Trends in Genetics

CellPress

Review

Evolutionary Cell Type Mapping with Single-Cell Genomics

Amos Tanay[1,*] and Arnau Sebé-Pedrós[2,3,*]

A fundamental characteristic of animal multicellularity is the spatial coexistence of functionally specialized cell types that are all encoded by a single genome sequence. Cell type transcriptional programs are deployed and maintained by regulatory mechanisms that control the asymmetric, differential access to genomic information in each cell. This genome regulation ultimately results in specific cellular phenotypes. However, the emergence, diversity, and evolutionary dynamics of animal cell types remain almost completely unexplored beyond a few species. Single-cell genomics is emerging as a powerful tool to build comprehensive catalogs of cell types and their associated gene regulatory programs in non-traditional model species. We review the current state of sampling efforts across the animal tree of life and challenges ahead for the comparative study of cell type programs. We also discuss how the phylogenetic integration of cell atlases can lead to the development of models of cell type evolution and a phylogenetic taxonomy of cells.

Highlights

Cell types are the basic functional units of multicellular organisms. So far, cell taxonomies remain Linnaean.

Single-cell transcriptomic methods enable the systematic characterization of cell diversity in understudied animal lineages. These data-driven cell catalogs should allow us to organize cell diversity into evolutionary classification schemes.

Single-cell sampling bias and other technical limitations can severely constrain our ability to integrate cell atlases across species. It is important to advance towards general sampling strategies and data standards.

Comparative analysis of cell type atlases uncovers transcriptional similarities across species, but are confounded by pleiotropy and non-independence of gene expression patterns, particularly at large phylogenetic distances.

The interrogation of cell type regulatory programs in closely related species should enable the development of quantitative models of cell type evolution and to assess the (potentially incongruent) evolutionary histories of the different components of these programs.

Specialized cells represent the fundamental level of organization in multicellular organisms [1]. The morphological and molecular regularities observed in cells have inspired analogies to the diversity of organisms and their hierarchical arrangement into different taxa. This analogy suggested that cellular taxonomy could be developed following similar principles to those underlying Linnaean species classification, with the existence of predefined, staggered ranks (but ignoring evolutionary relationships; Box 1). Indeed, we can characterize a cell type as a discrete entity that has unique morphological and functional properties (Box 2). We can also require a cell type to be reproducible – that is, to emerge stably across generations through embryonic development. Nevertheless, the hierarchical nature of cell types and the discrete nature of their classification remain more elusive: ontogeny and cell lineages within organisms are the major cell type-organizing forces, but these are remodeled continuously by the plasticity and **pleiotropy** (see Glossary) of gene regulatory programs across tissues. Because cell types are natural building blocks bridging molecular (gene level) and organismal (phenotypic) evolution, there is great interest in studying cell types as evolutionary units [2]. To this end, molecular profiling tools, particularly single-cell transcriptomics, hold the promise to bring cell type molecular phenotyping and classification to non-model species by building systematic atlases of cells in different animal lineages. Single-cell atlases not only can advance our understanding of the molecular and cellular biology of understudied animal groups but are also the necessary first step towards a comparative biology of cell type programs. Only through these cell type comparisons can we eventually understand and reconstruct cell type evolution.

In this review we examine recent advances in single-cell atlas building in non-traditional model species, and then discuss the challenges and opportunities ahead in the phylogenetic expansion of cell type mapping across the animal tree of life.

[1]Department of Computer Science and Applied Mathematics, and Department of Biological Regulation, Weizmann Institute of Science, 76100 Rehovot, Israel
[2]Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
[3]Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain

*Correspondence:
amos.tanay@weizmann.ac.il (A. Tanay) and arnau.sebe@crg.eu (A. Sebé-Pedrós).

Box 1. Linnaean Cell Type Classification

Linnaeus formalized a taxonomic classification system aimed at grouping, organizing, and naming organisms. Central to the Linnaeus system is the existence of predefined, staggered taxon ranks (kingdom, phylum, class, order, family, genus, species) from which a binomial (genus–species) nomenclature system is derived. Although this binomial naming is still widely used in taxonomy, modern systematics uses phylogenetic methods (very often molecular phylogenetics) to derive classifications that are explicitly based on evolutionary relationships. In addition, this phylogenetic systematics (also known as cladistics) often does not necessarily adhere to rigid, staggered ranks as universal principles, and instead focuses on the tree-like organization of species.

By analogy, cell type taxonomies remain Linnaean in the sense that they do not incorporate evolutionary considerations – whether or not implying the existence of staggered ranks, for example broad cell type identities [17]. In fact, cell type classifications very often include hierarchical tree representations based on, for example, transcriptome similarities. However, these trees do not (necessarily) represent any historical relationships (either ontogenetic or phylogenetic), in the same way as Linnaean ranks often do not conform to our current understanding of species phylogeny. A cladistic/phylogenetic cell type classification would incorporate information about the evolutionary relationships between cell types. However, to develop such cell type cladistics we will need cell type phylogenetic methods – ways to trace the evolutionary relationships and histories of cell types across species and phyla.

Cell Taxonomy

Cell type classification schemes vary in their granularity and in the degree of phylogenetic and anatomical generalization. That is, classifications may encompass only particular organs/species or represent phylogenetically and anatomically (even organism-level) wider frameworks [3]. Most proposed cell type classifications are hierarchical and, with few exceptions [4], use concepts and jargon borrowed from taxonomy (clades, lineages, trees, etc.), although they do not explicitly consider or try to convey evolutionary relationships between cell types [5].

From a historical perspective, the first efforts in cell type classification were based on the morphology, spatial tissue arrangement, cellular connectivity, and histological staining properties of cells. Using this information, multiple attempts were made in the pre-genomics era to develop global cell classification schemes [3], to systematically characterize cell types in specific taxa [6–8], and to use cell type number as a proxy for organismal complexity [9]. These classification frameworks were restricted in resolution and could not take into account functional or developmental considerations that are not readily represented morphologically. The advent of molecular profiling tools has extended the ability to characterize, identify, and classify cell types. Common strategies include detecting specific proteins, using antibody-based immunostaining (surface markers are still widely used for the molecular phenotyping of hematopoietic and immune cells [10]), or specific transcripts using RNA *in situ* hybridization (more rarely also by qPCR analysis [11]). While immunostaining is strongly constrained by the limited availability of antibodies, *in situ* hybridization with custom-synthesized probes has enabled the rapid extension of molecular profiling to a wide diversity of organisms, thus becoming a cornerstone of modern evo-devo studies [12–14]. A major limitation of these expression profiling tools is the limited scalability to dozens of markers and, most importantly,

Box 2. Broad-Sense and Narrow-Sense Cell Types

A cell type in the narrow sense is a subset of the cells within an organism that share (i) morphological properties, (ii) spatial tissue distribution or motility characteristics, (iii) signaling and metabolic input and output, and, in the case of progenitor cell types,(iv) differentiation and proliferation potential. The degree of homogeneity for each of these categories may vary and is currently defined *ad hoc*. Moreover, cell types may or may not be ontogenically coherent because they often emerge from different precursor developmental lineages. A cell type in the broad sense is a distribution of intracellular molecular properties such as transcriptional levels, protein abundance, or epigenomic landscapes. The distributions of cell type molecular properties are expected to be stable over timescales that exceed a cell cycle, and do not include the effect of transient fluctuations (e.g., circadian rhythm, physiological cycles, stress responses). Such broad-sense cell types can be defined based on unimodal or multimodal molecular atlases and, in many cases, they can be expected to approximate and refine narrow-sense cell types.

Glossary

Ambient RNA: RNA molecules released from bursting or leaky damaged cells. These RNAs can be biased for a specific cell type (e.g., epidermal cells are particularly sensitive) and will be captured during single-cell library preparation, representing a constant background in each single-cell transcriptome. Ambient RNA not only complicates single-cell RNA sequencing (scRNA-seq) analysis but is suspected to be a major source of batch effects across experiments.

