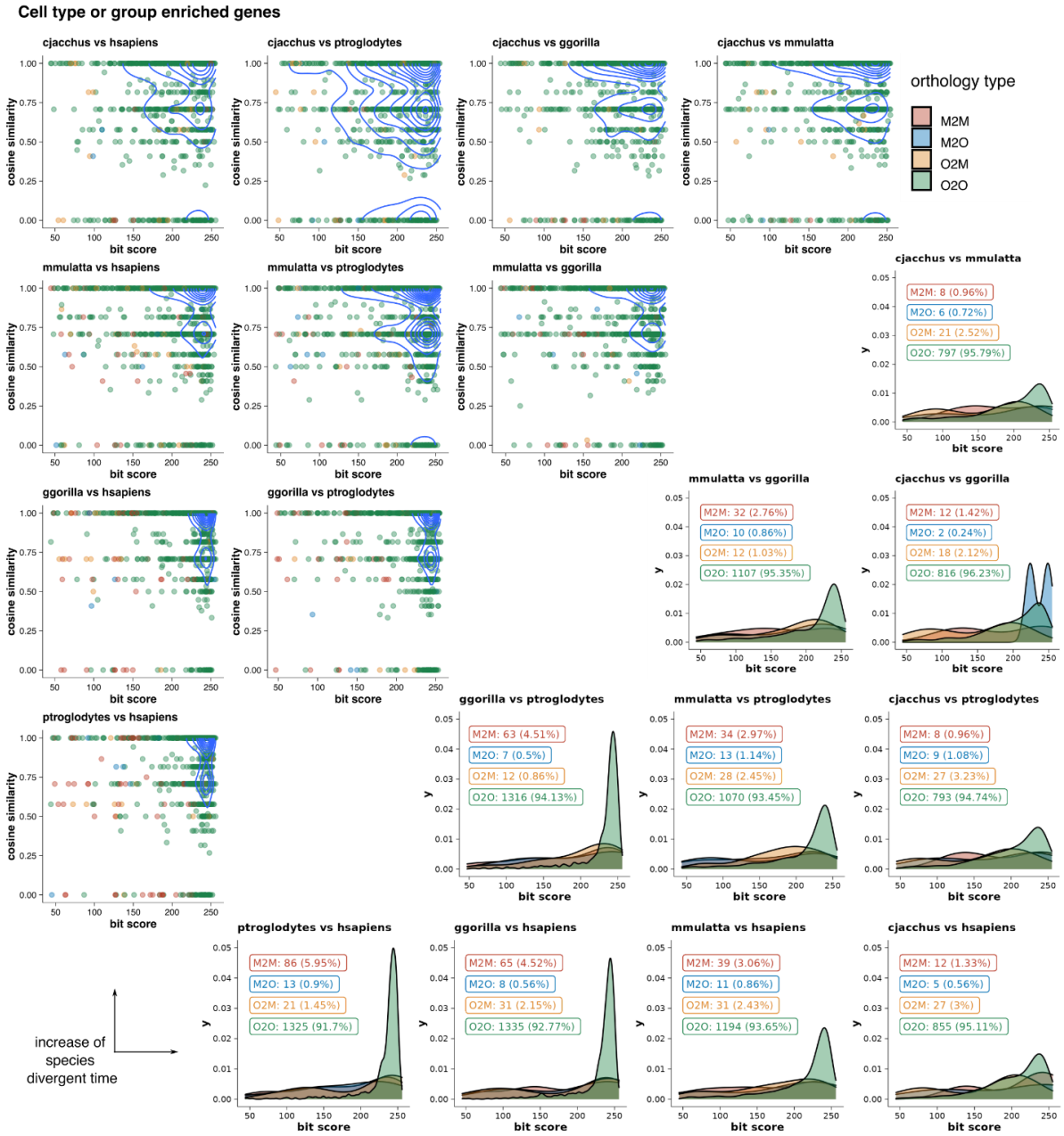

Figure 5 Comparing sequence conservation with cell type expression specificity specificity similarity across species distance for enriched genes. The top left triangle plots the

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. O2O: one-to-one, O2M: one-to-many, M2O: many-to-one, M2M: 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-O2O orthologs, encompass an unlimited number of species in a single analysis and consider the taxonomy resolution on which orthologs were computed.

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 in two cell types (Figure 6b). 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 [6]. Performing similar analysis on the reduced granularity data showed similar patterns, only with Endo, Micro_PVM and VLMC clustered mostly by species instead. Finally, with the increase of species divergence time, there was a general decrease in Jaccard-derived similarity for each cell type (Supplementary Figure 13).

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.

