nature ecology & evolution

Article

https://doi.org/10.1038/s41559-023-02281-9

Cell type evolution reconstruction across species through cell phylogenies of single-cell RNA sequencing data

Received: 6 June 2023

Accepted: 16 November 2023

Published online: 5 January 2024

Check for updates

Jasmine L. Mah & Casey W. Dunn

The origin and evolution of cell types has emerged as a key topic in evolutionary biology. Driven by rapidly accumulating single-cell datasets, recent attempts to infer cell type evolution have largely been limited to pairwise comparisons because we lack approaches to build cell phylogenies using model-based approaches. Here we approach the challenges of applying explicit phylogenetic methods to single-cell data by using principal components as phylogenetic characters. We infer a cell phylogeny from a large, comparative single-cell dataset of eye cells from five distantly related mammals. Robust cell type clades enable us to provide a phylogenetic, rather than phenetic, definition of cell type, allowing us to forgo marker genes and phylogenetically classify cells by topology. We further observe evolutionary relationships between diverse vessel endothelia and identify the myelinating and non-myelinating Schwann cells as sister cell types. Finally, we examine principal component loadings and describe the gene expression dynamics underlying the function and identity of cell type clades that have been conserved across the five species. A cell phylogeny provides a rigorous framework towards investigating the evolutionary history of cells and will be critical to interpret comparative single-cell datasets that aim to ask fundamental evolutionary questions.

Trees are a prevalent pattern in cell biology. Three distinct types of cell tree are often discussed. These are the cell lineage[1–3], the pattern of phenotypic similarity between cells[4–6] and the cell phylogeny[4,7–16]. Often, these trees are conflated because the implicit null expectation is congruence. Yet the most compelling evolutionary and developmental cell biology can defy null expectations and lead to incongruent topologies[17,18]. To explore this exciting biology, we need approaches to independently describe each of these distinct types of topology.

Cell phylogenies are the least well-explored of the three trees. Cells are evolutionary units and homologous cell types can be readily identified across organisms[12]. A cell phylogeny depicts the evolutionary history of cells: extant cells are placed at tip nodes, and interior nodes represent cells present in ancestral organisms. Much like in a gene tree, new cell types emerge by divergences that may occur between cell types or due to speciation[15]. This contrasts with a developmental lineage, where tip nodes are differentiated adult cells and interior nodes are precursor cells that divided during embryogenesis. Branching in a lineage depicts cell division.

Cell lineages are described by well-established methods for tracking cell division[1,3,19,20] and have been extensively studied in the context of embryology[1–3] and cancer[21–23]. Phenotypic similarity trees are likewise well-studied, as embodied in phenetic Linnaean-like cell taxonomies[4–6,24]. However, while adjacent methods have been pursued[3,8,25–27], there is far less work on building cell phylogenies using explicit evolutionary models[7,8,10]. Neighbour joining (NJ)[28], a hierarchical clustering algorithm, has been used to construct trees from cell phenotype data[9,22,29–31]. These trees describe distances between phenotypes and have been used as estimates of evolutionary history in the absence of

Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. e-mail: jasmine.mah@yale.edu

other approaches. However, while evolutionary conservation may lead to phenotypic similarity between closely related cell types, NJ trees do not explicitly model evolution.

There are multiple motivations for developing robust approaches to constructing cell phylogenies. Single-cell studies pursuing evolutionary questions across species have gained prominence, but without approaches to build a cell phylogeny most employ pairwise comparisons instead[32–36]. Pairwise comparisons, however, can be misleading when phylogenetic structure is present[37–39]. A cell phylogeny provides a concrete framework to study fundamental evolutionary questions, such as the origins of multicellularity[40], and will also establish clear biological criteria for defining cell types[4,11,12,15], just as phylogenetics has transformed the taxonomic organization of species. Importantly, a cell phylogeny allows us to directly apply phylogenetic comparative tools to long-held questions in cell biology[9,26]. We can, for instance, investigate the homology between the choanoflagellate and choanocyte collars[41,42] with an ancestral character state reconstruction.

Here we provide an approach to building cell phylogenies that applies an explicit evolutionary model to single-cell RNA sequencing (RNA-seq) data from multiple species. We use principal components (PCs) as phylogenetic characters and observe the phylogenetic signal in earlier components. We identify robust clades that allow us to assign a phylogenetic definition of cell type to a dataset of differentiated eye cells. PC loadings outline gene expression dynamics that describe the function and identity of cell type clades in the cell phylogeny.

Results

While single-cell count data may present a rich source of evolutionary information, it comes with challenges that must be addressed before phylogenetic methods can be applied. These challenges include normalization[43], integration[44,45] across samples, co-expression, noise and high multidimensionality. However, while tools designed to approach these challenges have been extensively developed in the single-cell field, a disconnect exists in that most phylogenetic models are not built for continuous counts, but for discrete characters. Our approach is to use Brownian motion, an evolutionary model commonly applied to continuous characters like morphology, to explicitly model the evolution of counts.

The first PCs have phylogenetic signal

We use single-cell RNA-seq (scRNA-seq) counts published by van Zyl et al.[33] as our focal dataset. These counts were obtained from cells of the aqueous humour and surrounding structures of the eye from five model organisms (*Homo sapiens*, *Macaca mulatta*, *Macaca fascicularis*, *Mus musculus* and *Sus scrofa*[33]; Extended Data Fig. 1, step 1). Strong interest and potential clinical implications[33] emerging from the investigation of the evolution of these cell types make this high-quality dataset an ideal candidate for phylogenetic analysis.

van Zyl et al.'s[33] analysis of this dataset produced 92 annotated cell clusters across all five species. In our analysis, we associate each cluster with its cell type label and call it a 'cell type group'. We used standard single-cell workflows[43–45] to normalize and integrate counts separately for each species and subsampled cells (Extended Data Fig. 1, steps 1–4). Matrices were then combined by gene homology to perform a cross-species integration (Extended Data Fig. 1, steps 5–6). The combined and cross-species-integrated matrix consisted of 919 cells and 2,000 genes (Extended Data Fig. 1, step 6). This matrix represented the 92 cell type groups across five species, each with 15–20 cell type groups, for which ~10 cells were sampled per cell type group. Following cross-species integration, a UMAP[46] revealed that cells clustered by cell type rather than species, consistent with successful cross-species integration (Extended Data Fig. 2).

Gene expression is noisy, correlated and highly multidimensional. To address the feature independence assumption of the uncorrelated Brownian motion model, we performed a principal component analysis (PCA) of this matrix (Extended Data Fig. 1, step 7) and used the PCs of gene expression as phylogenetic characters. This produced a 919-cell matrix with 919 PCs (Extended Data Fig. 1, step 7). PCs are orthogonal, and both dimensionality and noise can be reduced by trimming later PCs if the data is low-rank. Plotting the amount of variance described per PC revealed a clear elbow at PC 20 (Fig. 1). Approximately 27% of the variance is explained by the first 20 PCs, while there was negligible change in the amount of variance explained by the remaining 899 PCs (Fig. 1). This demonstrates the data is low-rank.

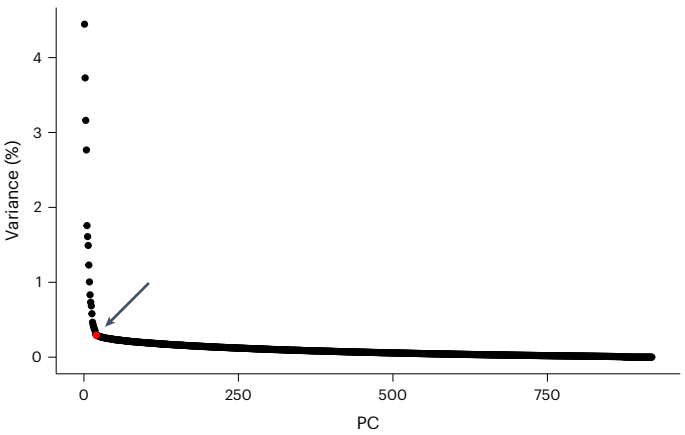


Fig. 1 | Percentage of variance described per PC. A total of 919 PCs were created by PCA of a 919-cell, 2,000-gene matrix. Approximately 27% of the variance resides in the first 20 PCs, which forms a distinct elbow in the plot (arrow). PC 20 is highlighted in red.

We used a Brownian motion model of continuous character evolution[47] to infer cell phylogenies from PCs derived from gene expression levels (Methods). We inferred unrooted phylogenies, where each tip is a cell type group represented by a single cell and branch lengths indicate the amount of expected evolutionary change in expression. We examined the impact of rank on the cell phylogeny by inferring trees from matrices with an increasing number of PCs (Fig. 2). Trends in tree length, tip versus internal edge length and star-ness of the phylogenies were observed. The plots of these trends exhibited elbows after PC 20 (Fig. 2), corresponding to the elbow in the PC variance plot (Fig. 1).

Total variance can be measured as tree length. As PCs are sorted in descending order of variance (Fig. 1), tree length increased most rapidly as early PCs were added (PCs 3–25), while growth in tree length slowed as later PCs were included (Fig. 2a).

We characterize the added variance by inspecting the behaviour of tip edge lengths and interior edge lengths as the number of PCs was increased (Fig. 2b). Phylogenetic signal, the variance that is shared among tips due to common ancestry, is represented as interior edge lengths in a phylogeny. In contrast, cell-specific variance, which may include cell-specific evolution, cell state and observational noise, is reflected in tip edge length. We hypothesized that much of this cell-specific variance is noise, and if so, is present as the higher-frequency signal that resides in later PCs. We found that tip edge length increased while interior edge lengths remained relatively constant as the number of PCs increased (Fig. 2b), suggesting that later PCs describe high-frequency cell-specific signal.

Finally, we examined how adding higher-frequency signal affected the phylogenetic signal of the phylogeny (Fig. 2c). Star phylogenies, which have long tip edges and negligible interior edge lengths, possess little to no phylogenetic signal. We find that by adding later PCs, the phylogeny became more star-like (Fig. 2c). This suggests that adding higher-frequency variance from later PCs drowns out the phylogenetic signal.

Clustering of cells into cell type clades improved substantially when the number of PCs used to infer the phylogeny was reduced (Fig. 3).

https://doi.org/10.1038/s41559-023-02281-9

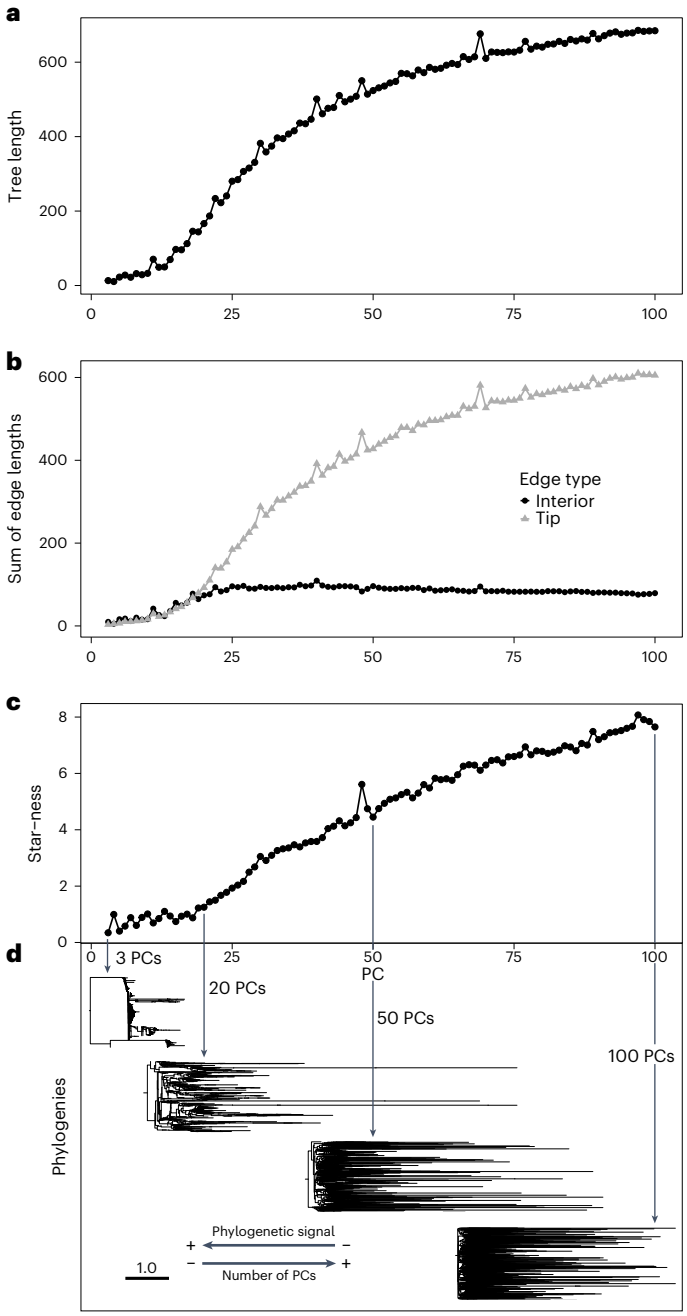


Fig. 2 | The inclusion of more PCs in the character matrix leads to a prevalence of noise, obscuring the phylogenetic signal. a–**c**, As the number of PCs increases, tree length increases (**a**), tip edge length increases while interior edge lengths remain constant (**b**) and the tree becomes more star-like (**c**). **d**, Examples of trees inferred from 3, 20, 50 and 100 PCs. The scale bar in **d** represents units of expected evolutionary change.

Clear grouping by cell type clade could be seen in the tree made from 20 PCs (Fig. 3a). A phylogeny inferred from 500 PCs exhibited poor grouping (Fig. 3b), while one inferred from all 919 PCs failed to produce cell type clades (Fig. 3c).

Identification of phylogenetic cell type clades

Based on the above trends, we inferred an unrooted branch-length phylogeny of 54 cells (Fig. 4), which was then midpoint-rooted. This phylogeny represented 54 cell type groups (1 cell per cell type group) from five species, each of which possessed 15–20 cell type groups (Fig. 4). Analogous to removing rogue taxa from a species phylogeny, this subset was created by trimming away the minority of cell type groups that exhibited instability or were present in only a single species (Methods and Extended Data Fig. 3). We used the first 20 PCs calculated from gene expression as phylogenetic characters (Fig. 1). In this phylogeny, cells from across the five species clustered by cell type despite varying branch lengths, indicating the presence of a strong cell type signal that was not overwhelmed by species-level or cell-specific signal (Fig. 4). In addition, cells clustered by cell type regardless of the expression of key marker genes that have traditionally been used to identify particular cell types. For instance, the pig juxtacanalicular canal tissue (JCT) cell, which exhibits a long branch (Fig. 4), does not express *CHI3L1*, a key marker gene expressed in human, macaques and mouse JCT cells[33]. Similarly, Schlemm's canal cells from pig and mouse exhibit more dominant expression of lymphatic marker genes than in humans and macaques[33], but still formed a single well-supported Schlemm's canal cell clade (Fig. 4).

Just as phylogenies have directly informed species taxonomies, the cell phylogeny allows us to phylogenetically define cell type by clade membership (Fig. 4). These cell type clades included corneal endothelium, JCT, ciliary muscle, pericyte, myelinating and non-myelinating Schwann cells, melanocyte, and Schlemm's canal cells. As indicated by the node coloured with a black circle in Fig. 4, myelinating and non-myelinating Schwann cells were sister to each other, forming a Schwann cell superclade (Fig. 4). While macrophages and natural killer T cells did not perfectly segregate into separate clades, they formed a single robustly supported superclade of immune cells (Fig. 4). Finally, a clade of vascular endothelia and collector channel cells were present as a sister group to Schlemm's canal cells (Fig. 4). We label the superclade encompassing these three cell types as the 'vessel endothelia' clade, following van Zyl et al.[33]. Ciliary muscle cells, pericytes, and the clade encompassing Schwann cells, melanocytes, vessel endothelia and immune cells formed a polytomy (Fig. 4).