Character state: in taxonomy and phylogenetics, a specific, form, category, or quantitative value in which a character (a heritable trait) is present in an organism. The uniform encoding of character states is important in comparative biology. In the case of cell type program comparisons, character states can be gene expression levels (either quantitative or binarized), sequence motif enrichments, regulatory element usage, etc.

Evolutionary rate: a parameter to describe the dynamics of change of a particular character, generally or in a specific lineage. For example, changes in DNA or protein sequence or divergence in gene expression levels [107].

Gene module: a set of genes that are coexpressed in a coordinated fashion in one or more cell types. A gene program could typically be associated with a specific biological function, which may or may not be reused in different cellular contexts. A program is regulated by one or more transcription factors (TFs).

Metacell: a group of single-cell genomic measurements that can be modeled mathematically as resampling of the same idealized cell. The combined statistics of a metacell can quantitatively define the molecular state of a cell by neutralizing single-cell sparse-sampling noise and stochastic, unregulated variation in cellular molecular composition.

Orthologs/paralogs: a pair of genes in two species are orthologs if they originated through speciation, and were therefore present as a single gene in the common ancestor of the two species. By contrast, paralogs are genes originating from a duplication event. When referring to paralogs within a species the term 'in-paralogs' is often

the need to define *a priori* the set of gene markers to study. This selection of marker genes is phylogenetically biased by previous studies on model species, and therefore ignores clade-specific mechanisms. Molecular profiling strategies have been very effective when used in the comparative study of embryogenesis and tissue/organ-level anatomical structures [15,16]. However, mapping the diversity of cell types across species in a truly systematic fashion was so far not possible.

An extension of the candidate marker gene profiling is the analysis of genome-wide gene expression using bulk transcriptomics. Pioneering studies in cell type bulk transcriptomics has provided systematic cell classification schemes [17], revealed the hierarchical structure of cell type transcriptomes [18], and enabled the first attempts at building phylogenies of closely related cell types based on their gene expression profiles [19]. However, these enrichment strategies are necessarily limited to cell lines [17,20–22] – that are virtually non-existent in the vast majority of organisms – or homogeneous cell populations isolated manually or by fluorescence-activated cell sorting (FACS) [10,23,24] – which require dedicated methods and do not ensure purity. High-throughput single-cell RNA sequencing methods (scRNA-seq) overcome many of these limitations, ultimately facilitating the minimally biased sampling and molecular characterization of thousands of single cells, and setting the stage for *in silico* reconstruction of cell type repertoires in species that were so far difficult to study.

Overall, the development of cell type classification tools is in a way analogous to that of species phylogenetic methods: from morphological to molecular characters (nucleotide or protein sequences). In the same way that it is difficult to resolve species phylogenies using morphological characters alone, only with molecular data can we aim at developing phylogenetically inclusive cell type classification schemes. However, the analogy ends here: modern taxonomy is explicitly based on the underlying evolutionary history of a species, very often incorporating molecular phylogenetics as a key tool for classification. Cell type taxonomies remain, to date, fundamentally Linnaean.

Animal Cell Type Mapping – Phylogenetic State of the Art

Since the first proof-of-concept scRNA-seq studies in the early 2010s ([25–27]; reviewed in [28]), we have witnessed the rapid proliferation of scRNA-seq analyses, with ever-growing numbers of cells and moving from descriptive cell type phenomenologies to perturbation assays, development and temporal differentiation dynamics, and spatial transcriptomics with single-cell resolution [29]. Today, cataloging the full repertoire of cell type programs in human tissues and development seems to be within reach [30,31], and important progress has already been made in cataloguing mouse cell types [32,33]. Applied to non-traditional model species, whole-organism scRNA-seq methods should pave the way to the systematic characterization and comparison of cell types across the animal and, consequently, can rapidly advance our understanding of cell type diversity, development, and evolution [34]. In addition, studying the spatial arrangement of these cell types with the emerging spatial transcriptomic technologies [29] can help to refine these cell atlas models.

Given a minimally biased single-cell sampling strategy (next section), we can use standardized pipelines to generate gene expression profiles for thousands of cells and to group such profiles into discrete, highly similar, transcriptional cell states. These data-driven cell groups/clusters constitute basic units that can be further developed, through biological interpretation, into cell type classification schemes. Following up on the phylogenetics analogy, scRNA-seq methodologies can have an impact on the study of cell type diversity and evolution analogous to the impact of whole-genome/transcriptome sequencing techniques on the resolution of the animal tree of life.

used, whereas paralogs in different species are called 'out-paralogs'. A gene that is orthologous to a group of paralogs in another species is often call a 'co-ortholog'. In addition, paralogs specifically originating through genome duplication are called 'ohnologs', whereas those resulting from hybridization events are termed 'homoeologs'. Further details on gene homology definitions are given in [109,110].

Pleiotropy: a single gene influencing multiple phenotypic traits. In the context of cell type programs, pleiotropic genes (and gene modules) are expressed in different cell types, and there are similarly pleiotropic *cis*-regulatory sequences.

Unique molecule identifiers (UMIs): short random sequences that are incorporated into transcripts during reverse transcription to accurately measure transcript molecular counts in scRNA-seq analysis. UMIs are aimed at mitigating the effect of library amplification bias in scRNA-seq protocols.

From a taxonomic perspective (Figure 1), whole-organism cell type atlases are currently available for seven major animal lineages (in most cases represented by a single species), including a ctenophore [35], two sponges [35,36], a placozoan [35], five cnidarians [37–40,113], an acoel [41], craniates (considering mouse whole-organ single-cell transcriptomes [32,42]), and platyhelminths [43–45]. Tissue-specific single-cell atlases are already available for model species, including multiple *Drosophila melanogaster* (Arthropoda) datasets [46–51] and *Caenorhabiditis elegans*