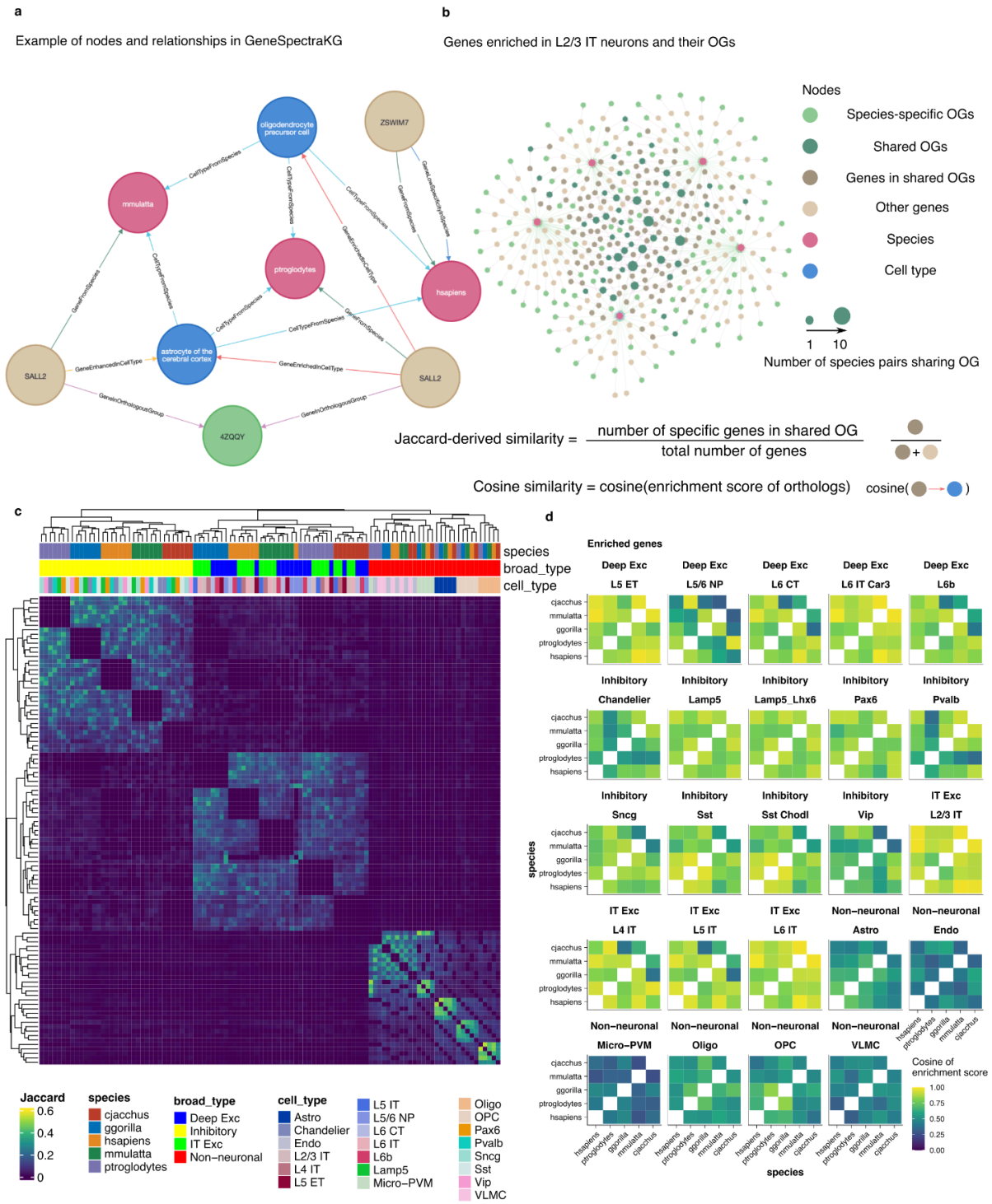

Figure 6 Contextualised analysis of cell type similarity between species using GeneSpectraKG. (a) An example showing all types of nodes and relationships in the GeneSpectraKG. (b) An example query showing all the genes that enrich in the L2/3 IT neurons from all species. Taken from a screenshot of the 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 [33,42]. 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.