Robustness of the cell phylogeny

We next aimed to measure the repeatability of the topology of the cell phylogeny. Because our phylogenetic characters are not weighted equally, bootstrap scores were not appropriate for our data[48] (Methods). Instead, we assessed technical repeatability with 'jumble scores' that summarize consistency of relationships given different addition orders of tips. Jumble scores are calculated from the clade frequencies of a set of trees ('jumble trees') generated by repeated maximum likelihood searches, each with a different starting tree. Like bootstrap scores, jumble scores nearing 100% indicate that the split is highly repeatable across the set of jumble trees. We found high support for splits that defined cell type clades, with most scores greater than 95% (Fig. 4). The subclades of the Schwann cell and vessel endothelia superclades were also well-supported, giving validity to their status of being closely related in a superclade relationship (Fig. 4). Several of the deeper splits along the backbone displayed much more variation across the jumble trees, leading to scores below 70% (Fig. 4). Because of this, broader relationships between clades at these splits generally remain poorly characterized, even as the cell type clades themselves are robustly supported (Fig. 4).

Given that inference was performed on a small subset of cells from the full matrix and that gene expression may vary even across cells of the same cell type, we also assessed the effect of cell subsampling on topological stability. Unlike most phylogenetic datasets, we have a large pool of biological replicates, allowing us to perform a single-cell jackknife (scjackknife) procedure. The cells at the tips of the 54-cell phylogeny (Fig. 4) were repeatedly randomly resampled from a larger dataset to create 439 scjackknife trees that varied in the identity of the sampled cells. Distinct, reproducible cell type clades emerged, and their repeatability was best described by measures that recognized heterogeneous degrees of clade presence (Fig. 4 and Extended Data Fig. 4). Calculating the transfer bootstrap expectation[49] score for clades

https://doi.org/10.1038/s41559-023-02281-9

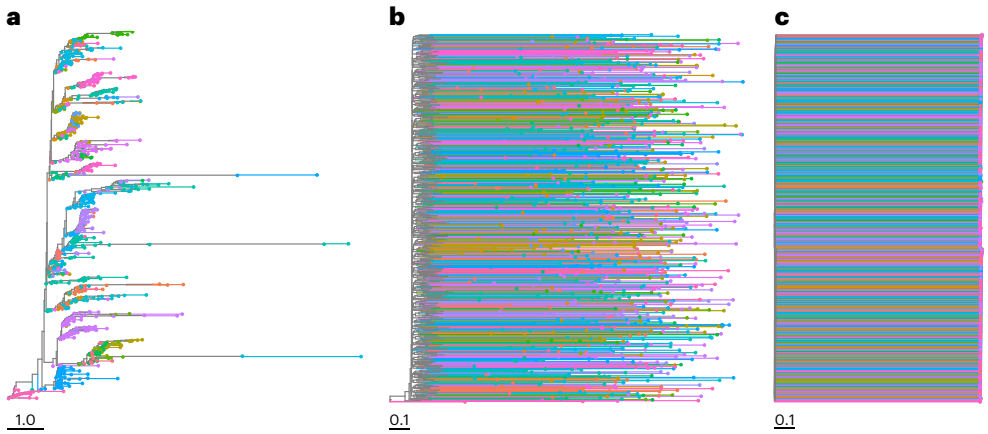


Fig. 3 | The formation of cell type clades improves when the number of PCs is reduced. a–**c**, Cell phylogenies of 919 cells inferred from 20 PCs (**a**), 500 PCs (**b**) and all 919 PCs (**c**). Improved clustering by cell type and increased internal structure can be seen in the 20-PC tree, which is subsequently reduced in the 500-PC tree and finally lost in the 919-PC tree. All three trees were inferred using the same data and parameters. Tip branches are coloured according to cell type. Cell type colours are consistent across all three trees. Scale bars indicate the amount of expected evolutionary change.

in the cell phylogeny from this set of scjackknife phylogenies revealed consistent support (≥70) for the cell type clades that were robustly supported by jumble scores (Fig. 4).

Tip stability is clade-specific

We assessed the topological stability of tips in the cell phylogeny by calculating the leaf stability index[50] (LSI) from the jumble trees (Fig. 4 and Extended Data Fig. 5). This index ranges from 0 (very low stability) to 1 (very high stability). We found that LSI varied closely by cell type clade and was highly concordant within cell type clades, suggesting that instability is cell type specific and describes the stability of the clade as a unit (Fig. 4 and Extended Data Fig. 5). The most stable cell type was the Schlemm's canal cells whereas the least stable was the melanocytes (Fig. 4 and Extended Data Fig. 5).

Averaging replicate cells stabilizes clade relationships

We inferred cell phylogenies where each tip indicates a cell type group represented as the average of multiple replicate cells, rather than as a sample of a single cell (Fig. 5). Almost all jumble scores in the resulting phylogeny were 100 and most scjackknife scores fell above 70, even amongst deep nodes. This increased topological repeatability was accompanied by increased tip stability (Extended Data Fig. 6). While LSI values remain highly concordant within cell type clades (Extended Data Fig. 6), differences between clades became less apparent as most indices approached the maximum score of 1 (Extended Data Fig. 6). New relationships within cell type clades also emerged with high repeatability. For instance, the collector channel cells now formed a clade nested within the vascular endothelia, with high jumble and scjackknife support (Fig. 5). The JCT cells were paraphyletic, but with low scjackknife support (Fig. 5). In Fig. 4, the ciliary muscle and pericyte clades were in a polytomy, but here they emerged with high repeatability as sister clades (Fig. 5).

Highly loaded genes exhibit cell type signals

To identify the genes whose expression variance contributes most to phylogenetic signal, we examined the most highly positively or negatively loaded genes for each PC. The most significantly enriched gene ontology (GO) terms that emerged from the most highly loaded genes revealed the presence of consistent cell type signals described by certain PCs (Table 1 and Supplementary Table 1). Particular cell type clades, such as the vessel endothelial clade, are reflected in GO terms that emerged from the most positively loaded genes (Supplementary Table 1). For instance, the positive loadings for PC 2 were dominated by GO terms describing vascular development (for example, 'angiogenesis', GO:0001525; 'vasculature development', GO:0001944; and 'blood vessel development', GO:0001568; Supplementary Table 1). Validated marker genes[33] for the Schlemm's canal, collector channel and vascular endothelial cells (for example, *PECAM1*) were present among the top ten most positively loaded genes for PC 2 (Supplementary Tables 2 and 3). Other cell type signals were present among the negative loadings, suggesting that the phylogenetic identity of these cell type clades was defined by downregulation of their marker genes among other cells (Table 1 and Supplementary Table 1). One example is the melanocyte clade, described by the emergence of melanocyte-specific GO terms from the most negatively loaded genes of PC 4. These GO terms included 'pigmentation' (GO:0043473), 'developmental pigmentation' (GO:0048066) and 'melanocyte differentiation' (GO:0030318; Supplementary Table 1). The melanocyte master transcription factor MITF[33,51,52] was also present among the top ten most negatively loaded genes for PC 4 (Supplementary Tables 4 and 5).

Comparison to distance-based phenetic trees

In the absence of explicit phylogenetic approaches, phenetic trees built using distance-based methods have stood in as important proxies for the cell phylogeny[9,29]. We used the NJ[28], unweighted pair group method with arithmetic mean[53] (UPGMA) and weighted pair group method with arithmetic mean[54] (WPGMA) methods to build phenetic trees from the averaged expression matrix (Extended Data Fig. 1, step D1) used for Fig. 5 (Extended Data Figs. 7–9). We find that distance-based methods also produce cell type clades (Extended Data Figs. 7–9); however, some cells in the UPGMA and WPGMA trees failed to group with their corresponding cell type clade (Extended Data Figs. 8 and 9). In addition, the immune and Schwann cells superclades are paraphyletic in the UPGMA tree (Extended Data Fig. 8), while the vessel endothelia superclade is polyphyletic in the WPGMA tree (Extended Data Fig. 9). The UPGMA tree, and to a degree the WPGMA tree, exhibit a more ladder-like pattern, with cell type clades nested within each other, while the NJ and cell phylogeny display greater balance (Fig. 5 and Extended Data Figs. 7–9). To further compare the topologies of the NJ and cell phylogeny, we performed the scjackknife experiment, but using NJ to calculate the scjackknife trees (Extended Data Fig. 10). Mapping these scores onto the topology of the averaged cell phylogeny (Fig. 5) reveals that, while support varies within cell type clades, there is broad agreement in the emergence of cell type clades and the relationships between them, with the exception of the JCT clade (Extended Data Fig. 10). This is consistent with results derived from other types of molecular data, where NJ and maximum likelihood approaches produce identical or very similar trees[55,56].

https://doi.org/10.1038/s41559-023-02281-9

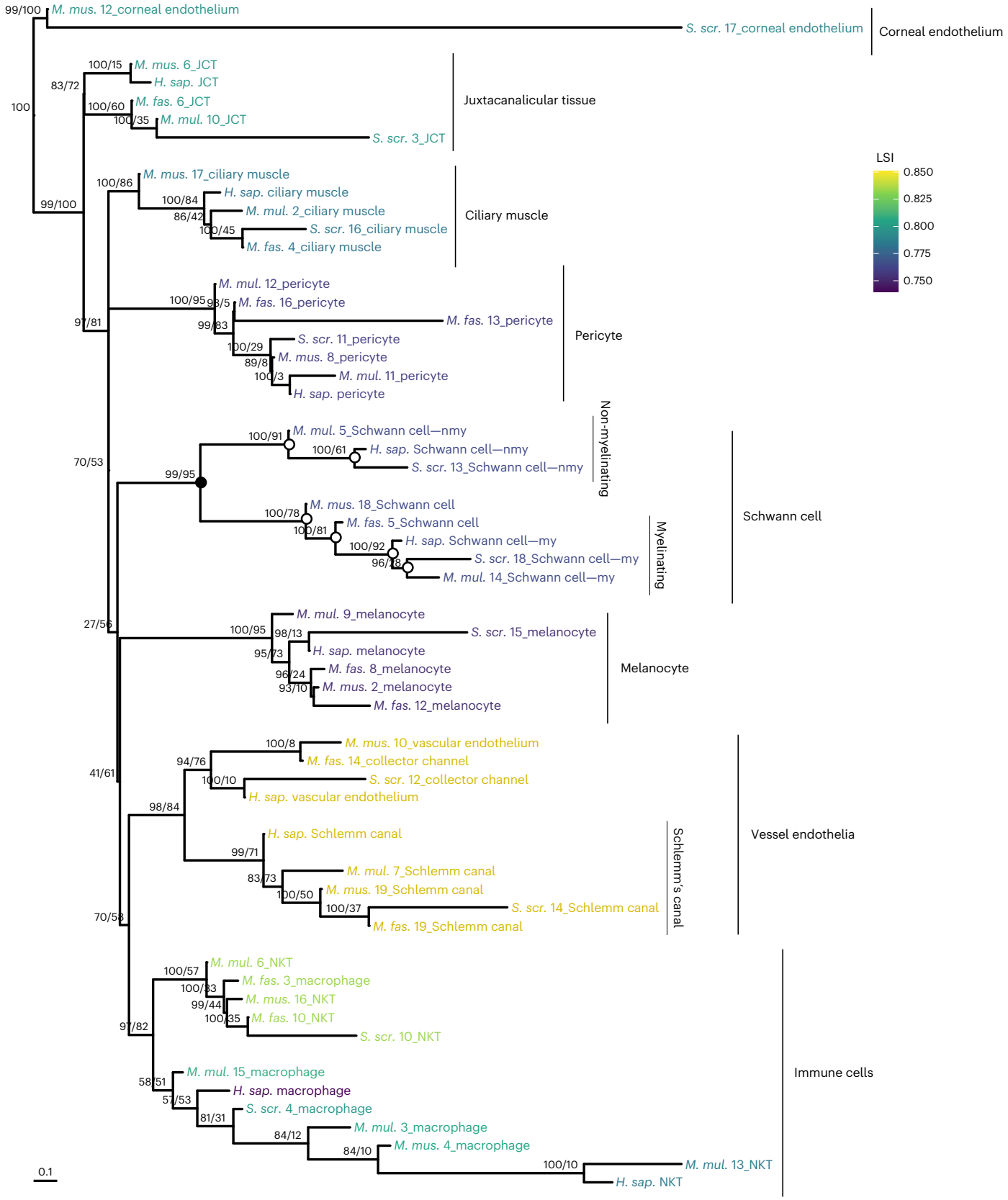


Fig. 4 | A cell phylogeny of aqueous humour cells. Fifty-four cells from five species cluster by cell type in the cell phylogeny. Species and cell type groups are labelled at the tips; numbers refer to the cluster number of the cell type group. Cell type clades are indicated by vertical bars. Superclades include the Schwann cell, immune cells and vessel endothelia clades. In the Schwann cell superclade, the black node is an example of a within-cell-type divergence, and the white nodes are examples of species-level divergences. Support values are printed as 'jumble score/scjackknife score'. The LSI is plotted as tip label colour. The scale bar indicates units of expected evolutionary change. *H. sap.*, *Homo sapiens*; *M. fas.*, *Macaca fascicularis*; *M. mul.*, *Macaca mulatta*; *M. mus.*, *Mus musculus*; *S. scr.*, *Sus scrofa*; my, myelinating; nmy, non-myelinating; NKT, natural killer T cell.