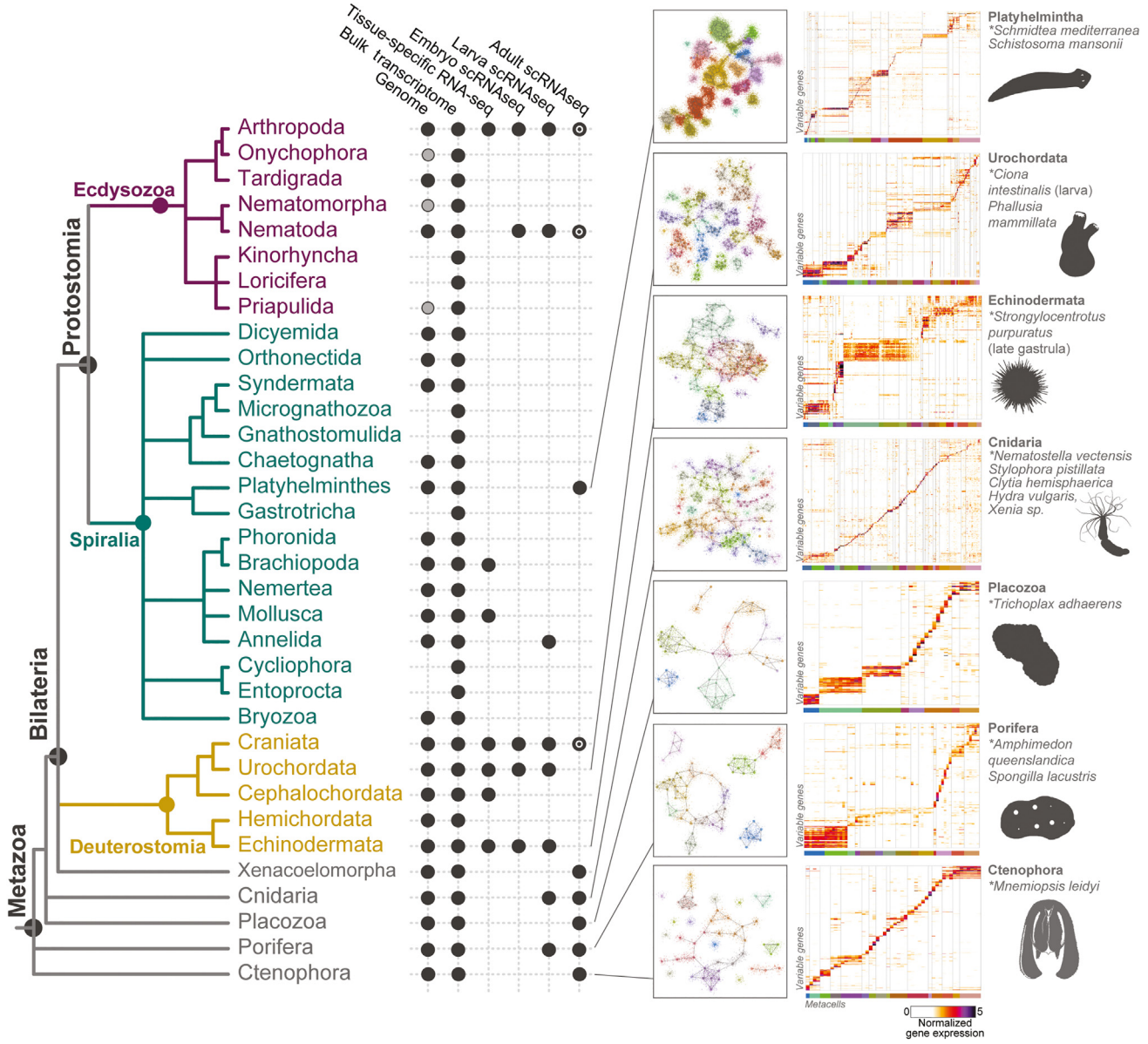

Figure 1. Single-Cell Genomics across Metazoa. (Left) The availability of genome sequences and/or bulk transcriptomes across major animal lineages, as well as embryonic, larval, and whole-adult single-cell atlases. Light-grey dots indicate ongoing genome projects; concentric circles indicate sampling of specific organs/tissues. (Right) Examples of seven whole-adult/larva cell type atlases in non-traditional model species are shown, including a 2D projection of metacells and the expression of highly variable genes across metacells. Asterisks indicate the species shown; colors are arbitrary and highlight similar metacells. Data from [35,38,44,53,56]; animal silhouettes from phylopic.org. Abbreviation: scRNA-seq, single-cell RNA sequencing.

(Nematoda) neuronal single-cell analyses. In addition, embryonic and larval stages have been sampled in two sea urchin species (Echinodermata) [52–54], in the marine polychaete *Platynereis dumerilii* (Annelida) [55], in the tunicates *Ciona intestinalis* [56] and *Phallusia mammillata* (Urochordata) [57], and again both in *D. melanogaster* and *C. elegans* [58,59]. The vertebrates are the most densely sampled lineage, including developmental single-cell atlases in four species (human, mouse, zebrafish, and *Xenopus*) [31,42,60–63] and brain single-cell data (in most cases for specific brain regions) for several mammals, reptiles, and teleosts [64–66].

It is interesting to compare the phylogenetic expansion of single-cell atlases to that of reference genomes over the past 5 years. In 2015, 15–17 years after the publication of the first animal draft genomes, Dunn and Ryan [67] reviewed the status of genome sequencing across animal lineages. By then, 212 genomes from 14 animal phyla were available at the US National Center for Biotechnology Information (NCBI). Today, 11 404 genomes from 27 animal phyla (Figure 1) are available or in progress, and many more are expected in the near future in the context of large-scale biodiversity sequencing initiatives such as the Darwin Tree of Life project and the Vertebrate Genomes project [68]. Importantly, taxon sampling biases persist, and 86% of these genomes are from vertebrates (6454) and arthropods (3383) – 95% if we include mollusks (518) and nematodes (253). By comparison, since the publication of the first high-throughput single-cell transcriptomics datasets in about 2015 [27,69–71], whole-organism single-cell atlases have been published for 13 non-model animal species. Given the fast pace in the scale and sophistication of single-cell methods in model species (and the wide availability of commercial solutions), the current taxonomic sampling of single-cell atlases across the animal tree of life seems rather modest.

A Natural History of Animal Cell Types – Challenges Ahead

There are several technical limitations that can explain the limited success of scRNA-seq analyses in different animal groups. We highlight here some of the challenges based on the initial experience of whole-organism single-cell atlasing in non-model species.

Single-Cell Sampling

Much attention has been devoted to the molecular biology (reverse transcription, amplification strategies, multiplexing, etc.) and physical implementation (droplets, microwells, etc.) of scRNA-seq. However, the dissociation and handling of single cells from intact organisms remains the main challenge in scRNA-seq analysis for non-model species. This not only affects the quality of the scRNA-seq data but can also introduce important biases against particular cell types, whose existence may remain hidden when working with previously unsampled organisms ('unknown unknowns'). For example, epidermal cells are considered to be particularly sensitive to dissociation, and the same is suspected for cnidocytes in cnidarians. How then do we approach a completely novel species to obtain the most complete and minimally biased single-cell atlas? There is no definitive answer to this question yet, but there are important considerations to take into account.

Tissue/organism dissociation generally involves an enzymatic digestion step. Digestion is limited to the shortest possible time (usually <30 minutes) because very aggressive dissociations can induce cell death/stress [72]. Several proteases have been successfully used, typically including trypsin, collagenases, chymotrypsin, cold-active proteases, and others. Digestion is followed by an evaluation of induced cell death (generally aiming at less than 5–10%) before single-cell transcriptome capture. In marine organisms, this strategy is usually combined with the use of calcium/magnesium-free seawater (CMFSW), and in some cases this removal of divalent cations is sufficient to effectively dissociate animals, as in some sponges and placozoans [35].

The 'fresh dissociation' strategy outlined previously has several downsides. First, although remaining tissue chunks are removed by serial filtering, cell doublets and triplets can be abundant and confound scRNA-seq data analysis. A solution to this is to use FACS to discard dead cells and cell doublets, as well as non-cellular debris. Coupled with direct cell lysis into multiwell plates, this strategy has been successful in several marine species [35,38]. A second problem is the differential sensitivity of cells to the dissociation treatments: some cells lyse early, whereas other may require more aggressive dissociation, and similarly some cells may not survive a FACS-sorting procedure (if not captured directly after sorting). Another side-effect of fresh dissociations is that **ambient RNA** is released by lysing cells (sometimes also called 'free-floating RNA'). This ambient RNA will be captured together with each single-cell transcriptome and, although methods to model this effect have been developed [73,74], high levels of ambient RNA can confound the analysis of complex atlases, in particular when noisy 'cells' can be misinterpreted as transitional states between cell types.

A potential solution to some of these problems is to immediately fix cells after dissociation, preventing any further transcriptional changes. Methanol fixation has been successfully used in different species [37,45,75], and mild formaldehyde fixations have also been used [76]. Fixed cells can be washed to reduce ambient RNA and, importantly, can be FACS-sorted and subsequently encapsulated by different methods. The downsides of fixation include reduced sensitivity (less molecules detected per cell) and cell loss during the fixation process. Furthermore, fixation does not prevent any sampling biases introduced during dissociation. To tackle this problem, Garcia-Castro *et al.* recently introduced a novel method (based on acetic acid, glycerol, and methanol; ACME) for simultaneous cell dissociation and fixation compatible with scRNA-seq [77]. An important additional consideration when dealing with marine organisms is osmotic stress. Freshly dissociated cells should be kept in marine water equivalents (such as CMFSW), but the high salt concentrations inhibit the reverse transcriptase, making it incompatible with scRNA-seq methods. Fixation prevents cells from bursting or developing osmotic shock responses.