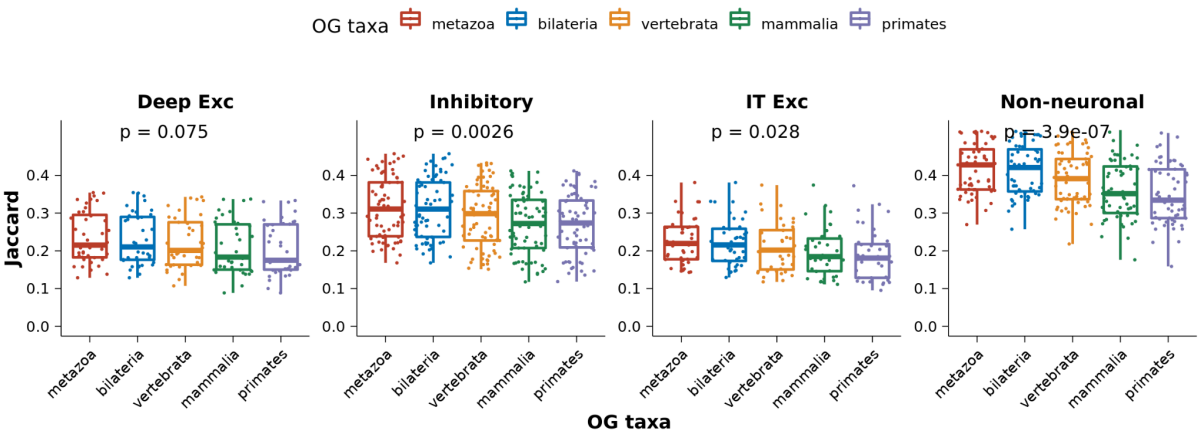

Figure 7 Impact of orthologous group taxon on cell type Jaccard-derived similarity. Each dot corresponds to the Jaccard-derived similarity score of the same cell type in one pair of species using OGs calculated at a certain taxonomy level, grouped by their broad type. The box plot shows the Jaccard-derived 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. This framework represents a conceptual advancement by adopting a holistic approach, moving beyond reductionist analyses that rely on strong assumptions. Specifically, it enables a dataset-specific examination of the ortholog conjecture; allows traversal analysis of different cell type resolutions and OG taxonomy levels; and takes different angles to analyse cell type similarity while inherently includes in-paralogs and species-specific genes. We further demonstrate that the gene classes can explain the species effect observed on cross-species scRNA-seq datasets, and analyse how the agreement between sequence conservation and expression specificity similarity vary for different species divergence time.

We found that genes with low cell type specificity and broad expression generally maintain such expression patterns between species, while genes showing some degree of cell type specificity or expression selectivity significantly diverge. We also showed a non-trivial percentage of genes that stayed specific switched cell type between species. Understanding the expression diversification of orthologs will not only be informative as is, but also provides a refined basis for developing new approaches for cross-species integration analysis. One possibility is to utilise the knowledge about which orthologs have more conserved specificity and distribution class between species as a prior for developing novel algorithms for cross-species alignment of evolutionary related cell types.

The cell type similarity analysis reveals that certain cell types have accumulated more species-specific expression signals, 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). This finding echoes the analysis by Liang et al., [43] 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 [43]. 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 the community begins to explore the evolutionary relationships of cell types between distantly related species. As more cases of gene emergence and loss exist between evolutionarily distant species, the importance of cell type-specific expression signals from species-specific genes will become more pronounced. Our KG-based cell type similarity measurements can effectively account for the varying numbers of non-O2O orthologs and species specific genes. Furthermore, the efficient database system provides flexibility and scalability that is crucial for atlas-level multi-species comparisons.

There are still several 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. Additionally, many cross-species expression differences could be evolutionarily neutral. The addition of gene regulation data would be helpful to disentangle the signals that are relevant to species adaptation.

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 [44]. Author cell ontology annotations were obtained from Cellxgene of the same dataset [45]. 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 [33]. 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.

Constructing metacells

Metacells partition the raw single cell data into disjoint, homogeneous groups ideally representing re-sampling from the same cellular state [36]. 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 O2O orthologs for each pair of species was calculated as follows: we first matched the O2O orthologs using ENSEMBL and counted the number of O2O orthologs in each class for both species. We then counted the intersection of O2O orthologs in all class combinations between the pair of species, and calculated the percentages respective to the total number of genes in that class for each species. 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 O2O 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 [46] (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.12209577. 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 [22] 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	Categories	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

1. Arendt, D. *et al.* The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).
2. Marioni, J. C. & Arendt, D. How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. *Annu. Rev. Cell Dev. Biol.* **33**, 537–553 (2017).
3. Song, Y., Miao, Z., Brazma, A. & Papatheodorou, I. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nat. Commun.* **14**, 6495 (2023).
4. Tanay, A. & Sebé-Pedrós, A. Evolutionary cell type mapping with single-cell genomics. *Trends Genet.* **37**, 919–932 (2021).
5. Shafer, M. E. R. Cross-Species Analysis of Single-Cell Transcriptomic Data. *Frontiers in Cell and Developmental Biology* **7**, (2019).
6. Jorstad, N. L. *et al.* Comparative transcriptomics reveals human-specific cortical features. *Science* **382**, eade9516 (2023).
7. Najle, S. R. *et al.* Stepwise emergence of the neuronal gene expression program in early animal evolution. *Cell* **186**, 4676–4693.e29 (2023).
8. Shafer, M. E. R., Sawh, A. N. & Schier, A. F. Gene family evolution underlies cell-type diversification in the hypothalamus of teleosts. *Nat Ecol Evol* (2021)

doi:10.1038/s41559-021-01580-3.