https://doi.org/10.1038/s41559-023-02281-9

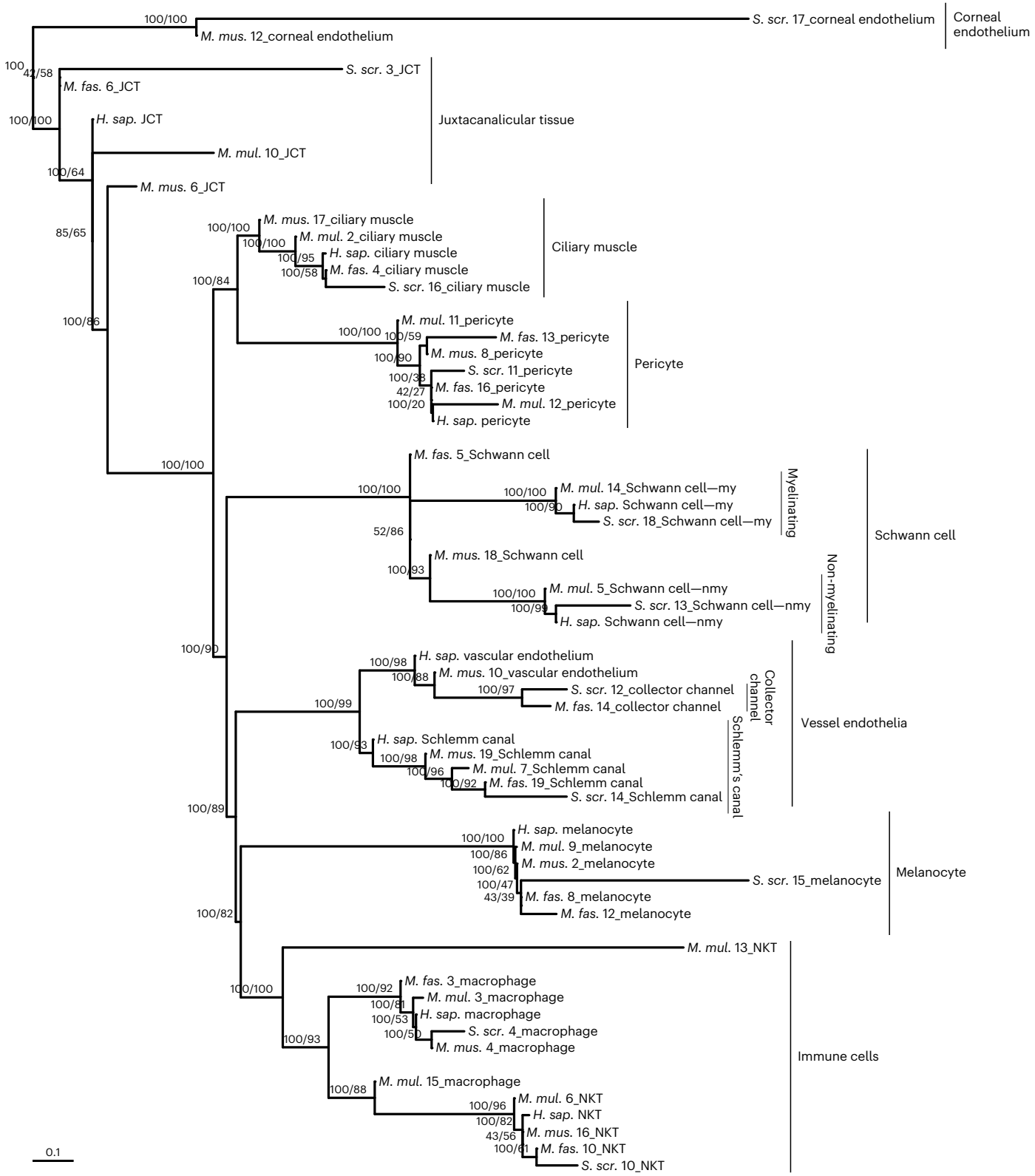


Fig. 5 | Averaging replicate cells stabilizes relationships between and within cell type clades. In this phylogeny, each tip corresponds to a cell type group represented by the average of ten cells. High jumble and scjackknife scores (plotted as 'jumble score/scjackknife score') characterize many nodes, including those along the backbone of the cell phylogeny. Cell type clades present in Fig. 4 are also present here and indicated with vertical bars. Species and cell type groups are labelled at the tips; numbers refer to the cluster number of the cell type group. The scale bar indicates units of expected evolutionary change. *H. sap.*, *Homo sapiens*; *M. fas.*, *Macaca fascicularis*; *M. mul.*, *Macaca mulatta*; *M. mus.*, *Mus musculus*; *S. scr.*, *Sus scrofa*; my, myelinating; nmy, non-myelinating; NKT, natural killer T cell.

Table 1 | Cell-type-specific signals are present among the most highly loaded genes in early PCs

PC	Positive loading	Negative loading
PC 1	Immune	Muscle
PC 2	Vessel endothelium	?
PC 3	Vessel endothelium	Muscle
PC 4	Fibroblast	Melanocyte
PC 5	Neural system	Immune
PC 6	?	?
PC 7	?	Neural system
PC 8	?	?
PC 9	?	Immune
PC 10	Vessel endothelium	?

Gene loadings for the first ten PCs were ordered by most positive or negative loading. GO terms enriched from the top 100 most highly loaded genes were related to the development, function and identity of specific cell type clades that emerged in the cell phylogeny. The correspondence between PC loading and cell type clade signal is provided. Loadings that had no immediately apparent cell type affinity are labelled '?'.

Discussion

Our findings suggest that the first PCs of single-cell gene expression data describe evolutionary variance across cells and that this variance can be used to construct cell phylogenies. Cells from distantly related species form well-supported cell type clades, which emerge only after removing the large number of low-variance PCs that describe fine, cell-specific differences. Liang et al.[8] was the first to statistically demonstrate tree-like evolutionary structure among cell transcriptomes. We aimed to build directly from that foundation by applying an explicit phylogenetic approach using an evolutionary model and examining cell type evolution across multiple species. Our ability to construct a robust cell phylogeny from single-cell expression levels is consistent with Liang et al.'s[8] finding that there is phylogenetic structure among cell transcriptomes[7–9].

The emergence of robust cell type clades in the cell phylogeny enabled us to provide a phylogenetic definition of cell type to the cells of the tree. Marker gene expression is traditionally used to identify cell types, but most marker genes have been identified and characterized in model organisms, especially humans. We find that in species distantly related to humans, phylogenetic signal is sufficient to identify such cells by membership in a cell type clade, even when these cells did not express these marker genes (for example, pig JCT[33]). The topology of the cell phylogeny provides a rich biological rationale to a phylogenetic classification of cells as these relationships describe the evolutionary mechanism by which new cell type identities arise.

We present several potential tools for the study of cell type evolution. The cell-level tree can be used to identify homologous cells by grouping into cell type clades (Fig. 4). Meanwhile, averaging expression levels increases topological stability, including among interior relationships, and so the averaged tree presents an avenue to examine the deeper evolutionary history between cell type clades (Fig. 5). Finally, we can use the GO terms that emerge from the most highly loaded genes to identify the cell type signal and investigate the expression dynamics of gene modules driving the function and identity of conserved cell type clades (Table 1).

Several ground-breaking papers[7–10,29] investigating cell phylogenies employed maximum parsimony or NJ to infer cell trees from discretized expression data. We seek to advance this progress by analysing continuous expression levels with an explicit model of evolution. By using continuous expression levels, we were able to harness established single-cell integration tools that resulted in the first cross-species cell phylogeny. Our approach builds from the use of non-model approaches[7,8,10] or distance-based methods[9,29] by using an explicit evolutionary model to infer the tree. We aim to provide a foundation for the development of more complex evolutionary models[57–60]. Such advanced models give us the opportunity to describe evolutionary scenarios that more closely match observed biology in a way that non-model or Brownian motion approaches cannot. For example, it is relevant that gene modules may be under stabilizing selection[57,58,61], as our use of PCs suggest that the units of analysis are cohorts of correlated genes. Finally, because our approach produces multiple trees rather than a single topology, we can apply and develop statistical tools to quantitatively investigate cell type evolution. A model-based approach can powerfully bridge the fields of cell biology, phylogenetics and evolutionary modelling, providing access to fresh tools and contributing a novel perspective to cell type evolution.

While there are many processes that may create branching patterns, several observations indicate that the signal in the first PCs contains evolutionary variance across cells that is appropriately analysed in a phylogenetic context. A 'cell type' is an identity[11,13] that transcends transient physiological conditions, tissue contexts, life histories and correlated within-species evolution[62], among many other factors. We find that the cell type clades that emerged in our cell phylogeny are consistent with well-established, independently verified[33] cell type identities that have persisted across five distantly related species (Figs. 4 and 5). These identities emerged even in the face of fluctuations in cell states, which may be embodied as cell-specific variance described by long tip branches (Figs. 4 and 5). Additionally, the LSI scores indicate that variance in the phylogenetic signal is uniquely specific to the cell type clade: cells of the same clade exhibit phylogenetic instability as a unit, possibly moving around the tree together as a clade (Fig. 4 and Extended Data Fig. 5). This provides a unique 'meta-phylogenetic' character independent from phenotypic similarity that can be used to characterize cell types, further supporting the utility of phylogenetic signal to define cell types.

The regulatory activity of transcription factors plays an important role in establishing cell type identity[11–14,63,64]. While we have identified transcription factors among the most highly loaded genes (for example, MITF) and have found significantly enriched molecular function GO terms that describe the activity of transcription factors (for example, transcription regulator activity (GO:0140110) or DNA-binding transcription factor activity (GO:0003700); Supplementary Table 6), not all correlated gene expression will directly correspond to a regulatory signal. However, the gene modules that emerge from expression constitute a phenotype that arises from underlying regulatory activity[65–67]. Through this pathway, PCs provide effective information on the evolutionary history of cell types.

The topology of a cell phylogeny allows us to infer specific events in the history of cell type evolution. The node at the base of the tree is an ancestral cell in a common ancestor, which then differentiated in the course of evolution to give rise to multiple distinct cells (Fig. 4). Other interior nodes allow us to infer the mechanism by which new cell types emerge. Much like in a gene tree, some interior nodes represent divergences between cell types and others divergences between species (Fig. 4). The node defining the Schwann cell superclade is an example of a within-cell-type divergence; an ancestral Schwann cell type programme split into the myelinating and non-myelinating subtypes (Fig. 4, black node). There was subsequent divergence through speciation events, in which both Schwann cell subtypes each split with the divergence of the five species (Fig. 4, white nodes). These evolutionary events provide the biological reasoning behind a phylogenetic definition of cell types[12]. The sister clade relationship of the myelinating and non-myelinating Schwann cell types suggests that these two cell types may have arisen as sister cell types during animal evolution[11–14].

The topology of the cell phylogeny is also consistent with hypothesized scenarios of cell type evolution[68]. Animal cell type evolution can be traced back to the emergence of the three germ layers and different germ-layer origins is a fundamental distinction between

cell types[69,70]. Eye development involves a complex series of events involving ectoderm-derived neural crest cells and mesoderm-derived cell types[71]. While vessel endothelia and corneal endothelia are both described as endothelia, they are not unified in a single clade in the cell phylogeny (Figs. 4 and 5). Accordingly, the corneal endothelium has a distinct embryological origin from vessel endothelia, arising not from the mesoderm but from the first wave of neural crest cells during eye development[71]. The corneal endothelium is among the earliest tissues to differentiate during eye development[72] and is placed as sister to all other cells in the rooted cell phylogeny (Figs. 4 and 5).

More fine-grained evolutionary stories are also present in the cell phylogeny. Our phylogeny hypothesizes a close evolutionary relationship between the Schlemm's canal, collector channel and vascular endothelium cells, which emerged as a robustly supported superclade ('vessel endothelia'; Figs. 4 and 5). Some early work suggested a neural crest origin of the Schlemm's canal[73,74], but embryological studies have since traced its emergence from the limbal blood vessels[75], congruent with its grouping with mesoderm-derived cells in the vessel endothelia clade. Collector channels are channels that are continuous with the Schlemm's canal, directly connecting the Schlemm's canal to the aqueous veins. Less work has been performed on their developmental origins, but their anlage appears to arise from straight vessels that sprout from the outer wall of the Schlemm's canal[76,77]. It has also been hypothesized that the collector channels develop from the intrascleral venous plexus[78,79]. Despite this close anatomical and functional relationship, the cell phylogeny hypothesizes that the collector channel is more closely related to the vascular endothelia of capillaries instead of the Schlemm's canal, with the collector channel forming a clade nested within the vascular endothelia (Fig. 5). Vertebrates possess two vascular systems—the blood and lymphatic systems. The mesoderm gives rise to blood vessel endothelia, which may then transition to a lymphatic fate by the expressions of specific marker genes, like *PROX1*[80]. The expression of these lymphatic marker genes in the Schlemm's canal (for example, *PROX1*, *CCL21*, *FLT4*/*VEGFR3*)[33] has led to the emerging hypothesis that the Schlemm's canal is a unique vessel with some affinity to the lymphatic system[75,81]. Collector channels do not express these genes—instead they express venular marker genes (for example, *ACKR1*)[33]. The topology of the cell phylogeny is consistent with the venous versus lymphatic split of the two vascular systems during vertebrate evolution and development, with the collector channel being of venous origin, while the Schlemm's canal may have a lymphatic association.

When the tips of the cell phylogeny are averages of replicate cells, topological stability within and between cell type clades notably increases (Fig. 5). The pericytes and ciliary muscle cells, both neural crest-derived, emerged as sister clades with high jumble and scjackknife scores (Fig. 5). However, in the eye they hold notably distinct biological roles[71]. While their link is understudied, extensive parallels have been drawn between pericytes and vascular smooth muscle cells[82–84]. Both are contractile mural cells that regulate blood flow[82] and both share key smooth muscle marker genes, including *ACTA2*[33,82]. It is in this smooth muscle identity that evident similarities between pericytes and ciliary muscle lie. *ACTA2* is also selectively expressed by ciliary muscle[33], and the canonical vascular smooth muscle cell regulators nitric oxide (NO[85]) and endothelin-1[86] govern the contraction of both ciliary muscle[87,88] and pericytes[89]. Neural crest cell types present a unique challenge due to potentially diverse evolutionary origins, but these similarities suggest a close evolutionary relationship between pericytes and ciliary muscle, consistent with phylogenetic topology (Fig. 5).

All cross-species studies are implicitly evolutionary[37–39], but despite growing excitement in comparative scRNA-seq we still lack a true phylogenetic comparative approach to cell biology. Here we provide a means that will enable an explicit phylogenetic approach to evolutionary questions involving potentially any cells that can be described by scRNA-seq data. Not only does this allow the leverage of phylogenetic comparative methods directly to cell biology, but the ability to infer the cell phylogeny completes the triptych of trees describing the cell lineage, phenotype similarity and evolutionary history of cells. These three trees, with their distinct but mechanistically related topologies, have long existed in separate fields of study. However, it is their deconvolution and then systematic comparison that will give rise to insights that cannot be achieved by any of the trees alone, as the most compelling biology exists where they are incongruent. For example, one principle of development observed in *Caenorhabditis elegans* is that cell transcriptomes exhibit an initial 'cell lineage signal' earlier in development[17]. However, as cell division progresses, this signal fades and the transcriptomes of cells of the same cell type, which are characterized by shared phenotypic characteristics, converge when adult terminal differentiation is approached[17]. This divergence in the topologies of the developmental lineage and phenotypic similarity trees may possibly be linked by the cell phylogeny: phenotypic similarity may have arisen from shared evolutionary ancestry, not development. The cell phylogeny will enable us to deconvolute the intertwined branching patterns of phenotype and development, while providing a unifying mechanistic understanding of the emergence of cell type diversity at both the developmental and deep evolutionary time scales. Together, as the topologies of these three trees are systematically explored, a phylogenetic approach to cell biology will provide a path to connect the branching pattern of cell evolution to the broader tree of life.

Methods

Several technical challenges exist in analysing gene expression data across species. These include normalization and integration of counts across species[43–45,62], the highly correlated nature of gene expression, the prevalence of noise, and the extreme multidimensionality of the number of characters. We approach these problems by using established scRNA-seq bioinformatics workflows to normalize and integrate counts[43–45], and PCA to rotate the data and reduce rank. This facilitates applying a Brownian-motion-based model to the data, and we use this explicit evolutionary model to infer the cell phylogeny.

Data

We used scRNA-seq counts from van Zyl et al.[33] as our pilot dataset (National Center for Biotechnology Information Gene Expression Omnibus accession number GSE146188). van Zyl et al.[33] sampled cells of the aqueous humour of the anterior segment of the eye from five adult model species: *Macaca mulatta* (GSE148374), *Macaca fascicularis* (GSE148373), *Mus musculus* (GSE146186), *Sus scrofa* (GSE146187) and *Homo sapiens* (GSE148371). Unique molecular identifier counts were downloaded as CSV files. This dataset was chosen for the uniformity of sampling, consistency of lab and sequencing protocols, the high quality of its cell type annotations, and the abundance of genomic resources available for the five model species. A file containing meta-data, including cluster assignment and cell type labels, was obtained from the Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP780/).

van Zyl et al.[33] identified cell types by clustering their human dataset and assigning a cell type label to each human cluster according to marker gene expression and histology. They then clustered the other non-human datasets and used a machine learning algorithm to identify clusters from non-human species corresponding to the labelled human clusters. Because correspondence between clusters was not always 1:1, some non-human species had multiple clusters identified to the same human cell type cluster. Similarly, some human cell type clusters were absent in other species, while other non-human clusters were not present in humans. The latter were given their own unique cell type label. In our analyses, we considered each labelled cluster in any one species a 'cell type group', following the labelling scheme of van Zyl et al.[33]. The number of cell type groups possessed by each species

ranged from 15 to 20. While gene expression levels underlie the basis of both the cell type labels and the cell phylogenies we infer here, van Zyl et al.[33] additionally verified the accuracy of these labels through histology. This verification of cell type identity, independent of gene expression, allowed us to distinguish grouping in the cell phylogeny due to any potential circularity versus expected similarity due to homology.