Another alternative is to completely avoid cell dissociation and instead focus on the analysis of single nuclei extracted from (usually flash-frozen) whole tissues, such as brain [78,79] and muscle fibers [80]. This strategy minimizes any alterations introduced by dissociation, and the use frozen tissues conveniently allows specimen sampling to be decoupled from single-cell processing – an important consideration for species that may not be readily maintained in the laboratory. Single-cell transcriptomes derived from nuclei show enriched intronic content (pre-spliced mRNAs) and a slight bias towards longer transcripts (hypothesized to be due to nuclear export dynamics), but overall the recovered cell type transcriptional profiles are highly concordant [81]. The main downside of single-nucleus RNA-seq (snRNA-seq) is an important loss of sensitivity relative to scRNA-seq because the nucleus contains only a fraction of the mRNAs of the whole cell. Interestingly, a recent snRNA-seq method improved sensitivity by a novel nuclear preparation method that retains the endoplasmic reticulum and numerous attached ribosomes [82]. Given the advantages of avoiding aggressive cell dissociations and compatibility with sample freezing, snRNA-seq may represent an alternative worth considering for sampling cell types in non-model species with minimal biases.

Overall, we are still far from an ideal, quasi-universal sampling strategy for whole-organism single-cell analysis. Many of the challenges in studying organisms from the field (including sample transport and preservation) are similar to those faced by clinical studies with patient samples [83], and therefore methods and protocols developed in this context can be highly relevant for cell atlasing in non-model species. In any case, it is important to emphasize the need to optimize sampling

(examining dissociation conditions, checking cell survival, measuring ambient RNA, etc.), particularly when working with species with little or no standard handling and processing protocols.

Data Analysis and Interpretation

The analysis and interpretation of whole-organism scRNA-seq data in non-model animal species present specific challenges. A complication not directly related with scRNA-seq data is the quality of the reference genome for the species under study. For the purpose of scRNA-seq data mapping, even relatively non-contiguous, fragmentary assemblies can suffice, but scRNA-seq is dramatically affected by missing or inaccurate gene annotations. Gene models are usually built using a combination of direct RNA-seq evidence and *de novo* prediction [84]. However, gene annotation pipelines often fail to model genes that are very poorly expressed and/or are very short. The regulatory peptides in the placozoan *Trichoplax adhaerens* provide an interesting example. These very short and (globally) poorly expressed genes were missed in the original gene annotation [85], and were then manually annotated in another study [86]. scRNA-seq analyses revealed that these secreted regulatory peptides [87] are very highly expressed in specific, low-abundance cell types, which explains why they were barely detected in bulk RNA-seq. Another common problem, particularly in small, gene-dense genomes, is incorrectly fused gene models which can effectively mask the detection of poorly expressed genes 'fused' to highly expressed genes. scRNA-seq data can be interrogated for evidence of such fusions. Finally, the most common gene annotation artifact are incorrect 3′-untranslated region (UTR) annotations. Given that the majority of scRNA-seq methods detect the 3′-ends of transcripts, this systematic bias in transcription end-site (TES) annotation can have dramatic effects on gene detection [38,88]. In this case, scRNA-seq data (which are, conveniently, strand-specific) can be used to reannotate TES. As a final note, it is possible to analyze scRNA-seq based on *de novo* assembled transcriptomes alone. Although the same problems of poorly expressed (or highly expressed but in a rare cell type) transcripts apply here, one can resolve a single-cell atlas with a deeply sequenced bulk transcriptome as a reference [36,37]. However, working with a reference genome is essential to gain insights into cell type programs beyond gene expression – that is, to interrogate the *cis*-regulatory elements underlying cell type transcriptional phenotypes (see following text).

A second challenge with non-model organism scRNA-seq data analysis is that very often the median number of transcripts detected per cell is low [<1000 **unique molecule identifiers (UMIs) per** cell]. Although this can be partially due to technical factors, there are biological differences between species/lineages regarding the median transcriptional output of a differentiated cell. For example, four species from different animal lineages sampled with the same technology (massively parallel single-cell RNA-sequencing, MARS-seq) resulted in radically different median transcript counts per cell [35,38]: ~950 in the cnidarian *Nematostella vectensis*, ~550 in the placozoan *Trichoplax adhaerens*, ~1300 in the ctenophore *Mnemiopsis leidyi*, and >4500 in the sponge *Amphimedon queenslandica*. Differences can be also pronounced within the same lineage, for example, compared to *Nematostella*, the study of the cnidarian *Xenia* sp. reported 1100 transcripts/cell, whereas in the highly regenerative *Hydra* this number is much higher, closer to 3000 transcripts/cell. A scRNA-seq analysis of *Drosophila* aging brains reported a threefold decrease in neuronal transcripts per cell between old and young brains [51], suggesting links between cellular age/differentiation and transcriptional output. The analysis of scRNA-seq datasets with low numbers of transcripts per cell requires pooling of information across single cells, and therefore must balance the need to summarize atlases by means of relatively few cell types (that can be represented as clusters of single cell profiles) and the wish to define quantitatively complex differentiation gradients and states. In this context, highly cohesive **metacells** [89] can be derived from large groups of single cells, and these metacells can be used as building blocks to power downstream analysis and atlas interpretations [35,38,113].

Another challenge of whole-organism scRNA-seq data is the extreme heterogeneity in transcripts detected per cell: some cells are (much) bigger and/or transcriptionally active than others. Therefore, analytical strategies may take this into consideration, starting with cell/non-cell discrimination. For example, the commonly used rank-UMI inflection point (knee-plot) strategy for distinguishing cells from empty droplets can systematically discard low-transcript-content cell types. Similarly, common analytical strategies for dimensionality reduction (e.g., principal component analysis, PCA) or feature selection (e.g., finding variable genes) cannot be robustly applied to these datasets directly, requiring specifically adapted computational approaches.

Finally, the interpretation of complex whole-organism single-cell atlases, especially in species with little prior information, must rely on powerful gene functional annotations pipelines – that may be based on protein domain architectures or gene name transfer based on orthology. In addition, annotating single-cell maps must be supported by data from exploratory tools for the interrogation of different cell clusters, and should be supported by iterative validation (e.g., through broad *in situ* hybridization analysis). *In situ* follow-up screens are also important to bridge the gap between cell type catalogs and organismal tissue and body anatomy so as to further understand the functional and spatial relationships among the transcriptionally defined cell types. An illustrative example of the need for such crosstalk is the recent characterization of neurons in the ctenophore *Mnemiopsis leidyi* [90], that correspond to previously unidentified transcriptional cell clusters in a scRNA-seq atlas [35]. However, it is still unfeasible to validate all potential cell types identified in a scRNA-seq study, exactly as it is unrealistic to aim at characterizing all genes in a newly sequenced genome. Moreover, any future phylogenetic expansion in single-cell atlases will require sampling of species that cannot be maintained in the laboratory and/or for which molecular tools are limited. In such a scenario, comparison between closely related, deeply characterized species can help in the interpretation of these novel atlases. Beyond that, the comparative analysis of single-cell maps is an indispensable first step towards developing cell type evolutionary models.

Comparative Analysis of Single-Cell Atlases

As the taxon sampling of single-cell atlases increases, classical evolutionary questions can be reappraised. First, effective comparative analysis of cell type programs [91] between species is necessary to standardize and organize the new phenotypic space of cell types and gene programs. Second, based on such new comparative frameworks, models for the dynamics of cell type evolution will need to be developed.