9. Geirsdottir, L. *et al.* Cross-Species Single-Cell Analysis Reveals Divergence of the Primate Microglia Program. *Cell* **179**, 1609–1622.e16 (2019).
10. Sebé-Pedrós, A. *et al.* Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* **173**, 1520–1534.e20 (2018).
11. Franjic, D. *et al.* Transcriptomic taxonomy and neurogenic trajectories of adult human, macaque, and pig hippocampal and entorhinal cells. *Neuron* **110**, 452–469.e14 (2022).
12. Tritschler, S. *et al.* A transcriptional cross species map of pancreatic islet cells. *Mol Metab* **66**, 101595 (2022).
13. Liu, T. *et al.* Cross-species single-cell transcriptomic analysis reveals pre-gastrulation developmental differences among pigs, monkeys, and humans. *Cell Discovery* **7**, 1–17 (2021).
14. Sonnhammer, E. L. L. *et al.* Big data and other challenges in the quest for orthologs. *Bioinformatics* **30**, 2993–2998 (2014).
15. Nehrt, N. L., Clark, W. T., Radivojac, P. & Hahn, M. W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.* **7**, e1002073 (2011).
16. Stamboulian, M., Guerrero, R. F., Hahn, M. W. & Radivojac, P. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics* **36**, i219–i226 (2020).
17. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. *PLoS Comput. Biol.* **12**, e1005274 (2016).
18. Rogozin, I. B., Managadze, D., Shabalina, S. A. & Koonin, E. V. Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biol. Evol.* **6**, 754–762 (2014).

19. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360–366 (2013).
20. Chen, X. & Zhang, J. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput. Biol.* **8**, e1002784 (2012).
21. Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. & Dessimoz, C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.* **8**, e1002514 (2012).
22. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Science Advances* vol. 7 Preprint at https://doi.org/10.1126/sciadv.abh2169 (2021).
23. Rosen, Y. *et al.* Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nat. Methods* (2024) doi:10.1038/s41592-024-02191-z.
24. Biharie, K., Michielsen, L., Reinders, M. J. T. & Mahfouz, A. Cell type matching across species using protein embeddings and transfer learning. *Bioinformatics* **39**, i404–i412 (2023).
25. Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10**, (2021).
26. Park, Y., Muttray, N. P. & Hauschild, A.-C. Species-agnostic transfer learning for cross-species transcriptomics data integration without gene orthology. *Brief. Bioinform.* **25**, (2024).
27. Domcke, S. & Shendure, J. A reference cell tree will serve science better than a reference cell atlas. *Cell* **186**, 1103–1114 (2023).
28. van Zyl, T. *et al.* Cell atlas of aqueous humor outflow pathways in eyes of humans and four model species provides insight into glaucoma pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 10339–10349 (2020).
29. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the*

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (Association for Computing Machinery, New York, NY, USA, 2016).

30. Suresh, H. *et al.* Comparative single-cell transcriptomic analysis of primate brains highlights human-specific regulatory evolution. *Nature Ecology & Evolution* **7**, 1930–1943 (2023).
31. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).
32. Evolution at the cellular level. *Nat Ecol Evol* **7**, 1155–1156 (2023).
33. Hernández-Plaza, A. *et al.* eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* **51**, D389–D394 (2023).
34. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
35. Baran, Y. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).
36. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. *Genome Biol.* **23**, 100 (2022).
37. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).
38. Mantica, F. *et al.* Evolution of tissue-specific expression of ancestral genes across vertebrates and insects. *Nature Ecology & Evolution* 1–14 (2024).
39. Stumpf, P. S. *et al.* Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell RNA sequencing. *Commun Biol* **3**, 736 (2020).
40. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* (2021) doi:10.1038/s41592-021-01336-8.
41. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment

search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

42. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
43. Liang, C., Musser, J. M., Cloutier, A., Prum, R. O. & Wagner, G. P. Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes. *Genome Biol. Evol.* **10**, 538–552 (2018).
44. Index of /publication_release/Great_Ape_MTG_Analysis. https://data.nemoarchive.org/publication_release/Great_Ape_MTG_Analysis/.
45. Cellxgene data for Comparative transcriptomics reveals human-specific cortical features. https://cellxgene.cziscience.com/collections/4dca242c-d302-4dba-a68f-4c61e7bad553 .
46. Lobentanzer, S. *et al.* Democratizing knowledge representation with BioCypher. *Nat. Biotechnol.* **41**, 1056–1059 (2023).