Normalization and integration within and across species

Standard Seurat (version 4.1.0)[44] workflows were leveraged for normalization and integration[43–45] in R (version 4.2.2; https://www.R-project.org/). For within-species analysis, sctransform[43] in Seurat was used to normalize the counts of each species' matrix for cell sequencing depth and variance stabilization, using a negative binomial model of counts (Extended Data Fig. 1, step 2). The Seurat function IntegrateData[45] was subsequently used to make counts comparable across within-species sample batches (Extended Data Fig. 1, step 3). The integrated Pearson residuals (from the 'integrated assay', 'scale.data' slot) resulting from normalization and integration were considered in all further analyses. The total number of cells sampled per species ranged from 24,023 (human) to 5,067 (mouse), resulting in large matrices (Extended Data Fig. 1, step 1). To reduce computational burden, matrices were subsetted such that each cell type group had an equal number of replicates (Extended Data Fig. 1, step 4). The cell type group with the fewest replicates across all five species were mouse B cells, for which ten cells were sequenced. Thus, when subsetting we randomly selected ten replicate cells per cell type group in all five matrices (Extended Data Fig. 1, step 4). The number of cell type groups per species ranged from 15 to 20, totalling 92 cell type groups represented by 920 cells across the five species (Extended Data Fig. 1, step 4).

For cross-species analysis, a mapping of all homologues shared between the five model species was obtained from Ensembl[90] BioMart, which identifies homologues using gene trees. Ensembl constructs these gene trees by a species-aware process that finds the consensus of NJ and maximum likelihood trees inferred from protein and DNA sequences in the Ensembl database (version 107, accessed September 2022)[90,91]. The five matrices were then joined using one-to-one orthologues common to all five species to create a combined matrix with 920 cells and 5,870 genes (Extended Data Fig. 1, step 5). One cell was removed after filtering for cells that expressed at least 200 genes and genes that were expressed in at least three or more cells. IntegrateData[45] was used a second time to integrate values across species in this combined matrix (Extended Data Fig. 1, step 6). Subsequent analysis was performed on the integrated Pearson residuals for the 2,000 most highly variable genes, a threshold drawn to reduce computational burden and prioritize genes with increased biological signal (Extended Data Fig. 1, step 6).

Successful integration was confirmed by producing a UMAP and annotating cells by species identity and cell type identity (Extended Data Fig. 2). Although overcorrection is a potential artefact, we did not find over-grouping of unrelated cell types in the cell phylogeny (Figs. 4 and 5). The final matrix before rotation and rank reduction had 919 cells and 2,000 genes (Extended Data Fig. 1, step 6).

Rotation and rank reduction of the data

Gene expression data is highly noisy, highly correlated and extreme in its multidimensionality. Most models of continuous character evolution are based on Brownian motion[92]. While Brownian motion models can generate correlated data, there are multiple technical challenges to the inference of continuous data with correlation[47]. In addition, the number of parameters required to be estimated for models of correlated continuous character evolution increases quadratically with each additional character, as the mean, variance and covariances of each character must each be estimated. This results in a small N:p ratio (where N represents the number of samples and p the number of parameters), which makes parameter estimation difficult. The number of samples in our single-cell dataset (919 cells; Extended Data Fig. 1, step 6) is far outweighed by the number of parameters that must be calculated when data are correlated. The extensive correlation among gene expression suggests that the data may be low-rank, and together with the above challenges it is therefore desirable to transform the data to this lower-rank space with less correlation.

We performed a PCA on the combined cross-species integrated matrix (Extended Data Fig. 1, step 7). We first centred the integrated data around the mean before performing the PCA (using 'prcomp' in R), then subsequently normalized the variance of the resulting PCs by dividing by their standard deviation. This produced a character matrix with 919 cells and 919 PCs (Extended Data Fig. 1, step 7). The rotation of our data to these new axes produced PCs that are independent and uncorrelated. We use these PCs as phylogenetic characters moving forward.

We achieved rank reduction by subsequently discarding a subset of the later PCs. PCs are sorted in descending order by the variance they explain. Earlier PCs encode broad, slow-changing patterns of variance (often referred to as low-frequency signal), while later PCs encode rapid, fine-grained variance (that is, high-frequency signal). In a phylogenetic framework, low-frequency changes occur slowly across the whole topology of the phylogeny, while high-frequency changes occur rapidly across all tips regardless of the greater structure of the tree. Here, we hypothesize that early PCs are enriched for phylogenetic signal, while later PCs represent cell-specific noise.

We plotted the variance described by each PC and confirmed that the data, although high-dimensional, were low-rank (Fig. 1). We identified the number of PCs to retain by plotting the variance described by each PC and identifying the elbow in the variance plot where including further PCs only minimally increases the amount of variance described (Fig. 1). This elbow occurred at PC 20, and so we retained the first 20 PCs for use as phylogenetic characters, trimming the rest.

By taking the first 20 PCs, we aim to reduce noise, enrich for phylogenetic signal and drastically reduce the number of characters from 2,000 to 20. Our procedure also greatly simplified parameter estimation, making the model computationally feasible. Centring the data before performing PCA collapsed all means to zero and, similarly, the use of orthogonal PCs set all covariances between characters to zero.

Phylogenetic inference using a Brownian motion model

We used an explicit model of evolution based on Brownian motion to infer our cell phylogenies[47,92]. While the use of Brownian-motion-based models is well-established in the inference of phylogenies from continuous characters, like morphology[47,92–95], evolutionary models of gene expression remain underdeveloped[57]. Phylogenetic models were originally designed upon assumptions based on mechanisms of species evolution rather than cells, and some studies indicate that the use of more complex Ornstein–Uhlenbeck models may better fit the heterogeneity of most gene expression datasets[57]. Our study is unique in that instead of using gene expression values directly, we use PCs calculated from gene expression values as our phylogenetic characters. In addition, we remove later PCs that may represent highly heterogeneous cell-specific signal. Further investigation exploring the behaviour of PCs with more complex evolutionary models would yield productive insights.

Because PCs are orthogonal and uncorrelated to each other, this allowed the data to better fit the feature independence assumption of the uncorrelated Brownian motion model[47], which we used to infer the cell phylogeny. While we also aimed to increase taxon independence with cross-species integration, other potential among-cell correlations may exist and future studies will greatly benefit from the development of more complex models of evolution (for example, a Brownian motion or Ornstein–Uhlenbeck model accounting for correlation and so on) to extend our uncorrelated Brownian motion approach.

To infer the cell phylogeny, we used contml from PHYLIP (version 3.698; https://phylipweb.github.io/phylip/). This program performs

phylogenetic inference of continuous data by maximum likelihood search, calculating likelihoods using the REML PIC algorithm[47], an evolutionary model explicitly based on Brownian motion.

PC sweep experiment

We performed a sweep experiment to better understand the effect of the number of retained PCs on characteristics of the phylogeny (Fig. 2). To do so, we created 98 919-cell matrices that varied by the number of PCs (Extended Data Fig. 1). From a 919-cell, 919-PC matrix (Extended Data Fig. 1, step 7), the 98 character matrices were made by serially increasing the number of PCs retained, from 3 PCs (PCs 1–3 retained) to 100 PCs (PCs 1–100 retained; Extended Data Fig. 1, step A1). It was not possible to infer a phylogeny from a character matrix that featured fewer than 3 PCs. We inferred cell phylogenies with contml using the 'C' (continuous character) option and leaving all other settings at default. For each of these 98 trees, we calculated total tree length (sum of all edge lengths), the sums of the tip and interior edge lengths, and the star-ness score (the ratio of the sum of tip edge lengths to the sum of interior edge lengths). These values were plotted in R to produce Fig. 2.

To inspect the emergence of cell type clades as the number of PCs were manipulated, we followed the above procedures and inferred a tree from a 919-cell, 919-PC matrix (Extended Data Fig. 1, step 7) subsetted to 20, 500 and 919 PCs (Fig. 3 and Extended Data Fig. 1, steps B1–3).

Inference of the focal cell-level phylogeny

We used contml to infer the focal phylogeny of Fig. 4. The 'C' option was used to specify continuous characters were to be analysed. The 'G' option was selected to search tree space by global rearrangement. Finally, the 'J' option ('Jumble') was selected to randomize the input order of the taxa 100 times and select the highest likelihood tree (see Technical repeatability: jumble scores for detailed discussion of the Jumble setting). All other contml settings were left at default. Finally, the cell phylogeny was midpoint-rooted.

Preliminary phylogenetic inference runs were performed on a 92-cell, 20-PC matrix (Extended Data Fig. 1, step C1.1). We started with a 919-cell, 919-PC matrix (Extended Data Fig. 1, step 7). We then subsetted the PCs to 20 PCs and randomly subsetted cells to one cell per cell type group per species to create a 92-cell, 20-PC matrix representing 92 cell type groups (Extended Data Fig. 1, step C1.1). After initial inference on this 92-cell matrix (Extended Data Fig. 3), we removed a minority of cell type groups that consistently failed to form monophyletic groups and destabilized the topology. This is similar to removing rogue taxa from a species phylogeny. These unstable cells were the corneal epithelium, beam A, beam X, fibroblast and B cells. van Zyl et al.[33] had found that some clusters in non-human species expressed a mix of both fibroblast and beam cell marker genes, making identification of these cell types difficult to disentangle across species. Similarly, in a subset of human samples, B cells were poorly sampled and in some cases had to be identified histologically when identification by gene expression alone failed[33]. We additionally removed cell type groups represented by only a single tip, as it is difficult to meaningfully interpret the relevance of such singletons in the context of other cell type clades. This produced our final 54-cell, 20-PC matrix (Extended Data Fig. 1, step C1.2), representing 54 cell type groups. Phylogenetic inference was performed on this matrix to produce the focal cell-level phylogeny in Fig. 4.

Technical repeatability: jumble scores

The bootstrap procedure, which is commonly used to assess the repeatability of maximum likelihood trees, is built upon an assumption that each phylogenetic character is weighted equally[48]. This is not true of our characters, as PCs are weighted in descending order by the amount of variance described. This means bootstrapping is not appropriate for our application.

Instead, we assessed the repeatability of the topology by perturbing the starting position of the maximum likelihood search in tree space. Should phylogenetic inference be repeatable, the same topology should be inferred regardless of where the maximum likelihood search was initiated. This perturbation was achieved by using the 'J' option ('Jumble') of contml, which created different starting trees by randomizing the input order of the taxa. The contml options we used are 'C' (continuous characters), 'J' (jumble 100×), 'G' (global rearrangement) and all other settings at default. We performed 250 contml runs on the same dataset, jumbling 100 times each run and inferring a phylogeny from each jumble. This produced 100 'jumble trees' per run, from which the phylogeny with the highest likelihood was selected and presented as the final output of the run. A percentage (5.2%) of contml runs failed, leaving us with 237 high-likelihood jumble trees. We then calculated the clade frequencies of these 237 jumble trees using plotBS (with type = 'phylogram' and method = 'FBP') from the phangorn[96] R package (version 2.10.0). These frequencies were plotted onto the nodes of the focal cell-level phylogeny of Fig. 4 as 'jumble scores', which represent the percentage of jumble trees in which that split was present.

The focal cell-level phylogeny was selected from among the 237 jumble trees by identifying the phylogeny with the greatest overall sum of jumble scores. Fourteen phylogenies fit this criterion. They possessed identical topologies (Robinson–Foulds distance[97] = 0) or minimal differences when branch length was taken into account (average branch-score distance[98] = 0.0298). The tree with the smallest sum of branch-score distances between the 14 trees was selected as the focal phylogeny presented in Fig. 4.

Biological repeatability: scjackknife scores

We expect that cell gene expression may exhibit considerable biological variability, even among cells of the same cell type group, and thus needed to measure the robustness of the topology to cell sampling. We assessed biological repeatability using a scjackknife procedure. This was made possible by the fact that single-cell datasets provide multiple replicate cells per cell type group. We performed a jackknife procedure, where we randomly drew one cell per cell type group per species from a 919-cell, 20-PC matrix (Extended Data Fig. 1, step C2.1), which possesses ten cells per cell type group. We selected cells from among the 54 cell type groups that were used to infer the phylogeny in Fig. 4. This produced a jackknifed matrix with 54 resampled cells and 20 PCs (Extended Data Fig. 1, step C2.2). We performed this jackknife procedure 500 times to create 500 new 54-cell matrices (Extended Data Fig. 1, step C2.2), where the specific identity of the cell for a particular cell type group had been randomly sampled. The number of PCs was kept consistent at 20 PCs. From each matrix, a phylogeny was inferred with contml using the same settings as described above for the jumble trees. After removing runs that errored, we retained 439 high-likelihood scjackknife trees.

We used the transfer bootstrap expectation score[49] to measure the consistency of the topology across the scjackknife trees, and used this as a measure of support for biological repeatability. The transfer bootstrap expectation is similar to Felsenstein's bootstrap[48], where the frequency of a clade is calculated across a set of trees, except instead of scoring the presence of a clade through a simple presence/absence index, it uses transfer distance to identify clades that are present but exhibit some variability across comparison trees. Transfer distance between a branch in the reference tree and a branch in the comparison tree is equal to the number of tips that must be removed ('transferred') to make the two branches equivalent[49]. This is useful for datasets that exhibit clear topological patterns, but whose clades exhibit small variation not captured by a Boolean presence/absence procedure. In Felsenstein's bootstrap, even if all but one taxon is present in the comparison clade, the bootstrap score is 0, which is not reflective of the high degree of similarity shared by the two clades (Extended Data Fig. 4). We do not expect that the stability exhibited by species trees

will be matched by a cell phylogeny. The cell phylogeny possesses taxa that exist at an entirely different level of organization and is built using highly variable gene expression, rather than molecular sequence data. Likewise, there is no precedence for choosing a significance threshold in a cell phylogeny. A previous study[49] found that a transfer bootstrap expectation threshold of 70% reliably identified branches concordant with the true tree of a species phylogeny built from molecular sequence data, and so we use this threshold to conservatively define well-supported clades. We calculated the transfer bootstrap expectation from the 439 scjackknife trees using BOOSTER (https://github.com/evolbioinfo/booster) and mapped these scores onto the focal phylogeny in Fig. 4 as 'scjackknife scores'.

LSI

The LSI[50] was calculated from the set of jumble trees produced for Fig. 4 using Phyutility[99] (version 2.2), with the command: 'java -jar phyutility.jar -ls -in jumble_trees.tre'. These values were plotted by colour onto the cell phylogeny in Fig. 4.