The first step in any cross-species cell type comparison is to define genome-wide gene orthology relationships among the species involved [92]. Accurate gene orthology is essential both for supervised, gene-focused comparisons (e.g., transcription factor usage), as well as in systematic cross-species analyses (e.g., cell type clustering or tree building). Importantly, cell type comparisons necessarily involve large multigene families whose phylogenetic relationships and orthology-based classifications are difficult to resolve and that often involve linage-specific expansions (e.g., transcription factors, ion channels, or adhesion proteins). Therefore, orthology inference is necessary to explicitly account for orthologous/paralogous relationships in gene expression matrix integration, for example by restricting some analyses to one-to-one **orthologs**.

Once a set of orthologs has been defined, we can broadly identify two strategies for scRNA-seq cross-species comparison. A first set of methods directly integrate single-cell transcriptomes from different species [93,94] and develop a representation of combined atlases, merged clusters, and common reference maps. These strategies are difficult to scale to multiple species and, most importantly, to large phylogenetic distances – where it is more difficult to infer gene orthologies and to formulate explicit cell homology hypotheses. An alternative is to first resolve

the cell type map in each species and then compare the aggregated expression of cell clusters/types across orthologs [38,66,95]. This strategy allows comparisons at different levels of clustering granularity (e.g., comparing broad cell types or cell type subclusters) and makes it easier to account for non-unique or redundant gene orthology relationships. A hybrid strategy consists of single-cell integration followed by cross-species cluster overlap in the integrated space [64,96]. 'Clustering-first' methods have been applied, for example, to the systematic comparison of neuronal cell type transcriptomes, revealing widespread conservation between mouse and human, and even between mammals and reptiles. Interestingly, broad neuronal transcriptome similarity extends even beyond vertebrates to the cnidarian *Nematostella vectensis* (Figure 2) [38].

Cell type trees are often used to represent cross-species cell type comparisons. In the same way that within-species cell trees convey the underlying cell type hierarchical structure [18], cross-species trees highlight cases of interspecies cell type transcriptome similarities that are stronger than other intraspecies cell similarities. Trees are generally built using hierarchical clustering based on a distance metric derived from gene expression (e.g., correlation, Jensen–Shannon divergence, Mahalabonis distance) [38,97]. A different strategy consists of binarizing gene expression data to then apply maximum parsimony reconstruction [19] or to compute overlap-based distances (e.g., Jaccard coefficient) [65]. Often, these cell type transcriptional similarity trees are interpreted as indicative of evolutionary affinity. A first complication with deriving cell type homologies from transcriptional similarities is that a formal phylogenetic reconstruction method for cell type transcriptomes is largely missing (but *cf* [98]). That is, we lack a model that describes the evolutionary divergence rates of the **character states** involved, for example the expression of one gene within a cell type, or the rate of gain/loss of coexpression for a group of genes. These evolutionary models can eventually be derived from the systematic comparison of cell type transcriptomes in closely related species (representing a ground truth for homologous cell types), in a way analogous to how DNA or amino acid substitution models were derived from sequence alignments.

A more acute problem is the fact that, in comparing transcriptomes, we are dealing with characters that are non-independent: genes are organized into coexpression modules that are highly variable in size. Therefore, cell type transcriptome similarities will be dominated by genes in

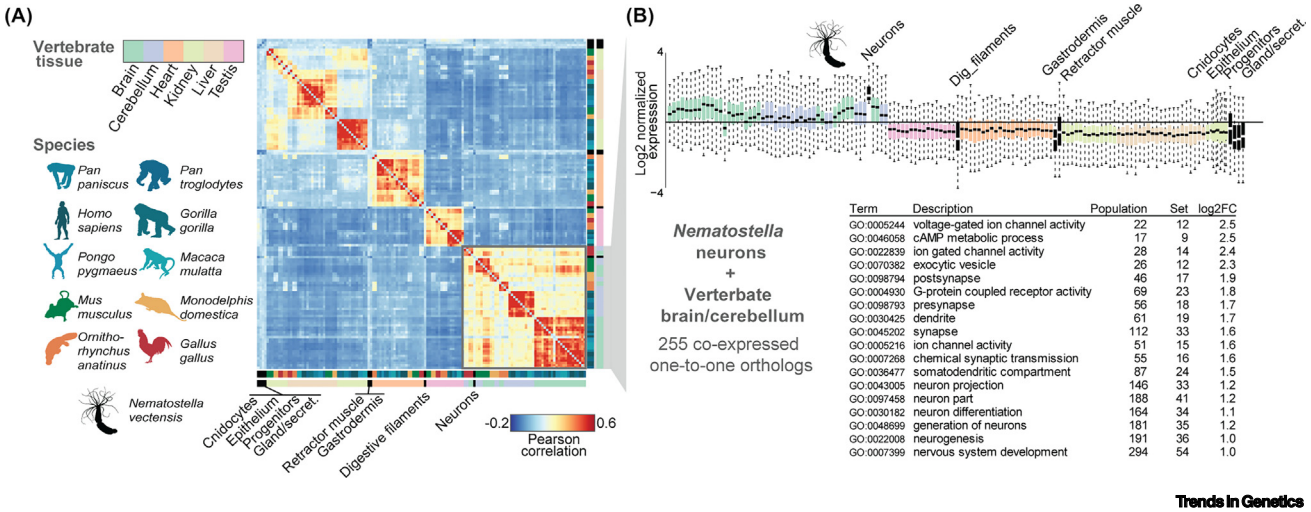

Figure 2. Comparing Cell Type Transcriptomes. (A) Correlation between organs/tissues of different vertebrate species (color-coded) and *Nematostella vectensis* (Cnidaria) broad cell types. (B) (Top) Expression 255 orthologs driving neuron and brain/cerebellum similarity across vertebrate tissues and *Nematostella* cell types. (Bottom) Enriched gene ontology terms among neuron/brain-coexpressed orthologs. Adapted, with permission, from [38].

large **gene modules**. Moreover, effector gene modules are likely to be prone to convergent recruitment by non-homologous cell types, and, reciprocally, specific cell type transcriptomes may diverge from homologous cell types by the acquisition of a single (but large) new gene module that dominates the transcriptome. The problems associated with comparative analyses of atlases are further exacerbated at large evolutionary distances [13] and when considering multiple species [99]. It is therefore important to devise taxon sampling strategies that will facilitate high-resolution modeling of cell type evolutionary dynamics, for example by focusing on a few, densely sampled clades. Beyond specific clades, it will be important to characterize recurring events of cell type innovation, loss, and merging, and to explore the dynamics of gene module remodeling, 'lateral transfer' of entire gene modules between cell type programs, and additional archetypical evolutionary scenarios that may still be uncharacterized.

Cell Type Molecular Evolution – Decoding the Evolutionary Dynamics of Cell Type Regulatory Programs

An alternative to comparing whole cell type transcriptomes is to focus on the expression of combinations of transcription factors (TFs), often called terminal selectors [100], as the key regulators of cell identity programs [2]. The implicit assumption is that TFs can represent a good proxy for the gene modules used in that particular cell type. A second underlying hypothesis when focusing on TFs is that regulatory similarities are more evolutionary constrained and therefore better approximate cell type homology (as compared to effector gene usage). However, we still do not know enough regarding the frequency by which cell-identity TFs can be replaced, especially given the intricate evolutionary history of TF gene families. Newly derived cell type atlases highlight the role of TFs from large multigene families (e.g., zf-C2H2, Ets, or Sox TFs). Such atlases also uncover multiple expressed **paralogs** that share very similar DNA-binding characteristics [101–103]. In addition, tens of different TFs are often expressed in a particular cell type (e.g., in sponge choanocytes [35]), making it difficult to determine which of them are the upstream drivers of cell type identity and what is the evolutionary significance of the conservation of a few of those TFs. Overall, we lack a systematic understanding of the extent to which TF expression is conserved between homologous cell types, and therefore of whether TF usage can be widely used for cell type evolutionary comparisons.