Averaging replicate cells

The 919-cell, 20-PC matrix (Extended Data Fig. 1, step C2.1) has approximately ten replicate cells per cell type group (92 cell type groups). We followed the above procedure to infer a tree and calculate jumble and scjackknife scores, but instead of selecting a single representative cell per cell type group we averaged values across multiple replicates. For the jumble trees, we averaged the PC loadings of all ten replicates, selecting the 54 cell type groups used in Fig. 4. This produced a matrix with 54 averaged cell type groups and 20 PCs (Extended Data Fig. 1, step D1). We performed 500 jumble tree runs and received 209 high-likelihood jumble trees, where each tip is a cell type group represented by the average of ten replicate cells. The 'C' (continuous character) and 'J' (jumble, 100×) options were used with contml. For the scjackknife trees, we randomly sampled 5 cells for each cell type group from the 919-cell, 20-PC matrix (Extended Data Fig. 1, step C2.1), selecting for the 54 cell type groups used in Fig. 4. Their PC loadings were averaged to create a new jackknifed matrix, with 54 averaged cell type groups and 20 PCs (Extended Data Fig. 1, step D2). We created 500 scjackknife matrices and received 393 high-likelihood scjackknife trees, where each tip indicates a cell type group represented by the average of 5 randomly chosen replicates. The same contml options were used as described above for the jumble runs. We used the best scoring jumble tree to plot jumble and scjackknife scores in Fig. 5. Five jumble trees tied for the highest sum of jumble scores; the most topologically representative among this subset was selected as the 'best' jumble tree following the same procedure as described for Fig. 4. We also calculated the LSI for the Fig. 5 tree using the same commands described above.

Analysis of highly loaded PC genes

We examined the identity of the genes with the highest positive and negative loadings for the first 20 PCs from the results of the PCA that produced the 919-cell, 919-PC matrix (Extended Data Fig. 1, step 7). We extracted the gene loadings, identified genes by their human gene symbol and sorted them by most positive and most negative loading scores for each PC. We then performed a GO enrichment analysis with GOseq[100], using the top 100 most positively or negatively loaded genes for each PC. *P* values were adjusted for multiple testing using the Benjamini–Hochberg[101] method. The 20 most significantly enriched biological process GO terms were used to discern the presence or absence of a cell type signal (summarized in Table 1). These GO terms are provided in Supplementary Table 1. We also include functional descriptions of the top ten most positively or negatively loaded genes for each PC that was curated from the National Center for Biotechnology Information Gene database. These genes and descriptions are summarized in Supplementary Table 3 for positive loadings and Supplementary Table 5 for negative loadings.

Comparison to phenetic trees built using distance-based methods

In R, distance matrices were calculated from the averaged matrix used for Fig. 5 (Extended Data Fig. 1, step D1) using the 'dist' function (method = 'euclidean') from the 'stats' package. The following commands were then used to create an NJ[28] tree ('nj' from the 'ape' package[102]), an UPGMA[53] tree ('upgma' with method = 'average', from the 'phangorn' package[96]) and a WPGMA[54] ('wpgma' with method = 'mcquitty', from the 'phangorn' package[96]) tree (Extended Data Figs. 7–9). We also replicated the scjackknife experiment using the 500 scjackknife matrices created for Fig. 5 (Extended Data Fig. 1, step D2), but calculating phenetic trees using NJ instead of an evolutionary Brownian motion model. These NJ scjackknife scores were plotted onto the cell phylogeny of Fig. 5 to facilitate a direct comparison of topologies (Extended Data Fig. 10).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

A GitHub repository (https://github.com/dunnlab/cellphylo) is provided containing select input, intermediate and output files ('cellphylo/analysis/') sufficient to reproduce the analyses.

Code availability

All custom code is available in our GitHub repository: https://github.com/dunnlab/cellphylo.

References

1. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
2. Martindale, M. Q. & Henry, J. Q. Intracellular fate mapping in a basal metazoan, the ctenophore *Mnemiopsis leidyi*, reveals the origins of mesoderm and the existence of indeterminate cell lineages. *Dev. Biol.* **214**, 243–257 (1999).
3. Spencer Chapman, M. et al. Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).
4. Tanay, A. & Sebé-Pedrós, A. Evolutionary cell type mapping with single-cell genomics. *Trends Genet.* **37**, 919–932 (2021).
5. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323 (2016).
6. Gilbert, E. et al. Molecular and cellular architecture of the larval sensory organ in the cnidarian *Nematostella vectensis*. *Development* **149**, dev200833 (2022).
7. Kin, K., Nnamani, M. C., Lynch, V. J., Michaelides, E. & Wagner, G. P. Cell-type phylogenetics and the origin of endometrial stromal cells. *Cell Rep.* **10**, 1398–1409 (2015).
8. Liang, C., Forrest, A. R. R. & Wagner, G. P. The statistical geometry of transcriptome divergence in cell-type evolution and cancer. *Nat. Commun.* **6**, 6066 (2015).
9. Musser, J. M. et al. Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *Science* **374**, 717–723 (2021).
10. Hughes, A. L. & Friedman, R. A phylogenetic approach to gene expression data: evidence for the evolutionary origin of mammalian leukocyte phenotypes. *Evol. Dev.* **11**, 382–390 (2009).
11. Wagner, G. P. *Homology, Genes, and Evolutionary Innovation* (Princeton Univ. Press, 2014).
12. Arendt, D. et al. The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).

13. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* **9**, 868–882 (2008).
14. Arendt, D. Evolution of eyes and photoreceptor cell types. *Int. J. Dev. Biol.* **47**, 563–571 (2003).
15. Arendt, D., Bertucci, P. Y., Achim, K. & Musser, J. M. Evolution of neuronal types and families. *Curr. Opin. Neurobiol.* **56**, 144–152 (2019).
16. Serb, J. M. & Oakley, T. H. Hierarchical phylogenetics as a quantitative analytical framework for evolutionary developmental biology. *Bioessays* **27**, 1158–1166 (2005).
17. Packer, J. S. et al. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).
18. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
19. Whitman, C. O. The embryology of *Clepsine*. *J. Cell Sci.* **s2-18**, 215–315 (1878).
20. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
21. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
22. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
23. Yang, D. et al. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* **185**, 1905–1923 (2022).
24. Levy, S. et al. A stony coral cell atlas illuminates the molecular and cellular basis of coral symbiosis, calcification, and immunity. *Cell* **184**, 2973–2987 (2021).
25. Seidel, S. & Stadler, T. TiDeTree: a Bayesian phylogenetic framework to estimate single-cell trees and population dynamic parameters from genetic lineage tracing data. *Proc. R. Soc. B* **289**, 20221844 (2022).
26. Zhao, Z.-M. et al. Early and multiple origins of metastatic lineages within primary tumors. *Proc. Natl Acad. Sci. USA* **113**, 2140–2145 (2016).
27. Moravec, J. C., Lanfear, R., Spector, D. L., Diermeier, S. D. & Gavryushkin, A. Testing for phylogenetic signal in single-cell RNA-seq data. *J. Comput. Biol.* **30**, 518–537 (2023).
28. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
29. Paganos, P., Voronov, D., Musser, J. M., Arendt, D. & Arnone, M. I. Single-cell RNA sequencing of the *Strongylocentrotus purpuratus* larva reveals the blueprint of major cell types and nervous system of a non-chordate deuterostome. *Elife* **10**, e70416 (2021).
30. Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
31. Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. & Shapiro, E. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* **1**, e50 (2005).
32. Tarashansky, A. J. et al. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10**, e66747 (2021).
33. 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. USA* **117**, 10339–10349 (2020).
34. Sebé-Pedrós, A. et al. Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat. Ecol. Evol.* **2**, 1176–1188 (2018).
35. Wang, R. et al. Construction of a cross-species cell landscape at single-cell level. *Nucleic Acids Res.* **51**, 501–516 (2023).
36. Chen, D. et al. Single cell atlas for 11 non-model mammals, reptiles and birds. *Nat. Commun.* **12**, 7083 (2021).
37. Dunn, C. W., Zapata, F., Munro, C., Siebert, S. & Hejnol, A. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl Acad. Sci. USA* **115**, E409–E417 (2018).
38. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985).
39. Grafen, A. The phylogenetic regression. *Phil. Trans. R. Soc. Lond. B* **326**, 119–157 (1997).
40. Sebé-Pedrós, A., Degnan, B. M. & Ruiz-Trillo, I. The origin of Metazoa: a unicellular perspective. *Nat. Rev. Genet.* **18**, 498–512 (2017).
41. Mah, J. L., Christensen-Dalsgaard, K. K. & Leys, S. P. Choanoflagellate and choanocyte collar-flagellar systems and the assumption of homology. *Evol. Dev.* **16**, 25–37 (2014).
42. Laundon, D., Larson, B. T., McDonald, K., King, N. & Burkhardt, P. The architecture of cell differentiation in choanoflagellates and sponge choanocytes. *PLoS Biol.* **17**, e3000226 (2019).
43. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
44. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
45. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
46. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. Preprint at http://arxiv.org/abs/1802.03426 (2020).
47. Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25**, 471–492 (1973).
48. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
49. Lemoine, F. et al. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).
50. Thorley, J. L. & Wilkinson, M. Testing the phylogenetic stability of early tetrapods. *J. Theor. Biol.* **200**, 343–344 (1999).
51. Levy, C., Khaled, M. & Fisher, D. E. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends Mol. Med.* **12**, 406–414 (2006).
52. van Zyl, T. et al. Cell atlas of the human ocular anterior segment: tissue-specific and shared cell types. *Proc. Natl Acad. Sci. USA* **119**, e2200914119 (2022).
53. Sokal, R. R. et al. *Principles of Numerical Taxonomy* (WH Freeman & Co, 1963).
54. Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **38**, 1409–1438 (1958).
55. Iwasa, M. A. & Suzuki, H. Evolutionary significance of chromosome changes in northeastern Asiatic red-backed voles inferred with the aid of intron 1 sequences of the *G6pd* gene. *Chromosome Res.* **10**, 419–428 (2002).
56. Leclaire, S., Menard, S. & Berry, A. Molecular characterization of *Babesia* and *Cytauxzoon* species in wild South-African meerkats. *Parasitology* **142**, 543–548 (2015).
57. Dimayacyac, J. R., Wu, S. & Pennell, M. Evaluating the performance of widely used phylogenetic models for gene expression evolution. Prepint at *bioRxiv* https://doi.org/10.1101/2023.02.09.527893 (2023).
58. Rohlfs, R. V., Harrigan, P. & Nielsen, R. Modeling gene expression evolution with an extended Ornstein–Uhlenbeck process accounting for within-species variation. *Mol. Biol. Evol.* **31**, 201–211 (2014).

59. Bertram, J. et al. CAGEE: computational analysis of gene expression evolution. *Mol. Biol. Evol.* **40**, msad106 (2023).
60. Harmon, L. J. et al. Early bursts of body size and shape evolution are rare in comparative data. *Evolution* **64**, 2385–2396 (2010).
61. Wagner, G. P. Homologues, natural kinds and the evolution of modularity. *Integr. Comp. Biol.* **36**, 36–43 (1996).
62. 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).
63. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462**, 587–594 (2009).
64. Hobert, O. Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. *Proc. Natl Acad. Sci. USA* **105**, 20067–20071 (2008).
65. Yin, W., Mendoza, L., Monzon-Sandoval, J., Urrutia, A. O. & Gutierrez, H. Emergence of co-expression in gene regulatory networks. *PLoS ONE* **16**, e0247671 (2021).
66. Spellman, P. T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
67. Wang, J. et al. Single-cell co-expression analysis reveals distinct functional modules, co-regulation mechanisms and clinical outcomes. *PLoS Comput. Biol.* **12**, e1004892 (2016).
68. Hall, B. K. Germ layers, the neural crest and emergent organization in development and evolution. *Genesis* **56**, e23103 (2018).
69. Hashimshony, T., Feder, M., Levin, M., Hall, B. K. & Yanai, I. Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature* **519**, 219–222 (2015).
70. Steinmetz, P. R. H., Aman, A., Kraus, J. E. M. & Technau, U. Gut-like ectodermal tissue in a sea anemone challenges germ layer homology. *Nat. Ecol. Evol.* **1**, 1535–1542 (2017).
71. Gage, P. J., Rhoades, W., Prucka, S. K. & Hjalt, T. Fate maps of neural crest and mesoderm in the mammalian eye. *Invest. Ophthalmol. Vis. Sci.* **46**, 4200–4208 (2005).
72. Williams, A. L. & Bohnsack, B. L. Neural crest derivatives in ocular development: discerning the eye of the storm. *Birth Defects Res. C Embryo Today* **105**, 87–95 (2015).
73. Rodrigues, M. M., Katz, S. I., Foidart, J. M. & Spaeth, G. L. Collagen, factor VIII antigen, and immunoglobulins in the human aqueous drainage channels. *Ophthalmology* **87**, 337–345 (1980).
74. Pandolfi, M. Coagulation factor VIII: localization in the aqueous outflow pathways. *Arch. Ophthalmol.* **94**, 656–658 (1976).
75. Kizhatil, K., Ryan, M., Marchant, J. K., Henrich, S. & John, S. W. M. Schlemm's canal is a unique vessel with a combination of blood vascular and lymphatic phenotypes that forms by a novel developmental process. *PLoS Biol.* **12**, e1001912 (2014).
76. Ramírez, J. M. et al. Schlemm's canal and the collector channels at different developmental stages in the human eye. *Cells Tissues Organs* **178**, 180–185 (2004).
77. Ashton, N. Anatomical study of Schlemm's canal and aqueous veins by means of neoprene casts. Part I. Aqueous veins. *Br. J. Ophthalmol.* **35**, 291–303 (1951).
78. Smelser, G. K. & Ozanics, V. The development of the trabecular meshwork in primate eyes. *Am. J. Ophthalmol.* **71**, 366–385 (1971).
79. Krohn, J. Expression of factor VIII-related antigen in human aqueous drainage channels. *Acta Ophthalmol. Scand.* **77**, 9–12 (1999).
80. Francois, M., Harvey, N. L. & Hogan, B. M. The transcriptional control of lymphatic vascular development. *Physiology* **26**, 146–155 (2011).
81. Aspelund, A. et al. The Schlemm's canal is a VEGF-C/VEGFR-3-responsive lymphatic-like vessel. *J. Clin. Invest.* **124**, 3975–3986 (2014).
82. Trost, A. et al. Brain and retinal pericytes: origin, function and role. *Front. Cell. Neurosci.* **10**, 20 (2016).
83. Alarcon-Martinez, L. et al. Capillary pericytes express α-smooth muscle actin, which requires prevention of filamentous-actin depolymerization for detection. *Elife* **7**, e34861 (2018).
84. Etchevers, H. C., Vincent, C., Le Douarin, N. M. & Couly, G. F. The cephalic neural crest provides pericytes and smooth muscle cells to all blood vessels of the face and forebrain. *Development* **128**, 1059–1068 (2001).
85. Ignarro, L. J. et al. Mechanism of vascular smooth muscle relaxation by organic nitrates, nitrites, nitroprusside and nitric oxide: evidence for the involvement of S-nitrosothiols as active intermediates. *J. Pharmacol. Exp. Ther.* **218**, 739–749 (1981).
86. Bouallegue, A., Daou, G. B. & Srivastava, A. K. Endothelin-1-induced signaling pathways in vascular smooth muscle cells. *Curr. Vasc. Pharmacol.* **5**, 45–52 (2007).
87. Kamikawatoko, S. et al. Nitric oxide relaxes bovine ciliary muscle contracted by carbachol through elevation of cyclic GMP. *Exp. Eye Res.* **66**, 1–7 (1998).
88. Lepple-Wienhues, A., Stahl, F., Willner, U., Schäfer, R. & Wiederholt, M. Endothelin-evoked contractions in bovine ciliary muscle and trabecular meshwork: interaction with calcium, nifedipine and nickel. *Curr. Eye Res.* **10**, 983–989 (1991).
89. Rucker, H. K., Wynder, H. J. & Thomas, W. E. Cellular mechanisms of CNS pericytes. *Brain Res. Bull.* **51**, 363–369 (2000).
90. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
91. Vilella, A. J. et al. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
92. Felsenstein, J. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution* **35**, 1229–1242 (1981).
93. Parins-Fukuchi, T. Use of continuous traits can improve morphological phylogenetics. *Syst. Biol.* **67**, 328–339 (2018).
94. Parins-Fukuchi, T. Bayesian placement of fossils on phylogenies using quantitative morphometric data. *Evolution* **72**, 1801–1814 (2018).
95. Caumul, R. & Polly, P. D. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (Marmota, Rodentia). *Evolution* **59**, 2460–2472 (2005).
96. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
97. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
98. Kuhner, M. K. & Felsenstein, J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11**, 459–468 (1994).
99. Smith, S. A. & Dunn, C. W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**, 715–716 (2008).
100. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
101. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
102. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

Acknowledgements
We express our gratitude to members of the Dunn lab, including S. Church, for their invaluable feedback throughout this project. We would also like to thank J. Musser, G. Wagner, D. Stadtmauer, A. Chavan, D. Adams and L. Revell for insightful comments that greatly improved our manuscript and analyses. We thank the Yale Center for Research Computing for guidance and the cloud resources provided as a part of the AWS Cloud Credit for Research Program at Yale. J.L.M. acknowledges funding from the Gruber Foundation (Gruber Science Fellowship) and the Natural Sciences and Engineering Research Council of Canada (NSERC PGS-D).

Author contributions
J.L.M. and C.W.D. conceptualized this study and developed its methodology. J.L.M. designed and conducted the formal analyses, investigation and visualization, wrote the manuscript and performed revisions. C.W.D. additionally supervised the project and contributed text, feedback and edits.

Competing interests
The authors declare no competing interests.

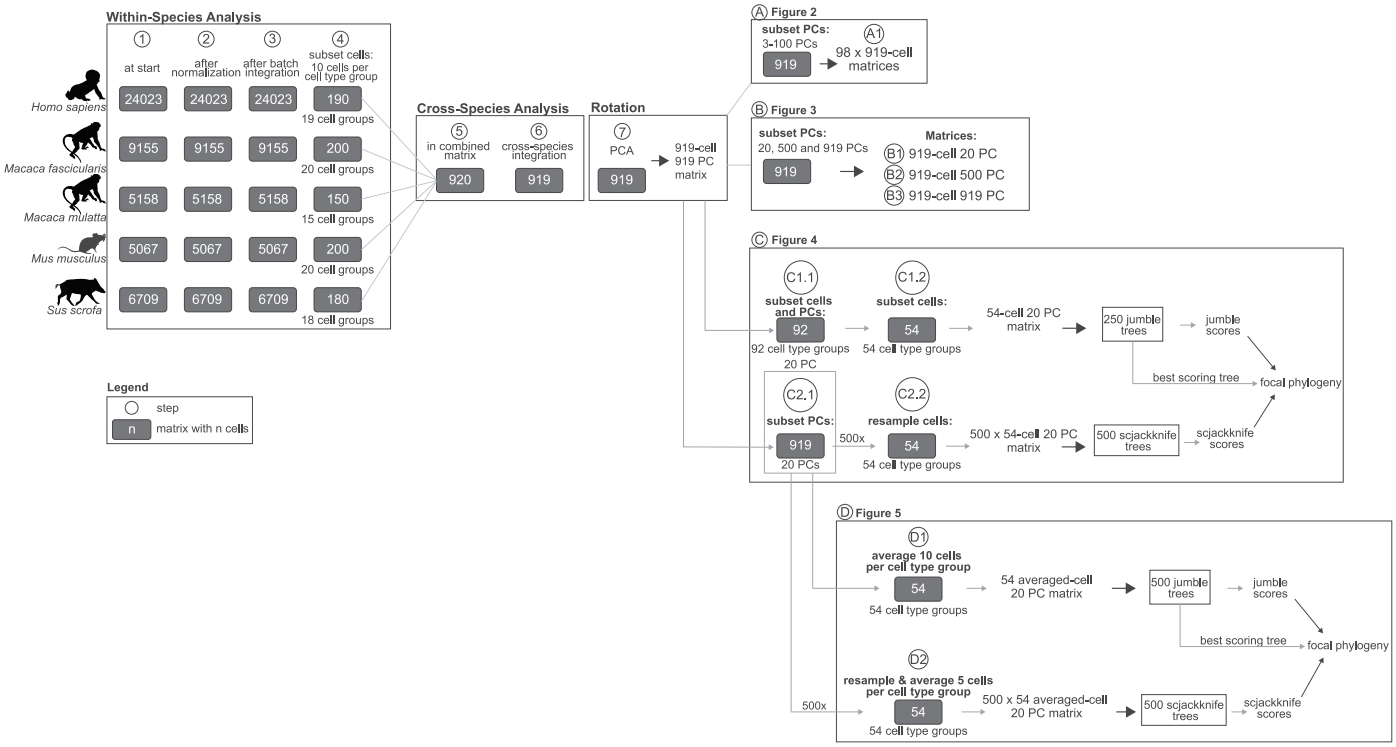
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41559-023-02281-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-023-02281-9.

Correspondence and requests for materials should be addressed to Jasmine L. Mah.

Peer review information *Nature Ecology & Evolution* thanks Xavier Grau-Bové and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

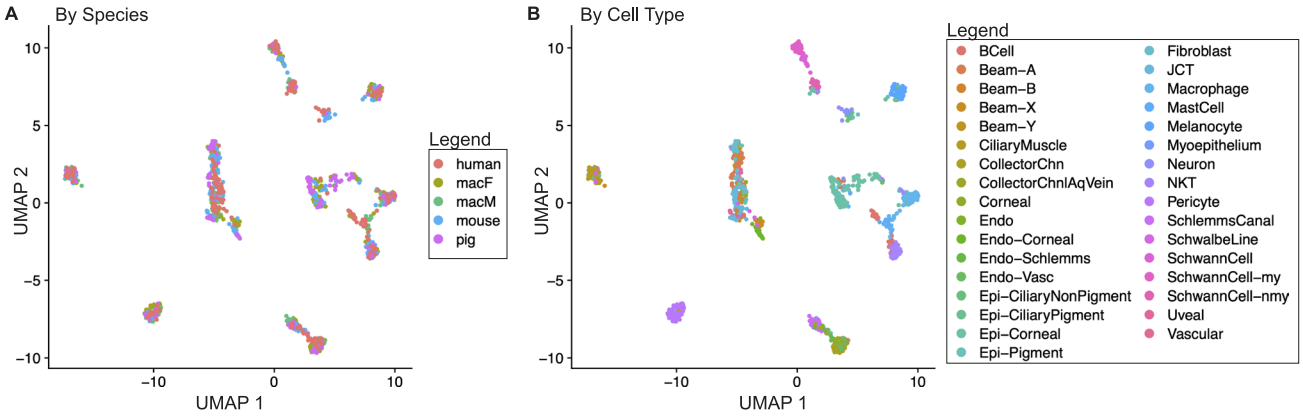
https://doi.org/10.1038/s41559-023-02281-9



Extended Data Fig. 1 | Matrices created at each step of the workflow. Each grey box represents a matrix, with the number of cells indicated. Matrices were processed separately for each species ('Within-Species Analysis' - steps 1-4), then combined into a single multi-species matrix ('Cross-Species Analysis' - steps 5-6) and rotated (step 7) to produce a 919 cell 919 principal component matrix. The matrices used for Figures 2–5 were subsequently built by subsetting from a 919 cell 919 PC matrix. Note that the step 7 matrix is re-calculated for each figure.

The PCA (via the R function 'prcomp') is reproducible, but small differences are possible due to machine rounding error. Because of this, all input files for each figure are provided in the git repo (https://github.com/dunnlab/cellphylo) to ensure reproducibility. A step-by-step walk through to produce the results described by Figures 2, 4 and 5 are provided in the repo. The animal silhouettes were from PhyloPic (https://www.phylopic.org).

https://doi.org/10.1038/s41559-023-02281-9

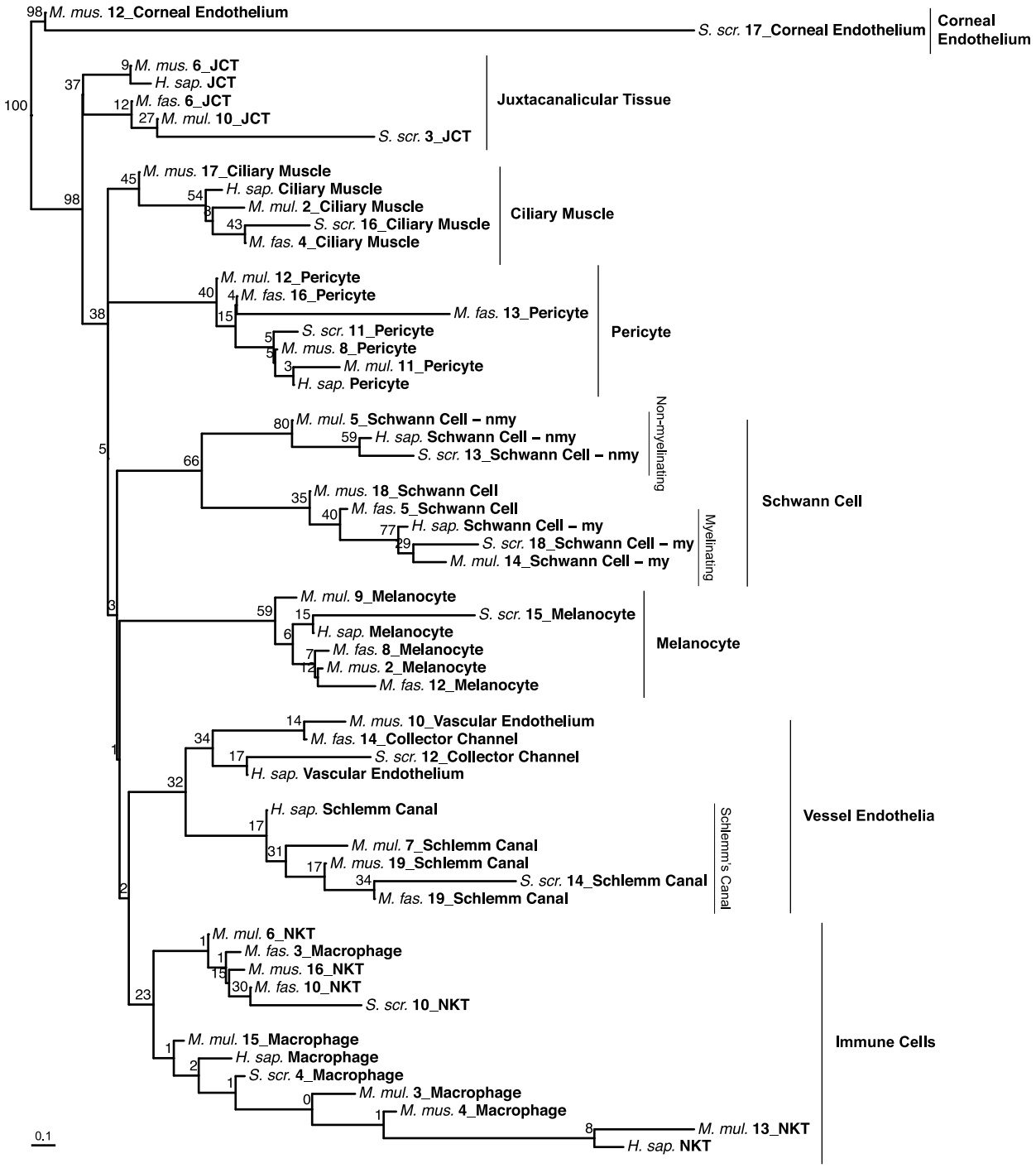


Extended Data Fig. 2 | Cells cluster into multi-species cell type clusters after cross-species integration. A UMAP plot of cells, colored by A. species identity and B. cell type identity. macF, *Macaca fascicularis*, macM, *Macaca mulatta*, CollectorChn, collector channel cell, CollectorChnAqVein, collector channel aqueous vein cell, Endo, endothelium, Endo-Corneal, corneal endothelium, Endo-Schlemms, Schlemm's canal endothelium, Endo-Vasc, vascular endothelium, Epi-CiliaryNonPigment, non-pigmented ciliary epithelium, Epi-CiliaryPigment, pigmented ciliary epithelium, Epi-Corneal, corneal epithelium, Epi-Pigment, pigmented epithelium, JCT, juxtacanalicular tissue, NKT, natural killer T cell, my, myelinating, nmy, non-myelinating.

https://doi.org/10.1038/s41559-023-02281-9

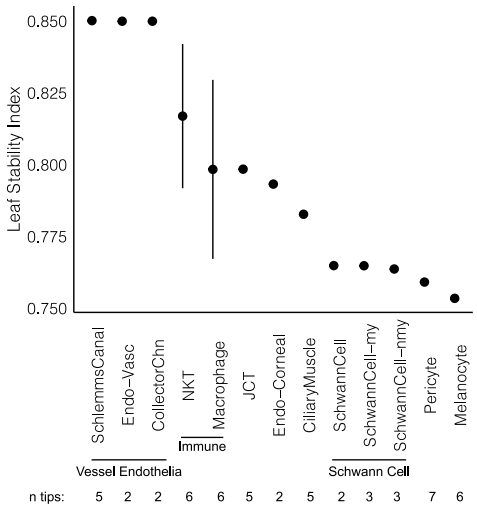
Extended Data Fig. 3 | A cell phylogeny of 92 aqueous humor cells. This phylogeny was created using the same methods as the 54 cell phylogeny of Figure 4, but was inferred from the 92 cell 20 PC matrix (Extended Data Fig. 1, step C1.1) that included one cell per cell type group per species for all 92 cell type groups. This matrix is more encompassing than the 54 cell 20 PC matrix (Extended Data Fig. 1, step C1.2), as it includes the unstable cell type groups that were excluded from the final analysis (Methods). Jumble scores are plotted at the nodes. The scale bar indicates units of expected evolutionary change. H. sap., *Homo sapiens*, M. fas., *Macaca fascicularis*, M. mul., *Macaca mulatta*, M. mus., *Mus musculus*, Sus. scr., *Sus scrofa*, my, myelinating, nmy, non-myelinating, JCT, juxtacanalicular tissue, NKT, natural killer T cell.