TFs control gene expression by recognizing and binding to short sequence motifs (6–12 bps) located in *cis*-regulatory regions (promoters and enhancers) of downstream genes [104]. Cell type transcriptional identity is strongly recapitulated by sequence motif enrichment [35,38] (Figure 3), representing the *cis*-regulatory embedding of the cell type program [100]. In addition, with a few exceptions (such as zf-C2H2 TFs), the binding sequences of TFs are very often conserved across large phylogenetic distances [102,103,105]. This opens the possibility of comparing cell types not through their gene expression profiles but instead through the set of regulatory sequences that define the cell type program. A recent study pioneered the idea of cross-species *cis*-regulatory sequence comparison [106]. Working on melanoma cell lines in different vertebrate species, Minnoye *et al.* uncovered a highly conserved *cis*-regulatory program that involves combinations of four TF-binding (SOX10, TFAP2A, MITF, and ETS) motif enhancer regions that most often showed little or no global sequence conservation. Building on this dissection of melanoma enhancer motif syntax, the authors further modeled the effect of evolutionary mutations in enhancer function, as defined by accessibility.

The different elements that constitute a cell type gene expression program (TFs, effector genes, TF binding sites, regulatory connections, etc.) do not necessarily have congruent evolutionary histories [107,108], in the same way as gene trees are not always in agreement with species

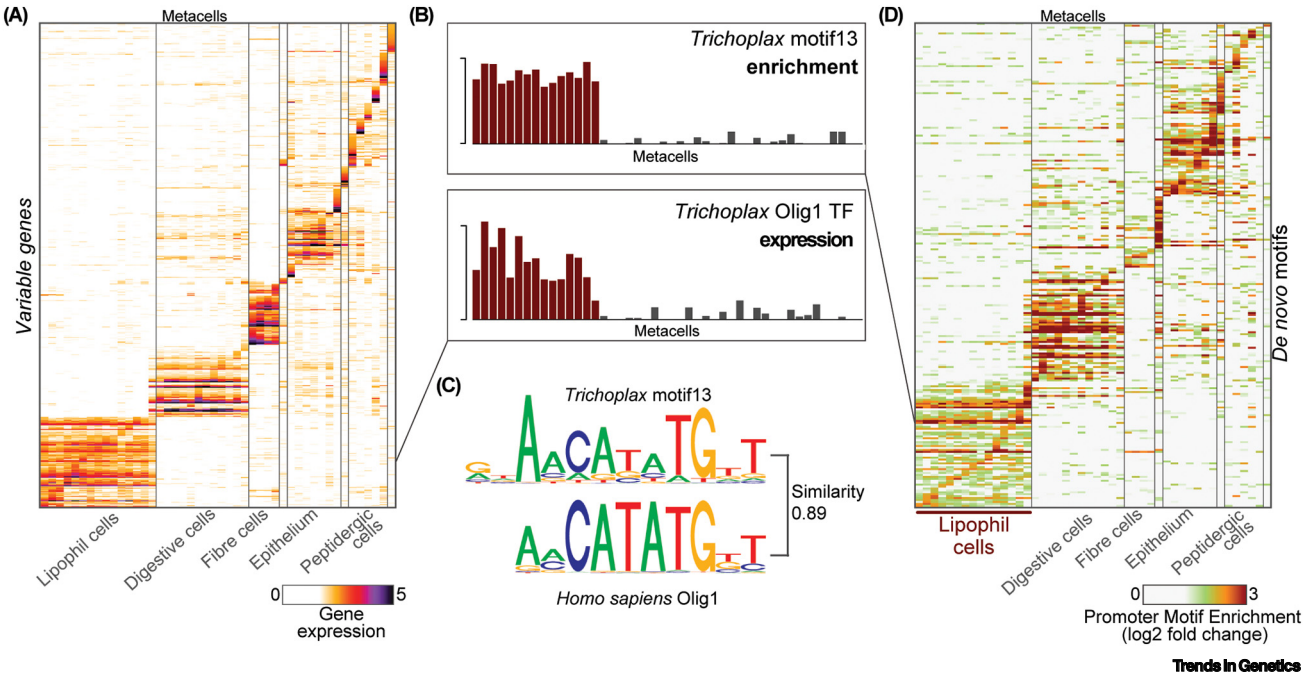

Figure 3. Genomic Embedding of Cell Type Transcriptional Programs. (A) *Trichoplax adhaerens* (Placozoa) cell type gene expression map. (B) Comparison of over-representation of a specific sequence motif in the promoter of metacell-expressed genes (motif enrichment) and the expression across metacells of the Olig1 transcription factor, the potential binder of motif13. (C) Comparison of *Trichoplax* motif13 and the highly similar binding site for *Homo sapiens* Olig1, suggesting that motif13 may also represent the binding site for *Trichoplax* transcription factor (TF) Olig1. (D) Motif enrichment heatmap in *Trichoplax* metacell-specific promoter sequences. Based on data from [35].

trees (as a result of horizontal gene transfers, incomplete lineage sorting, etc.). By combining sequence motif analysis with high-resolution chromatin accessibility data [109] and even single-cell accessibility data [110], we should be able to systematically reconstruct cell type gene regulatory networks in non-model species. Disentangling and comparing regulatory programs in multiple closely related species will enable the development of quantitative models of cell type evolution, including **evolutionary rates** of distinct regulatory characters: TF usage/replacement, sequence motifs, regulatory interactions, gene module composition [88], and more. These models should constitute the basis of a future cell type phylogenetics and will help to address important questions in cell type evolution: are these evolutionary rates universally conserved [111], or are there particularly 'fast-evolving' cell type programs? How robust are cell type genetic networks, and which components are particularly evolvable? Finally, identifying slowly evolving regulatory characters (e.g., sequence motifs) could help to formulate better cell type homology hypotheses.

From a broad perspective, the feasibility of inferring a comprehensive cell type phylogeny across species and phyla remains largely unclear. Such feasibility depends on the timescales and degree of constraint for traits that can be inferred robustly; in that respect, it is crucial to consider systematically all of the multiple molecular scales underlying gene regulation and cell type specification (TFs, regulatory elements, gene modules). In any case, comparison of cell type programs within densely sampled clades will be an essential first step to help us to define which characters can be used to infer cell type phylogenies and at what resolution. At the very least, such comparative studies will reconstruct specific histories and identify general trends in the evolution of cell type programs.

Concluding Remarks: Towards a Cell Type Tree of Life

Whole-organism scRNA-seq analysis holds the promise to develop comprehensive catalogs of cell types in phylogenetically diverse systems. Reference cell type molecular atlases will crucially advance our understanding of the biology of understudied animal groups [112]. This is in a way similar to how the sequencing and annotation of genomes of unsampled animal lineages uncovers novel biology and enables comparative genomic studies [67]. The most immediate challenge is to develop methodological standards to build cell type maps in the most unbiased and consistent manner (see Outstanding Questions). Only with dense and technically compatible phylogenetic sampling we will be able to start a systematic comparative study of cell type programs. Based on this sampling, cell type comparative biology will enable the development of cell type phylogenetic models and can promote our understanding of genetic changes associated with cellular novelty. Overall, this can offer transformative insights linking classical models of (genomic) molecular evolution with an intermediate molecular phenotype: cell types and their associated gene regulatory networks.

Acknowledgments

A.T. was supported by the European Research Council (ERC; grant 724824). A.S-P. was supported by the ERC under the EU Horizon 2020 Research and Innovation Programme (grant agreement 851647), the Ministerio de Ciencia e Innovación, Centro de Excelencia Severo Ochoa (ref. SEV-2016-0571), and the Agencia Estatal de Investigación. We thank Xavier Grau-Bové, Anamaria Elek, and Sebastián R. Najle for critical reading of the manuscript, as well as two anonymous reviewers for their valuable comments.

Declaration of Interests

The authors declare no conflicts of interest.

References

1. Arendt, D. (2008) The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* 9, 868–882
2. Arendt, D. *et al.* (2016) The origin and evolution of cell types. *Nat. Rev. Genet.* 17, 744–757
3. Willmer, E.N. (1970) *Cytology and Evolution*, Academic Press
4. Xia, B. and Yanai, I. (2019) A periodic table of cell types. *Development* 146, dev169854
5. Schwartz, G.W. *et al.* (2020) TooManyCells identifies and visualizes relationships of single-cell clades. *Nat. Methods* 17, 405–413
6. Baguñá, J. and Romero, R. (1981) Quantitative analysis of cell types during growth, degrowth and regeneration in the planarians *Dugesia mediterranea* and *Dugesia tigrina*. *Hydrobiologia* 84, 181–194
7. Bode, H. *et al.* (1973) Quantitative analysis of cell types during growth and morphogenesis in *Hydra*. *Wilhelm Roux. Arch. Entwickl. Mech. Org.* 171, 269–285
8. Simpson, T.L. (1984) *The Cell Biology of Sponges*, Springer New York
9. Valentine, J.W. *et al.* (1994) Morphological complexity increase in metazoans. *Paleobiology* 20, 131–142
10. Novershtern, N. *et al.* (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* 144, 296–309
11. Hirano, M. *et al.* (2013) Evolutionary implications of a third lymphocyte lineage in lampreys. *Nature* 501, 435–438
12. Tessmar-Raible, K. *et al.* (2007) Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. *Cell* 129, 1389–1400
13. Steinmetz, P.R.H. *et al.* (2012) Independent evolution of striated muscles in cnidarians and bilaterians. *Nature* 487, 231–234
14. Ogino, K. *et al.* (2011) Distinction of cell types in *Dicyema japonicum* (phylum Dicyemida) by expression patterns of 16 genes. *J. Parasitol.* 97, 596–601
15. Martín-Durán, J.M. *et al.* (2018) Convergent evolution of bilaterian nerve cords. *Nature* 553, 45–50
16. Sacerdot, C. *et al.* (2018) Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* 19, 166
17. Breschi, A. *et al.* (2020) A limited set of transcriptional programs define major cell types. *Genome Res.* 30, 1047–1059
18. Liang, C. *et al.* (2015) The statistical geometry of transcriptome divergence in cell-type evolution and cancer. *Nat. Commun.* 6, 6066
19. Kin, K. *et al.* (2015) Cell-type phylogenetics and the origin of endometrial stromal cells. *Cell Rep.* 10, 1398–1409
20. Cherbas, L. *et al.* (2011) The transcriptional diversity of 25 *Drosophila* cell lines. *Genome Res.* 21, 301–314
21. Brown, J.B. *et al.* (2014) Diversity and dynamics of the *Drosophila* transcriptome. *Nature* 512, 393–399
22. Forrest, A.R.R. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature* 507, 462–470
23. Alié, A. *et al.* (2015) The ancestral gene repertoire of animal stem cells. *Proc. Natl. Acad. Sci. U. S. A.* 112, E7093–E7100
24. Sogabe, S. *et al.* (2019) Pluripotency and the origin of animal multicellularity. *Nature* 570, 519–522
25. Tang, F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382
26. Islam, S. *et al.* (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21, 1160–1167
27. Jaitin, D.A. *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, 776–779
28. Svensson, V. *et al.* (2018) Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13, 599–604
29. Marx, V. (2021) Method of the Year 2020: spatially resolved transcriptomics. *Nat. Methods* 18, 9–14
30. Regev, A. *et al.* (2017) The human cell atlas. *eLife* 6, e27041
31. Cao, J. *et al.* (2020) A human cell atlas of fetal gene expression. *Science* 370, eaba7721
32. Han, X. *et al.* (2018) Mapping the mouse cell atlas by microwell-seq. *Cell* 172, 1091–1107
33. Schaum, N. *et al.* (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562, 367–372

Outstanding Questions

What are the optimal single-cell sampling strategies for generating unbiased and highly compatible cell atlases in different species? Is it possible to develop universal sampling methods and standards?

What is the ideal granularity by which one should define a cell type? How should we handle a differentiation continuum, or cells that respond to multiple signals?

What characters can we use to approximate the evolutionary relationships between cell types (e.g., global gene expression versus TF expression)? What is the best way to codify the states for these characters (e.g., binarizing gene expression)?

More generally, what are the evolutionary rates of the different elements of a cell type identity program (TF usage, gene modules, *cis*-regulatory regions, TF binding motifs)? Are these rates the same in all cell types and in all organisms?

How does the underlying genome evolutionary dynamics (gene duplication/loss, mutation in regulatory sequences, etc.) impact on the emergence of novel cell types? Reciprocally, how robust are cell type identity programs to these genomic changes? What types of genomic changes most frequently underlie the cooption of gene models between cell types?

What is the role of regulatory innovation (e.g., combinatorial usage of TFs, use of distal regulatory elements, splicing, etc.) in the diversification of cell types? Do organisms with low cell type diversity use a different regulatory logic (e.g., promoter versus enhancer regulation) from those with a large diversity of cell types? Similarly, do particularly diverse cell types (e.g., neurons or gland/secretory cells) use distinct encoding strategies?

34. Marioni, J.C. and Arendt, D. (2017) How single-cell genomics is changing evolutionary and developmental biology. *Annu. Rev. Cell Dev. Biol.* 33, 537–553
35. Sebé-Pedrós, A. *et al.* (2018) Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat. Ecol. Evol.* 2, 1176–1188
36. Musser, J.M. *et al.* (2019) Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *BioRxiv* Published online September 5, 2019. https://doi.org/10.1101/758276
37. Chari, T. *et al.* (2021) Whole animal multiplexed single-cell RNA-seq reveals plasticity of *Clytia medusa* cell types. *BioRxiv* Published online February 7, 2021. https://doi.org/10.1101/2021.01.22.427844
38. Sebé-Pedrós, A. *et al.* (2018) Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-seq. *Cell* 173, 1520–1534
39. Hu, M. *et al.* (2020) Lineage dynamics of the endosymbiotic cell type in the soft coral Xenia. *Nature* 582, 534–538
40. Siebert, S. *et al.* (2019) Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science* 365, eaav9314
41. Duruz, J. *et al.* (2021) Acoel single-cell transcriptomics: cell type analysis of a deep branching bilaterian. *Mol. Biol. Evol.* 38, 1888–1904
42. Cao, J. *et al.* (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502
43. Li, P. *et al.* (2021) Single-cell analysis of *Schistosoma mansoni* identifies a conserved genetic program controlling germline stem cell fate. *Nat. Commun.* 12, 485
44. Fincher, C.T. *et al.* (2018) Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* 360, eaaq1736
45. Plass, M. *et al.* (2018) Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 1723, eaaq1723
46. Hung, R.-J. *et al.* (2020) A cell atlas of the adult *Drosophila* midgut. *Proc. Natl. Acad. Sci. U. S. A.* 117, 1514–1523
47. Rust, K. *et al.* (2020) A single-cell atlas and lineage analysis of the adult *Drosophila* ovary. *Nat. Commun.* 11, 5628
48. Slaidina, M. *et al.* (2020) A single-cell atlas of the developing *Drosophila* ovary identifies follicle stem cell progenitors. *Genes Dev.* 34, 239–249
49. Allen, A.M. *et al.* (2020) A single-cell transcriptomic atlas of the adult *Drosophila* ventral nerve cord. *eLife* 9, e54074
50. Croset, V. *et al.* (2018) Cellular diversity in the *Drosophila* midbrain revealed by single-cell transcriptomics. *eLife* 7, e34550
51. Davie, K. *et al.* (2018) A single-cell transcriptome atlas of the aging *Drosophila* brain. *Cell* 174, 982–998
52. Massri, A.J. *et al.* (2020) Developmental single-cell transcriptomics in the *Lytechinus variegatus* sea urchin embryo. *BioRxiv* Published online November 13, 2020. https://doi.org/10.1101/2020.11.12.380675
53. Foster, S. *et al.* (2020) A single cell RNA sequencing resource for early sea urchin development. *Development* 147, dev191528
54. Paganos, P. *et al.* (2021) Single cell RNA sequencing of the *Strongylocentrotus purpuratus* larva reveals the blueprint of major cell types and nervous system of a non-chordate deuterostome. *BioRxiv* Published online March 16, 2021. https://doi.org/10.1101/2021.03.16.435574
55. Achim, K. *et al.* (2018) Whole-body single-cell sequencing reveals transcriptional domains in the annelid larval body. *Mol. Biol. Evol.* 35, 1047–1062
56. Horie, R. *et al.* (2018) Shared evolutionary origin of vertebrate neural crest and cranial placodes. *Nature* 560, 228–232
57. Sladitschek, H.L. *et al.* (2020) MorphoSeq: full single-cell transcriptome dynamics up to gastrulation in a chordate. *Cell* 181, 922–935
58. Packer, J.S. *et al.* (2019) A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* 365, eaax1971
59. Karaiskos, N. *et al.* (2017) The *Drosophila* embryo at single-cell transcriptome resolution. *Science* 358, 194–199
60. Pijuan-Sala, B. *et al.* (2019) A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 490–495
61. Wagner, D.E. *et al.* (2018) Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 4362, eaar4362
62. Briggs, J.A. *et al.* (2018) The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 5780, eaar5780
63. Farrell, J.A. *et al.* (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* 3131, eaar3131
64. Hodge, R.D. *et al.* (2019) Conserved cell types with divergent features in human versus mouse cortex. *Nature* 573, 61–68
65. Shafer, M.E.R. *et al.* (2021) Gene family evolution underlies cell type diversification in the hypothalamus of teleosts. *BioRxiv* Published online January 11, 2021. https://doi.org/10.1101/2020.12.13.414557
66. Tosches, M.A. *et al.* (2018) Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* 360, 881–888
67. Dunn, C.W. and Ryan, J.F. (2015) The evolution of animal genomes. *Curr. Opin. Genet. Dev.* 35, 25–32
68. Koepfli, K.-P. *et al.* (2015) The Genome 10K project: a way forward. *Annu. Rev. Anim. Biosci.* 3, 57–111
69. Pollen, A.A. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058
70. Klein, A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201
71. Zeisel, A. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142
72. Denisenko, E. *et al.* (2020) Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* 21, 130
73. Lun, A.T.L. *et al.* (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* 20, 63
74. Heaton, H. *et al.* (2020) Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* 17, 615–620
75. Alles, J. *et al.* (2017) Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* 15, 44
76. Rosenberg, A.B. *et al.* (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182
77. García-Castro, H. *et al.* (2021) ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. *Genome Biol.* 22, 89
78. Habib, N. *et al.* (2017) Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* 14, 955–958
79. Lake, B.B. *et al.* (2016) Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* 352, 1586–1590
80. Petrany, M.J. *et al.* (2020) Single-nucleus RNA-seq identifies transcriptional heterogeneity in multinucleated skeletal myofibers. *Nat. Commun.* 11, 6374
81. Ding, J. *et al.* (2020) Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* 38, 737–746
82. Drokhlyansky, E. *et al.* (2020) The human and mouse enteric nervous system at single-cell resolution. *Cell* 182, 1606–1622
83. Slyper, M. *et al.* (2020) A single-cell and single-nucleus RNA-seq toolbox for fresh and frozen human tumors. *Nat. Med.* 26, 792–802
84. Mudge, J.M. and Harrow, J. (2016) The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.* 17, 758–772
85. Srivastava, M. *et al.* (2008) The *Trichoplax* genome and the nature of placozoans. *Nature* 454, 955–960
86. Nikitin, M. (2015) Bioinformatic prediction of *Trichoplax adhaerens* regulatory peptides. *Gen. Comp. Endocrinol.* 212, 145–155

87. Varoqueaux, F. *et al.* (2018) High cell diversity and complex peptidergic signaling underlie placozoan behavior. *Curr. Biol.* 28, 3495–3501
88. Feregrino, C. and Tschopp, P. (2021) Assessing evolutionary and developmental transcriptome dynamics in homologous cell types. *BioRxiv* Published online Febuary 10, 2021. https://doi.org/10.1101/2021.02.09.430383
89. Baran, Y. *et al.* (2019) MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* 20, 206
90. Sachkova, M.Y. *et al.* (2021) The unique neuronal structure and neuropeptide repertoire in the ctenophore *Mnemiopsis leidyi* shed light on the evolution of animal nervous systems. *BioRxiv* Published online March 31, 2021. https://doi.org/10.1101/2021.03.31.437758
91. Shafer, M.E.R. (2019) Cross-species analysis of single-cell transcriptomic data. *Front. Cell Dev. Biol.* 7, 175
92. Altenhoff, A.M. *et al.* (2016) Standardized benchmarking in the quest for orthologs. *Nat. Methods* 13, 425–430
93. Butler, A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420
94. Liu, J. *et al.* (2020) Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat. Protoc.* 15, 3632–3662
95. La Manno, G. *et al.* (2016) Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* 167, 566–580
96. Bakken, T.E. *et al.* (2020) Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. *BioRxiv* Published online April 4, 2020. https://doi.org/10.1101/2020.03.31.016972
97. Merkin, J. *et al.* (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* 338, 1593–1599
98. Musser, J.M. and Wagner, G.P. (2015) Character trees from transcriptome data: Origin and individuation of morphological characters and the so-called 'species signal'. *J. Exp. Zool. B Mol. Dev. Evol.* 324, 588–604
99. Dunn, C.W. *et al.* (2018) Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl. Acad. Sci. U. S. A.* 115, E409–E417
100. Hobert, O. (2008) Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. *Proc. Natl. Acad. Sci. U. S. A.* 105, 20067–20071
101. Nitta, K.R. *et al.* (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* 4, e04837
102. Weirauch, M.T. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443
103. Lambert, S.A. *et al.* (2019) Similarity regression predicts evolution of transcription factor sequence specificity. *Nat. Genet.* 51, 981–989
104. Spitz, F. and Furlong, E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626
105. Sebé-Pedrós, A. *et al.* (2013) Early evolution of the T-box transcription factor family. *Proc. Natl. Acad. Sci. U. S. A.* 110, 16050–16055
106. Minnoye, L. *et al.* (2020) Cross-species analysis of enhancer logic using deep learning. *Genome Res.* 30, 1815–1834
107. Tschopp, P. and Tabin, C.J. (2017) Deep homology in the age of next-generation sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* 372, 20150475
108. Shubin, N. *et al.* (2009) Deep homology and the origins of evolutionary novelty. *Nature* 457, 818–823
109. Vierstra, J. *et al.* (2020) Global reference mapping of human transcription factor footprints. *Nature* 583, 729–736
110. Cusanovich, D.A. *et al.* (2018) A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* 174, 1309–1324
111. Carvunis, A.-R. *et al.* (2015) Evidence for a common evolutionary rate in metazoan transcriptional networks. *eLife* 4, e11615
112. Dunn, C.W. *et al.* (2015) The hidden biology of sponges and ctenophores. *Trends Ecol. Evol.* 30, 282–291
113. Levy, S. *et al.* (2021) A stony coral cell atlas illuminates the molecular and cellular basis of coral symbiosis, calcification, and immunity. *Cell* Published online April 28, 2021. https://doi.org/10.1016/j.cell.2021.04.005