https://doi.org/10.1038/s41559-023-02281-9



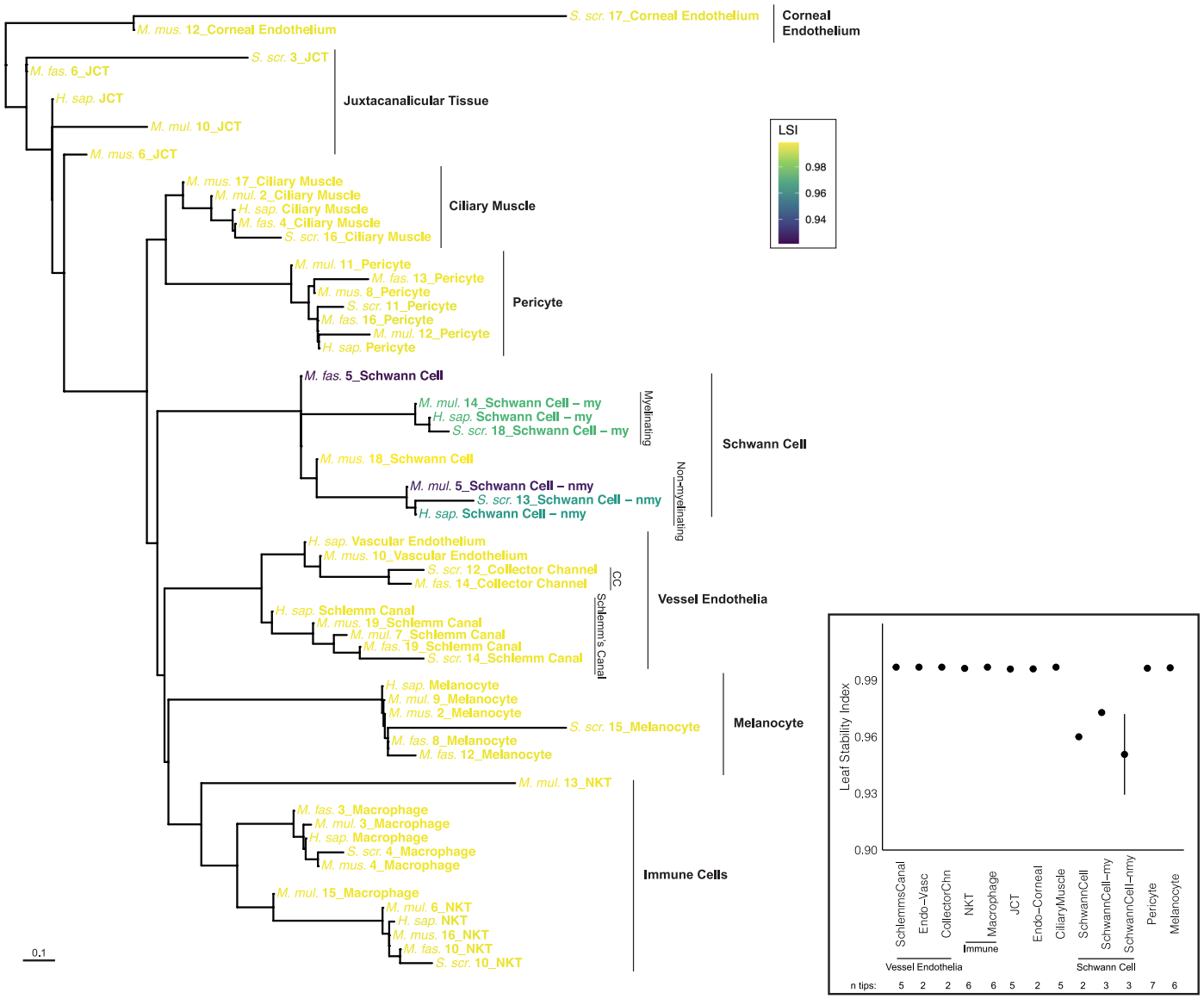
Extended Data Fig. 4 | Felsenstein's bootstrap is not a suitable measure of biological repeatability for a cell phylogeny. The 54 cell phylogeny (Fig. 4) is annotated with Felsenstein's bootstrap scores calculated from scjackknife trees. While scjackknife trees still produced cell type clades, cell clades among scjackknife trees frequently varied by a few cells. This subtle variability is not well captured by traditional bootstrap scores, which mark clades as present/absent based on the presence of all tips, without acknowledging the degree of similarity. Species and cell type groups are labeled at the tips; numbers refer to the cluster number of the cell type group. Cell type clades are indicated by vertical bars. The scale bar indicates units of expected evolutionary change. H. sap., *Homo sapiens*, M. fas., *Macaca fascicularis*, M. mul., *Macaca mulatta*, M. mus., *Mus musculus*, Sus. scr., *Sus scrofa*, JCT, juxtacanalicular tissue, my, myelinating, nmy, non-myelinating, NKT, natural killer T cell.

https://doi.org/10.1038/s41559-023-02281-9



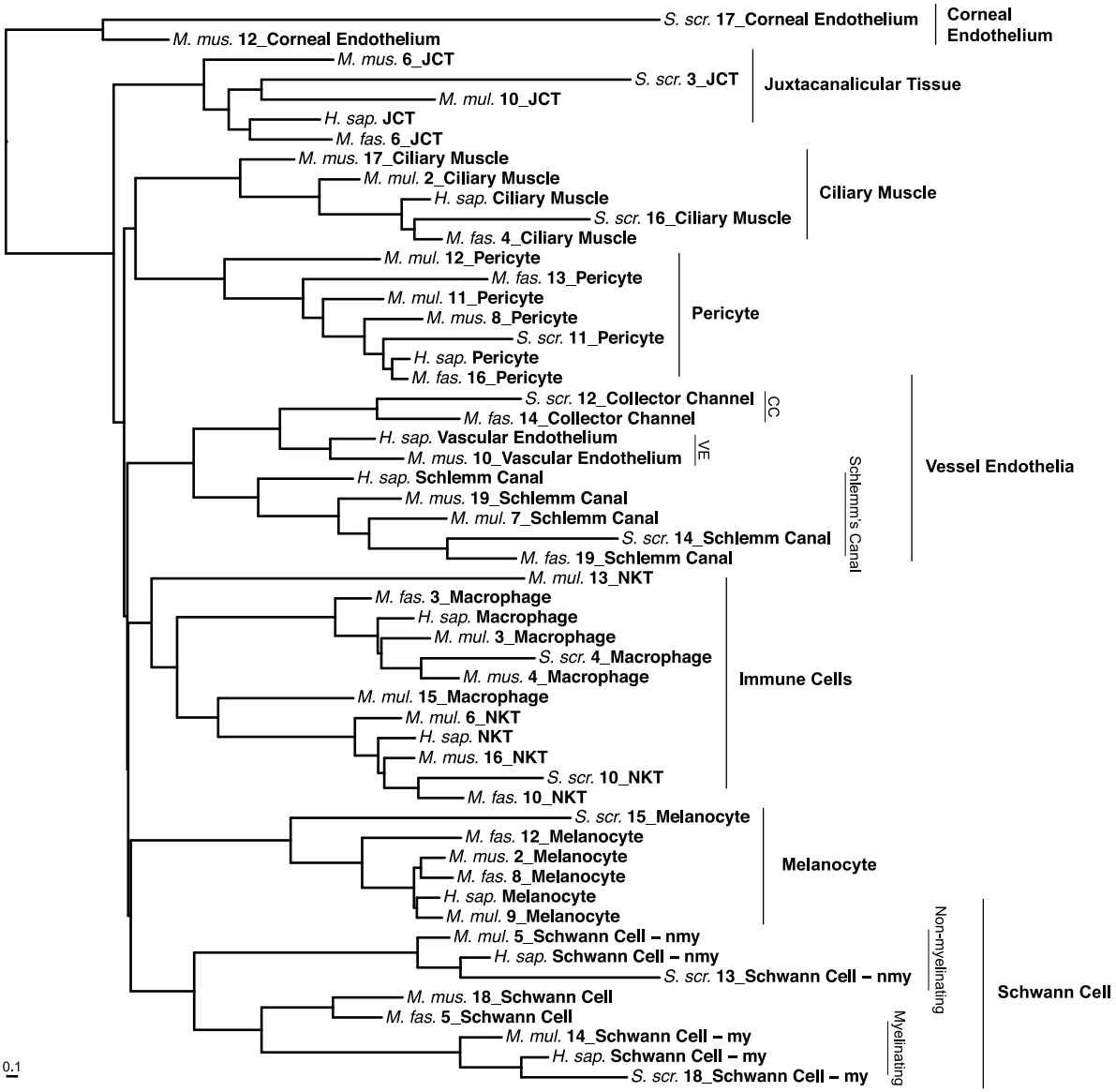
Extended Data Fig. 5 | Leaf stability index by cell type. Leaf stability indices (LSI) calculated from the Fig. 4 phylogeny are highly concordant with negligible spread within each cell type clade, with the exception of immune cells (NKT and macrophages). Mean LSI for each cell type is plotted and vertical lines indicate standard deviation. The number of tips per cell type label is indicated along the bottom as 'n tips'. The standard deviation for n tips < 3 is not shown. Cell type labels are labeled along the x-axis. Superclades are indicated with horizontal lines. Endo-Vasc, vascular endothelium, CollectorChn, collector channel cell, NKT, natural killer T cell, JCT, juxtacanalicular tissue, Endo-Corneal, corneal endothelium, my, myelinating, nmy, non-myelinating.

https://doi.org/10.1038/s41559-023-02281-9



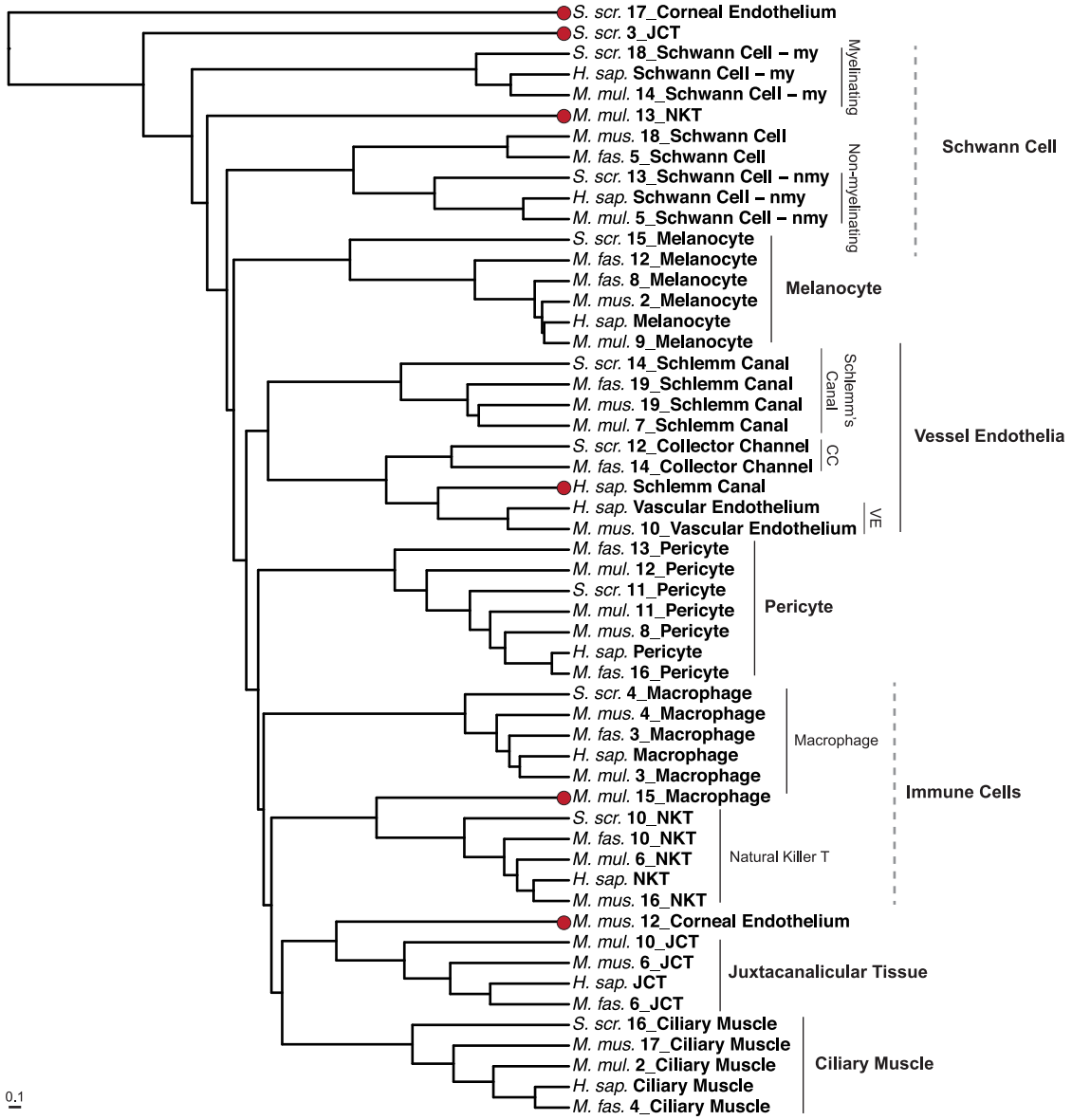
Extended Data Fig. 6 | The averaged tree exhibits high tip stability. The leaf stability index (LSI) is plotted as tip color onto the Fig. 5 tree. Most tips exhibit an LSI close to 1, the maximum score. Species and cell type groups are labeled at the tips, with numbers referring to the cluster number of the cell type group. Cell type clades are indicated by vertical bars. The scale bar indicates units of expected evolutionary change. Inset: Mean LSI for each cell type is plotted and vertical lines indicate standard deviation. The number of tips per cell type label is indicated along the bottom as 'n tips'. The standard deviation for n tips < 3 is not shown. Cell type labels are labeled along the x-axis. Superclades are indicated with horizontal lines. H. sap., *Homo sapiens*, M. fas., *Macaca fascicularis*, M. mul., *Macaca mulatta*, M. mus., *Mus musculus*, Sus. scr., *Sus scrofa*, my, myelinating, nmy, non-myelinating, JCT, juxtacanalicular tissue, CC, collector channel, NKT, natural killer T cell, Endo-Vasc, vascular endothelium, CollectorChn, collector channel cell, Endo-Corneal, corneal endothelium.

https://doi.org/10.1038/s41559-023-02281-9



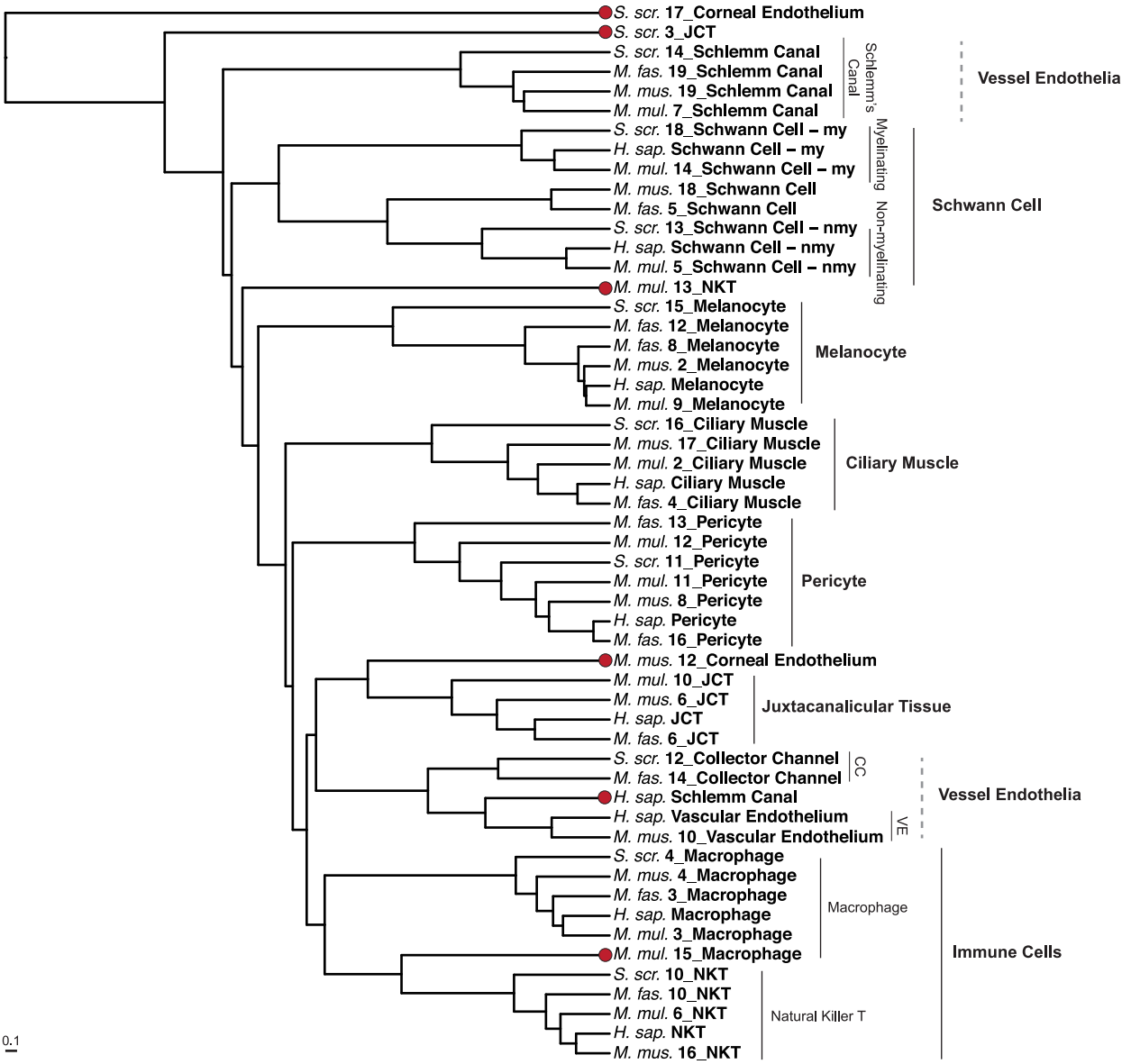
Extended Data Fig. 7 | Neighbour joining tree calculated from averaged expression levels. The neighbour joining method was used to calculate a phenetic tree from the averaged matrix (Extended Data Fig. 1, step D1) used to infer the Fig. 5 cell phylogeny. Cell type clades are indicated with vertical bars. Species and cell type groups are labeled at the tips, with numbers indicating the cluster number of the cell type group. H. sap., *Homo sapiens*, M. fas., *Macaca fascicularis*, M. mul., *Macaca mulatta*, M. mus., *Mus musculus*, Sus. scr., *Sus scrofa*, my, myelinating, nmy, non-myelinating, JCT, juxtacanalicular tissue, CC, collector channel, VE, vascular endothelium, NKT, natural killer T cell.

https://doi.org/10.1038/s41559-023-02281-9



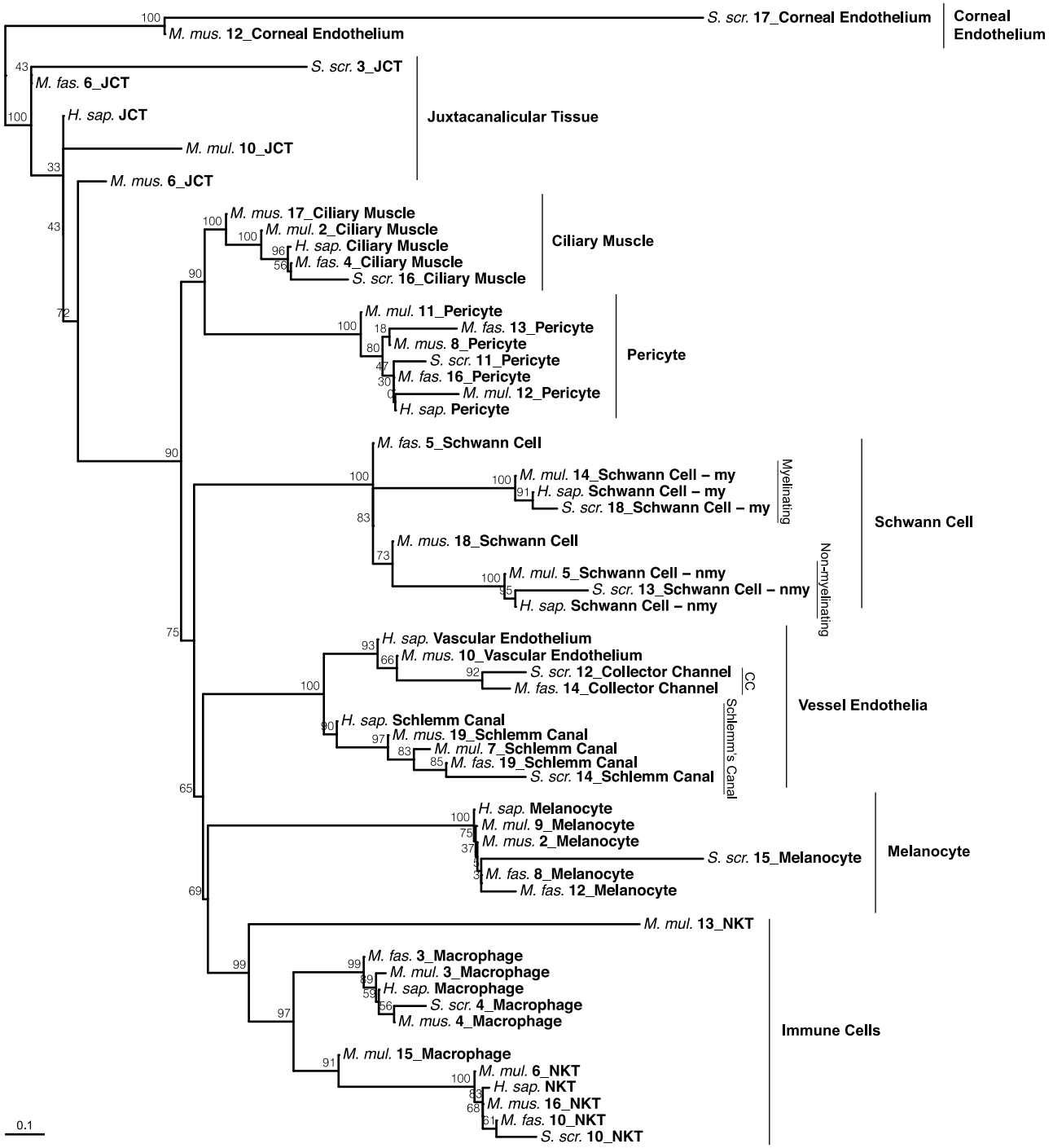
Extended Data Fig. 8 | UPGMA tree calculated from averaged expression levels. UPGMA was used to calculate a phenetic tree from the averaged matrix (Extended Data Fig. 1, step D1) used to infer the Fig. 5 cell phylogeny. Species and cell type groups are labeled at the tips, with numbers indicating the cluster number of the cell type group. Cell type clades are indicated with vertical bars. Red dots highlight cells that failed to group with their corresponding cell type clade. The Schwann cells and immune cells superclades are paraphyletic, as indicated with dashed lines. H. sap., *Homo sapiens*, M. fas., *Macaca fascicularis*, M. mul., *Macaca mulatta*, M. mus., *Mus musculus*, Sus. scr., *Sus scrofa*, my, myelinating, nmy, non-myelinating, JCT, juxtacanalicular tissue, CC, collector channel, VE, vascular endothelium, NKT, natural killer T cell.

https://doi.org/10.1038/s41559-023-02281-9



Extended Data Fig. 9 | WPGMA tree calculated from averaged expression levels. WPGMA was used to calculate a phenetic tree from the averaged matrix (Extended Data Fig. 1, step D1) used to infer the Fig. 5 cell phylogeny. Species and cell type groups are labeled at the tips, with numbers indicating the cluster number of the cell type group. Cell type clades are indicated with vertical bars. Red dots highlight cells that failed to group with their corresponding cell type clade. The vessel endothelia superclade is polyphyletic, as indicated with dashed lines. H. sap., *Homo sapiens*, M. fas., *Macaca fascicularis*, M. mul., *Macaca mulatta*, M. mus., *Mus musculus*, Sus. scr., *Sus scrofa*, my, myelinating, nmy, non-myelinating, JCT, juxtacanalicular tissue, CC, collector channel, VE, vascular endothelium, NKT, natural killer T cell.

https://doi.org/10.1038/s41559-023-02281-9



Extended Data Fig. 10 | The topologies of the cell phylogeny and neighbour joining tree are broadly similar. Neighbour joining was used to create scjackknife trees. The resulting scjackknife scores were plotted onto the Fig. 5 cell phylogeny. There is broad agreement at nodes defining cell type clades and the relationships between them. H. sap., *Homo sapiens*, M. fas., *Macaca fascicularis*, M. mul., *Macaca mulatta*, M. mus., *Mus musculus*, Sus. scr., *Sus scrofa*, my, myelinating, nmy, non-myelinating, JCT, juxtacanalicular tissue, CC, collector channel, NKT, natural killer T cell.

nature portfolio

Corresponding author(s): Jasmine Mah

Last updated by author(s): Nov 11, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	A mix of custom code and PHYLIP (v. 3.698; Felsenstein, retrieved from: https://phylipweb.github.io/phylip/) was used for the analysis. All custom code is written in R (v. 4.2.2) or bash (v. 5.2.15(1)-1) and is available in the git repo: https://github.com/dunnlab/cellphylo. Version numbers for all R packages are listed in the `sessionInfo_R.txt` file in the repo. All code required to reproduce the results of the paper is described in the R notebooks provided in the `analysis` folder: 1_Wrangle_Data.Rmd, 2_Create_Matrices.Rmd, 3_PC_sweep_analysis.Rmd, 4_Jumble_analysis.Rmd, 5_scJackknife_analysis.Rmd, 6_Average_cells.Rmd, 7_GO_and_phenetic_trees.Rmd

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data used for analyses is publicly available and was originally collected and published by van Zyl et al. 2020 (PNAS: 117 (19) 10339-10349, doi: 10.1073/pnas.2001250117). We are unaware of any restrictions. The count matrices published by van Zyl et al. were downloaded from NCBI GEO (GSE146188) and a meta data file for the study was obtained from the Broad Institute Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP780). The accession number for each count matrix is: GSE146186 (Mus musculus), GSE146187 (Sus scrofa), GSE148371 (Homo sapiens), GSE148373 (Macaca fascicularis), and GSE148374 (Macaca mulatta). In addition, select input, intermediate and output files sufficient to reproduce the analyses have been deposited in a git repo (https://github.com/dunnlab/cellphylo).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This information has not been collected as it is not relevant to our study.
Reporting on race, ethnicity, or other socially relevant groupings	This information has not been collected as it is not relevant to our study.
Population characteristics	This information has not been collected as it is not relevant to our study.
Recruitment	This information has not been collected as it is not relevant to our study.
Ethics oversight	This information has not been collected as it is not relevant to our study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study presents an approach to building cell phylogenies from comparative, single-cell RNAseq datasets using an explicit model of evolution. We examine the challenges inherent to applying evolutionary models to gene expression data and find a solution by performing rotation and rank reduction of the data with PCA. We use the resulting principal components (PCs) as phylogenetic characters. We perform a sweep analysis investigating the effect of trimming principal components and find that the evolutionary signal is present in the first PCs. We infer a cell phylogeny from a large single-cell RNAseq dataset obtained from five model species and find that robust cell type clades emerge, allowing us to define "cell type" by phylogenetic topology. We perform a "jumble analysis" by varying the initial starting tree of the maximum likelihood search. After processing, we produced datasets of 237 (Fig. 4) and 209 (Fig. 5) jumble trees from which we identified our focal phylogeny in Figures 4 and 5 and calculated technical repeatability scores ("jumble scores"). We also performed a "single-cell jackknife analysis" where we repeatedly resampled the cells at the tips of the cell phylogeny. This produced datasets of 439 (Fig. 4) and 393 (Fig. 5) 'scjackknife' trees, from which we calculated biological repeatability scores ("scjackknife scores"). The cell phylogeny was inferred at the cell-level (Fig. 4) and from averaged values (Fig. 5). The topology of these two cell phylogenies were consistent with known evolutionary developmental relationships between cell types. We also examine the gene expression dynamics underlying patterns of broad variation across the cell phylogeny by investigating the most highly loaded genes contributing to the PCs using gene ontology enrichment analysis. We find cell type-specific signal among these genes. Finally, we make a comparison of the cell phylogeny (Fig. 5) to phenetic trees built using distance-based methods (Extended Data Figs. 6-9).
Research sample	We used an existing scRNAseq dataset produced by van Zyl et al. 2020 (PNAS: 117 (19) 10339-10349, doi: 10.1073/pnas.2001250117), accessible at GSE146188. The accession number for each count matrix is: GSE146186 (Mus musculus), GSE146187 (Sus scrofa), GSE148371 (Homo sapiens), GSE148373 (Macaca fascicularis), and GSE148374 (Macaca mulatta). These

	datasets are single cell transcriptomes produced by sequencing cells present in the aqueous humor of the eye and surrounding eye tissue (van Zyl et al.).
Sampling strategy	We performed runs to infer 250 jumble trees (reduced to 237 after processing) and 500 scjackknife trees (439 after processing) to calculate the repeatability scores for Figure 4. For Figure 5, we performed runs for 500 jumble trees (209 after processing) and 500 scjackknife trees (393 after processing) for Figure 5. As we could not predict the number of trees that would remain after processing we selected the number of runs of sufficient size to produce enough trees for informative scores.
Data collection	Publicly available data (GSE146188) produced by van Zyl et al. 2020 (PNAS: 117 (19) 10339-10349, doi: 10.1073/pnas.2001250117) was used for our analyses. The accession number for each count matrix is: GSE146186 (Mus musculus), GSE146187 (Sus scrofa), GSE148371 (Homo sapiens), GSE148373 (Macaca fascicularis), and GSE148374 (Macaca mulatta). Detailed descriptions of the data collection procedure can be found in the original paper by van Zyl et al.
Timing and spatial scale	This information has not been collected as we did not collect new data.
Data exclusions	We excluded unstable cell types (similar to rogue taxa) and cell types present only in a single species ("singletons") from the final cell phylogenies (Fig. 4 and 5). Unstable cell types were identified after an initial analysis on a larger dataset containing all cell types, where these "rogue" cell types consistently failed to form clades and destabilized the topology (Extended Data Fig. 3). These unstable cell types were the corneal epithelium, beam A, beam X, fibroblast and B cells. We excluded singletons because they are difficult to meaningfully interpret in the context of other cell type clades.
Reproducibility	All custom code required to reproduce the analyses are given in the git repo: https://github.com/dunnlab/cellphylo. The code for specific figures and results are described in R notebooks in the `analysis` folder: 1_Wrangle_Data.Rmd, 2_Create_Matrices.Rmd, 3_PC_sweep_analysis.Rmd, 4_Jumble_analysis.Rmd, 5_scJackknife_analysis.Rmd, 6_Average_cells.Rmd, 7_GO_and_phenetic_trees.Rmd. In addition select input, intermediate and final output files are provided in the `analysis` directory to facilitate full reproduction of the analyses. All attempts to repeat the analyses successfully produced results consistent with our conclusions.
Randomization	Randomization was used to subset matrices and perform the jumble tree and scjackknife analyses. The R function base::sample() was used to perform random subsetting of matrices and resampling of cells to produce the scjackknife matrices. Random seeds required for input to contml for the jumble analysis were created with `$(shuf -i 1-2000000000 -n 1)` in a bash script (see `cellphylo/analysis/scripts/jumble/script_jumble_parallel_54.sh` in the git repo).
Blinding	This information has not been collected as it is not relevant to our study. We did not collect any new data.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging