Cell
Leading Edge

CellPress

Perspective

A reference cell tree will serve science better than a reference cell atlas

Silvia Domcke[1,*] and Jay Shendure[1,2,3,4,*]
[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
[2]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA
[3]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA
[4]Howard Hughes Medical Institute, Seattle, WA, USA
*Correspondence: domcke@uw.edu (S.D.), shendure@uw.edu (J.S.)
https://doi.org/10.1016/j.cell.2023.02.016

SUMMARY

Single-cell biology is facing a crisis of sorts. Vast numbers of single-cell molecular profiles are being generated, clustered and annotated. However, this is overwhelmingly *ad hoc*, and we continue to lack a principled, unified, and well-moored system for defining, naming, and organizing cell types. In this perspective, we argue against an atlas or periodic table-like discretization as the right metaphor for a reference taxonomy of cell types. In its place, we advocate for a data-driven, tree-based nomenclature that is rooted in a "consensus ontogeny" spanning the life cycle of a given species. We explore how such a reference cell tree, inclusive of both lineage histories and molecular states, could be constructed, represented, and segmented in practice. Analogous to the taxonomic classification of species, a consensus ontogeny would provide a universal, stable, and extendable framework for precise scientific communication, both contemporaneously and across the ages.

CLASSIFICATION IN BIOLOGY

In the preface to *The Order of Things*,[1] the French philosopher Michel Foucault illustrates the limitations of taxonomic classifications by quoting a fictional "certain Chinese encyclopedia" that states "animals are divided into: (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camel hair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies." Foucault's point in citing this absurd taxonomy is that we often assume that an accepted classification scheme represents an objective reality, but there are infinite alternatives. This matters, he argues, because our systems of classification both reflect *and* direct our thinking: how we apply labels, and how we relate those labels to one another, shapes our conception of the underlying phenomena.

To the extent that biology engages in "summarizing" the natural world, we should give careful consideration to how we frame the task, out of the infinite ways it might be done. Species taxonomy provides an excellent example. The Aristotelian system, which persisted for two millenia, classified animals into a dozen groups spanning a scale of relative "perfection". By the early 18th century, this was complemented by a polynomial, largely descriptive, and highly fluid nomenclature.[2] For example, the European honey bee might be known as "*Apis pubescens, thorace subgriseo, abdomine fusco, pedibus posticis glabris utrinque margine ciliatis,*" which translates as "furry bee, grayish thorax, brownish abdomen, black legs smooth with hair on both sides".[3] These unwieldy descriptors, lacking a shared organizational framework or information on how classes relate to one another, bring to mind the contemporary practice of naming cell types by combinations of differentially expressed marker genes.

In the late 18th century, species classification was revolutionized by Linnaeus, who began organizing species based on a binomial nomenclature ("genus" + "species", e.g., *Homo sapiens*). One of the first scientists to use paper-based index cards, Linnaeus was able to efficiently incorporate new species, as well as new information about already-named species, into a stable framework.[4] Although there are recent advocates for starting anew with a nomenclature grounded solely in molecular phylogenetics, Linnaeus' system is proving stubborn to displace. This is in part because it already meets key criteria, such as enabling precise communication (contemporaneously, as well as with past and future researchers) and the integration of new information into a stable, widely accepted backbone.

The invention of the microscope in the 17th century opened up an entirely new dimension in biology through the recognition that cells are the basic structural and functional unit of all living things, and moreover that all cells arise from other cells.[5] Cells were initially classified by physical appearance.[6] As microscopy improved and new kinds of measurement (immunohistochemical, electrophysiological) and understanding (functional, developmental, evolutionary) emerged, cell type nomenclature became increasingly muddled, and ultimately subfield-specific.

Within the past decade, new technologies have enabled the routine profiling of the mRNA contents of single cells. As the

Box 1. Terminology

Cell type: A recurring pattern of developmental origin and potential within and across cell lineage trees of individuals of a given species, generally reflected in shared molecular properties.

Cell state: Variations in molecular phenotypes within a cell type that do not impact its developmental potential (e.g., cell cycle, stochastic fluctuations).

Cell identity: An individual cell as characterized solely by its molecular phenotypes at a given moment in time.

Cell lineage: The relationships among cells of an individual organism as defined solely by the series of cell divisions that begins with a single zygote.

Cell trajectory: Ordering of cells' developmental relationships inferred solely from similarity in molecular phenotypes, which might or might not recapitulate developmental cell lineage relationships.

practitioners of such methods have expanded from a handful to thousands of labs, the repertoire of associated computational tools has also grown. However, although there are excellent algorithms for organizing single-cell profiles into manageable numbers of "clusters", researchers are largely left to their own devices with respect to naming clusters and relating them to other datasets.

This is unfortunate for several reasons. First, it leads to considerable repetition of effort, often in the form of time-intensive literature and web searches (e.g., the unsystematic practice of "googling" differentially expressed genes). Second, it is the wild west out there, with no widely accepted standards around annotation quality or nomenclature. Although we are increasingly adept at integrating datasets and transferring labels computationally, this risks simply propagating potentially suboptimal descriptors. Third, the resulting corpus is heavily biased toward the systems in which the data is being generated (a complex function of scientific interest, resource allocation, and technical factors), rather than being anchored in a natural distribution. Fourth, it represents a missed opportunity, as it doesn't feel like we are moving toward any consensus or cohesion that mirrors Linnaeus's index cards, where new information can simply be added to a stable backbone.

As more data is generated, the situation is becoming progressively worse. Not only do we lack a unified system of cell type classification,[7] we also lack consensus on which is the most useful unit for classification or what the terms that we routinely use actually mean. For example, cell "type", "state", and "identity" are often used interchangeably, as are cell "lineage" and "trajectory".[8] How we define these terms in the context of this perspective is summarized in Box 1.

In our opinion, we should be pushing for a cell type nomenclature that meets some of the same key criteria as Linnaean taxonomy, as well as additional ones, including: (1) accommodating all cells arising during the life cycle of a given organism; (2) accommodating inter-individual variation, both normal and disease-related; (3) relating cell types to one another in a biologically meaningful way; (4) being stable to the incorporation of new data or new data types; and (5) being constructed in a largely, if not entirely, data-driven manner. As we will be studying cells for a long time, a nomenclature that meets these criteria is necessary for precise scientific communication in the present and across the ages. How can we ensure that the cell type labels we use in papers today can be accurately interpreted by researchers working 20 or 200 years from now?

What is the right organizing principle?

What are the options for an "organizing principle" around which we might base a universal, stable, and useful nomenclature? To date, cell type names have primarily derived from historical, morphological, functional, molecular, evolutionary, and/or developmental distinctions. However, they are not consistently based on any one of these. Rather, each name seems to be a result of historical contingency. For example, the retina includes cell types named by shape/position (horizontal cells), function (photoreceptors) and scientist (Müller glia).

This mixing of organizing principles is also an explicit choice of the most serious effort to date to create a universal nomenclature. Cell Ontology is a curated resource, analogous to Gene Ontology, with over 2,000 cell type classes. No single organizing principle is prioritized, but rather the classes correspond to different cellular properties (e.g., functional, histological) or broad "lineage" classes (e.g., "blood cells").[9] In our view, we should be more consistent, choosing a "primary" principle, even if others are given secondary consideration. Given the known discrepancies between different organizing principles (e.g., two cells may have similar molecular profiles or functions but be developmentally unrelated), this is necessary to end up at an unambiguous, universal framework.

What is the "right" choice for this principle? Let us consider the more obvious options in turn: historical, morphological, functional, evolutionary, molecular, and developmental (Figure 1). We immediately reject historical footnotes, as there is no practical utility to naming cell types after specific researchers. We also reject morphology—although all cells could be systematically grouped by morphology, the resulting classification would be decidedly less meaningful than alternatives.

A system in which shared physiological function is the primary characteristic merits consideration,[10] and is arguably the dominant principle in the nomenclature we have inherited. Even cells with non-functional names are often still primarily defined by their function (e.g., B cells). Such functional definitions boil down all aspects of cellular biology into a brief *raison d'être*. However, our knowledge and assignment of cells' primary function is inherently subjective and decidedly incomplete. Although some gaps could be filled through systematic approaches (e.g., inferences based on gene expression modules), this risks simply propagating biases of current knowledge.

One can also take an evolutionary perspective, defining cell types as "a set of cells in an organism that change in evolution together, partially independent of other cells, and are evolutionarily more closely related to each other than to other cells".[11] This definition relies on the concept of an evolutionarily stable "core regulatory complex" (CoRC), composed of key transcription factors (TFs), miRNAs, RNA binding proteins, etc. Although undoubtedly useful for certain goals, we argue against an evolutionary perspective for our purposes. First, identifying

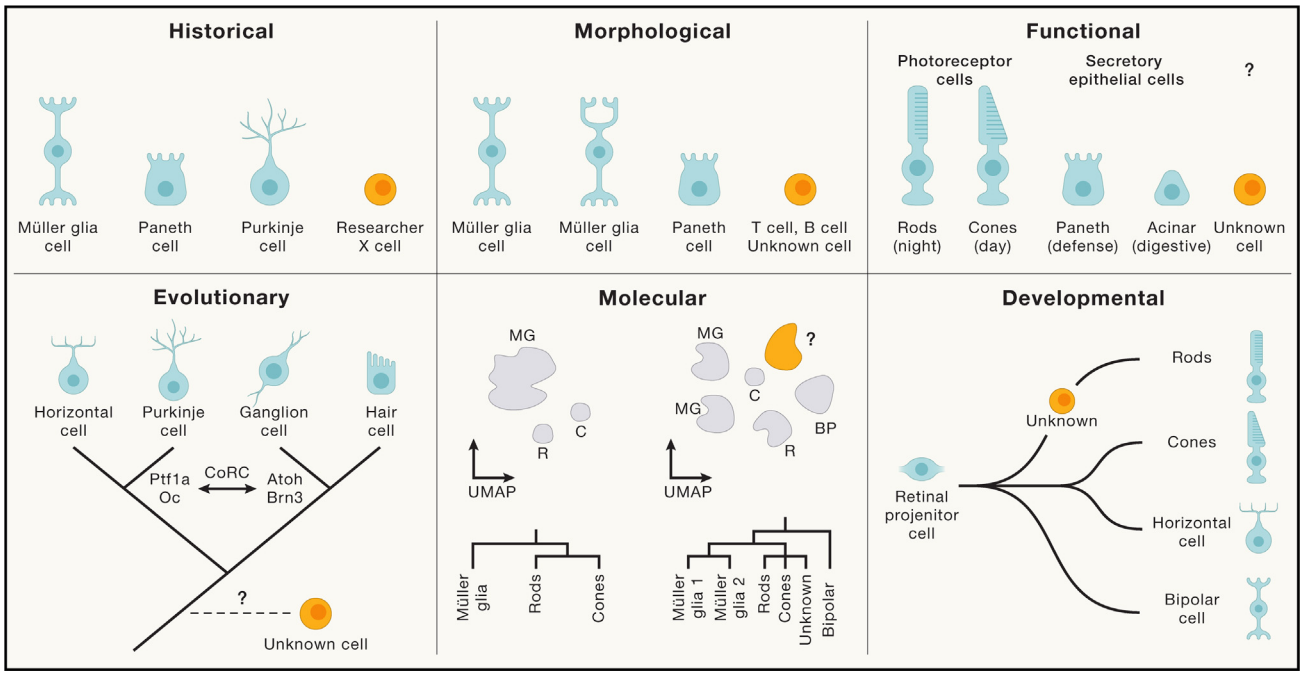

Figure 1. Potential organizing principles for cell type taxonomy
An ideal principle would: arrange cells in a biologically meaningful way; be stable to the incorporation of new data; accommodate all cells arising during the organism's natural life cycle; and be fully data-driven.

orthologous cell types across vast evolutionary distances is considerably more challenging for cell types than for gene sequences. Second, the CoRC definition focuses on terminally differentiated cell types, deemphasizing the way(s) that each cell type came to be within a developing organism. Third, the use of CoRCs as the distinguishing feature of each cell type is diffuse, as it is subjective which factors would be included or what level of difference would delineate CoRCs from one another. Fourth, the overwhelming majority of biomedical research is focused on humans and a few model organisms. Giving primacy to evolution comes at the expense of precision in defining relationships among cell types within this handful of intensely studied species.

How about molecular profiling? The paradigmatic example is the "cluster of differentiation" (CD) system for classifying immune cells based on cell-surface markers. Single-cell transcriptomics (scRNA-seq) has recently emerged as means of scalably sampling *all* mRNAs from single cells or nuclei. Although the resulting profiles are sparse, they can nonetheless be clustered into groups corresponding to the expected "cell types" of the tissue-of-origin. As scRNA-seq has grown in use, the field increasingly relies on this approach to organize cells into "periodic tables" or hierarchically clustered dendrograms.[12,13] Although laborious googling of marker genes is gradually being displaced by automated annotation via label transfer,[14] these methods tend to rely on predefined marker genes or labeled reference datasets as input, which themselves do not follow any systematic nomenclature.

On one hand, molecular profiling meets some of our key criteria. It is systematic, comprehensive, and data-driven. The very fact that cells often cluster into discrete groups rather than a messy continuum supports the concept of discretized cell types. These cell types can be conceptualized as basins of attraction in a Waddington landscape, with the measured transcriptome reflecting but one aspect of cell identity underlying a cell's stability within a given basin.[15]

On the other hand, there are limitations to this framing. A first concern is it is not necessarily stable to the incorporation of new data, be it from cancerous tissue, or a different developmental stage, or even from the *same* sample, as clustering is sensitive to the technology used, batch effects, depth of profiling, etc. Different groupings may result based on what is being measured (e.g., chromatin vs. RNA vs. protein). It also forces us to choose a resolution at which to define cell types, as one can cluster and sub-cluster *ad infinitum.* Although the decision of when to stop can be reached in a principled way, even this seems to risk drawing less on any underlying biological reality than on the human impetus to organize any observed heterogeneity into a set of discrete, namable things. Of note, these challenges have been addressed in part by new methods to build harmonized cell type hierarchies out of different atlases and update existing references as cell type resolution increases due to increased sampling.[16]

A second concern is that the discretization of cell types fails to incorporate continuous forms of molecular heterogeneity, e.g., spatial or temporal gradients in gene expression. This issue—of being unable to see a clear boundary between cell types and needing to rely on experts who might disagree on where to put the threshold—is an acknowledged challenge for automated label transfer.[16]

A third concern has less to do with how cell types are defined and named than how they are related to one another. Although a periodic table does relate cell types to one another through rows and columns, and dendrograms capture cell type similarities along one biological dimension, this is reminiscent of the pre-Linnaean species nomenclature based on arbitrary external similarities. Though the descriptive properties that form the basis for the classification might be technically accurate, they are not inherently rooted in any specific axis of biology that we care about, whether functional, evolutionary, or developmental.

Finally, we come to developmental relationships. Lineage represents the ground truth of a cell's history and future—where it comes from and what it will give rise to: clear, quantifiable relationships between cells as they arose within the context of one individual. The paradigmatic example is *Caenorhabditis elegans*, where an invariant lineage, painstakingly documented by visual observation, allows for each and every cell in this organism's life history to be precisely named.[17] For example, "MS paapaaa", also known as M1, is a pharyngeal motor neuron, whose descent from the MS founder cell is reflected by the letters corresponding to the (a)nterior or (p)osterior daughter of successive cell divisions. Such centering on development reminds us that the discrete clustering of cell types is an illusion, borne of the fact that scRNA-seq provides a snapshot rather than a movie, i.e., the data samples a specific moment in an organism's life cycle and is blind to time.

However, *C. elegans* has a number of features that make it uniquely well-suited to a lineage-based nomenclature, including: (1) the invariant nature of its wild-type development; (2) adult *C. elegans* consists of only $\sim 10^3$ cells; (3) the lineage is known due to the confluence of its invariance, the organism's transparency, and the sheer persistence of Sulston. In a few instances, *C. elegans* also illustrates how purely lineage-based descriptors might obscure the functional homologies among cells with disparate lineage histories. For example, IL1 and IL2 neurons derive from different founder cells, yet are practically indistinguishable transcriptionally; the same is true of subsets of muscle cells.[17,18]

For more complex, non-transparent organisms, it is implausible that visual observation will yield a complete cell lineage tree. As an alternative, we and others have developed methods that explicitly record cell lineage via clonal tagging[19] or evolving barcodes.[20] However, even if these methods worked as well as we can possibly imagine, the best outcome would be a complete cell lineage tree of an individual in which only the terminal nodes are annotated by cell identity. As we would remain blind to the identities of inferred ancestors (not to mention "extinct" lineages, e.g., due to apoptosis), it is unclear how such a tree alone would yield a satisfactory cell type nomenclature. Finally, naming cells by each cell division is not necessarily useful for organisms consisting of trillions of cells, especially if those divisions do not even play out in an invariant manner.

Toward a molecularly annotated consensus ontogeny

In summary, none of these potential organizing principles are satisfactory, at least not in isolation. However, the last two (molecular, developmental) have appealing features. Might they be combined?

Two maturing technical paradigms are relevant here. The first involves "whole organism" scRNA-seq profiling of multiple developmental stages of model organisms, coupled to computational inference of transcriptional trajectories.[21] Such studies can yield "pseudo-trees" encompassing the development of large, opaque model organisms, including fly, frog, zebrafish, and mouse.[22,23,24,25] Although these pseudo-trees will undoubtedly continue to improve in resolution and completeness, they remain inherently inferential in terms of the cell identities linked between successive time points, which has clear limitations. For example, asymmetric cell divisions are obscured when stitching datasets derived from multiple individuals, rather than following one individual's development over time.

The second involves molecular recorders that go beyond lineage to record phenomena such as transcriptional activity, cell state, signaling activity, cell-cell communication, etc.[26,27] Such methods enable reconstructing cell lineage histories in a way that recovers the identities not only of terminal nodes but also of inferred ancestors.

We envision that these technologies can be combined. Imagine that a series of progressively older embryos is subjected to a flavor of molecular recording that yields a comprehensive lineage tree, with rich molecular states for terminal nodes (e.g., scRNA-seq) but also information about the molecular identities of inferred ancestors. Such trees could be merged across individuals to yield a molecularly annotated consensus ontogeny of a given species (Figure 2). The output of such an integration exercise might resemble what was beautifully achieved by Packer et al. for *C. elegans* by merging transcriptional pseudotrees with the Sulston lineage.[18]

Assuming it can be constructed, a molecularly annotated consensus ontogeny would provide an excellent framework for cell type classification. First, it could be generated in an entirely data-driven, fully describable way. Second, both *variation within* (e.g., due to stochastic factors, genetic variation) or *deviation from* (due to disease) the consensus could be represented by "summary statistics" of individual branches or by alternative branches, respectively. Third, in contrast to current atlases which are biased toward specific systems by an amalgam of factors, a consensus ontogeny spanning the life cycle would include *all* "normal" cell states and relate them to one another. Finally, as this consensus ontogeny is rooted in a naturally bounded, comprehensive, reproducible process, it would represent a stable backbone onto which additional information could be layered.

How would a consensus ontogeny be structured and represented? We recognize that the practical aspects of the proposed concept are likely still vague to most readers. Below we conceptualize a potential structure for a consensus ontogeny, address the integration of individuals and data types, and suggest a data-driven cell type nomenclature. We then provide a practical example of how we envision experimentally deriving and visually representing such a tree. Finally, we address outstanding challenges for the proposed framework, in particular for *H. sapiens*, and possible solutions.

How might a consensus ontogeny of cell types be structured and represented?

In considering how dense lineage trees might be summarized, we find inspiration in how human demographic histories are modeled by population geneticists (Figure 3), with subsets of

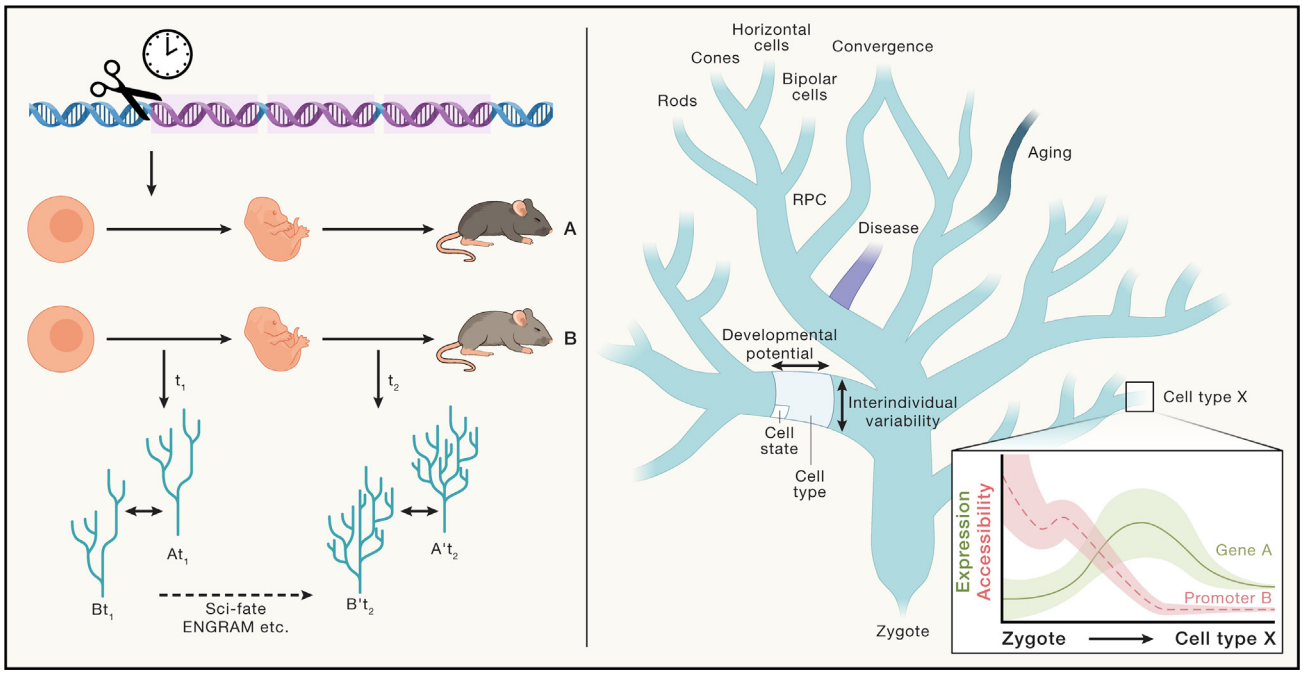

Figure 2. Conceptualizing the construction of a "consensus ontogeny"

Left: The envisioned consensus ontogeny will require: (1) time-resolved lineage tracing, e.g., by writing DNA barcodes that uniquely mark daughter cells, ideally at every cell division, starting at the single-cell zygote stage; (2) co-assays of cell lineage and molecular state at multiple developmental time points; (3) methods to bridge gaps in molecular states between time points, such as *in vivo* sci-fate[28,29] or recording of prior cell states,[26, 27] (4) integration of trees derived from many individuals.

Right: Conceptualization of a "consensus ontogeny" of cell types. In addition to summarizing the entirety of organismal development, such a tree could incorporate any number of molecular state measurements, enabling the representation of phenotypic differences, including both intra-individual and inter-individual variation.

reproductively isolated individuals conceptualized as branch segments of a tree. The parameters defining each branch segment might include the number of founder individuals, their age structure, birth/death rates, etc., while the overall model summarizes how each segment relates to other segments. Building on concepts from Stadler and colleagues,[30] we can imagine summarizing complex cell lineage trees in an analogous manner. Essentially, subsets of cells similar with respect to their past, present, and future (i.e., with respect to both molecular state and lineage), might be "bundled" into branch segments defined by certain parameters (e.g., number of founding cells, proliferation rate, cell division motifs, molecular state, etc.). Within cross-sections of each branch segment, cells might be heterogeneous with respect to cell cycle phase, stochastic differences, etc., while along the branch segment, cells might be heterogeneous with respect to a longitudinal component of continuous differentiation. Aspects of such heterogeneity might be correlated with the fate(s) of any given cell's descendants, analogous to incomplete lineage sorting in demographic histories.

Naturally, there are differences between cellular ontogeny and classic phylogenetics, but these arguably make development easier to model than demographic history. First, there is no recombination between individuals in the case of development. Second, demographic history only happened once, whereas development can be repeated and measured many times under controlled conditions, including intermediary time points. Third, we can experimentally perturb development to probe the processes underlying each feature of the consensus ontogeny. Fourth, whereas ground truth is inaccessible in phylogenetics due to the passage of time, future scientists will have access to the same underlying reality of development as we do; as such, as technologies and computational methods improve, so will the reference.

If each individual has a slightly different lineage history, how do we arrive at a consensus tree for each species? An objective approach would be to define resolution in terms of aspects of the tree that are consistent across individuals, i.e., "bundling" patterns that are invariant across individuals. We also imagine that the comprehensive lineage tree of even one individual will contain abundant "variations on a theme" that can be summarized, e.g., the consistent aspects of how each nephron is built. In fact, it is precisely the characteristics of such lineage trees that are invariant (or at least statistically bounded) that should define the right level of resolution for a "reference cell tree". In contrast with one-off single-cell datasets, such a consensus representation would be a ground truth framework onto which other datasets could be projected, analogous to projection of local chromatin states onto the reference genomes constructed by the Human Genome Project.

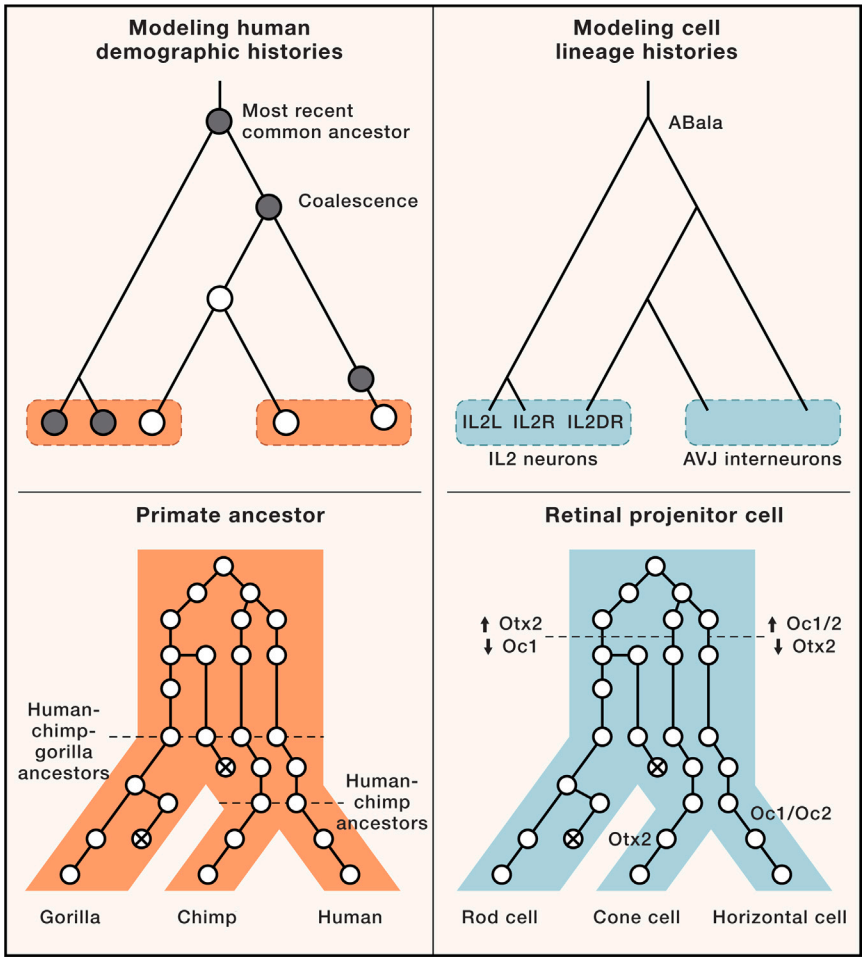

Figure 3. Parallels between modeling demographic and cell lineage histories

Top Left: Coalescent model in a subdivided population, adapted from Masatoshi.[31] Unlike development, no replication of the "experiment" is possible in population genetics, as there is only one run of history available to be studied. In addition, the starting conditions of the experiment are unknown. Therefore, we need to model the past using a suitable stochastic model, the coalescent.[32] Eventually, all lineages coalesce into a single lineage, the most recent common ancestor (MRCA).

Top Right: Schematic of an analogous model of development. All cells and cell types can be traced back to an MRCA, the single-cell zygote. In some cases, molecularly similar cell types originate from well-separated lineages.

Bottom Left: In incomplete lineage sorting, an ancestral gene copy fails to coalesce (looking backwards in time) into a common ancestral copy until deeper than previous speciation events, i.e., the tree produced by a single gene is discordant from the species-level tree.

Bottom Right: The probability that a cell's descendant will follow a certain lineage trajectory can differ before a cell type split occurs, e.g., through fluctuations in TF levels. Crossed circles are "extinct" lineages.

Distinctions from a Waddington landscape and traditional cell lineage trees

A key difference between a reference cell tree and a "reference cell atlas" is that the tree would be continuous, i.e., all segments would be connected to other segments, flowing back to a common root, the single-cell zygote. But it would also differ from the continuous Waddington landscape, the dominant paradigm for representing the molecular basis of cellular differentiation for decades.[33] The Waddington landscape conceptualizes cell types as attractor states and illustrates the probability of developmental trajectories, as well as cell type stability and transition likelihoods, which are inherently challenging to measure in practice. In contrast, a consensus ontogeny of cell types would more explicitly summarize the molecular and lineage paths in this landscape as they are actually traversed in wild-type development.

To be clear, we do not envision the consensus ontogeny as a literal depiction of binary splits representing every cell division like the Sulston tree. A purely bifurcating lineage tree would fail to incorporate concepts such as state divergence and convergence.[34] In contrast, as articulated by Wagner and Klein, state manifolds (i.e., a continuum of cell states) do include these concepts and are arguably more useful than a purely bifurcating lineage tree. A consensus ontogeny, grounded in lineage tracing information and roughly resembling the shape of a tree, would be inclusive of both representations. It would: (1) be able to represent a variety of recurrent motifs including but not limited to binary splits, such as asymmetric cell divisions,[35] transcriptional state convergence, recurrent "subprograms", etc.; (2) not seek to represent every cell division of an individual's development, but rather bundle the key patterns, both within and across individuals, into summary representations, which take the form of branch segments; and (3) leave room for "demographic" aspects of development, such as how many cells are the "founders" of each branch, clonal dominance, etc.

Layering on additional molecular phenotypes

By practical necessity, the envisioned consensus ontogeny would require at least some systematically acquired representation of the molecular phenotypes of cells at any position throughout the tree, to facilitate integration across individuals, the bundling of recurrent patterns into cell types, relating these to existing cell type nomenclature, etc. However, it would also represent a backbone onto which more molecular measurements could be added. Which molecular phenotypes would best enhance the tree? Currently we can measure mRNA, chromatin accessibility, protein, and epigenetic marks, as well as spatial location, at single-cell resolution. The dynamic range for these phenotypes differs within and across cells, as does the ease of measurement and interpretation, and the perceived role and temporal order in changing cell fates. Ultimately, we do not need to decide at this point which molecular phenotypes

to include. Indeed, a key benefit of such a reference, grounded in the reproducible process of development, is that even decades from now, measurements we haven't conceived of yet could be layered on to it.

Single-cell co-assays of any molecular characteristic and lineage would be sufficient for its incorporation into a reference tree. Co-assays already exist for mRNA and lineage and chromatin accessibility and lineage, although not yet at the required temporal resolution or depth of coverage.[36,37,38] However, given how rapidly single-cell and genome-editing technologies are advancing, we are optimistic that at least for model organisms, dense lineage trees in which the transcriptome is measured in nearly all endpoint cells are within reach. Once the consensus ontogeny is constructed from many such trees, we may no longer need to co-measure lineage alongside molecular phenotypes in order to place cells with reasonable confidence to specific locations on the tree.

In addition to molecular phenotypes, our tree needs to explicitly accommodate further continuous forms of heterogeneity that tend to be discarded by discretized cellular taxonomies, e.g., spatial gradients and biased differentiation potentials. Biased differentiation potentials may be linked to spatial position, possibly but not necessarily reflected in a cell's current transcriptome. *C. elegans,* for which we know ground truth, provides some insight about congruence (or lack thereof) between these features. A systematic comparison of anatomical position and single-cell transcriptomes of the anatomically defined canonical 118 neuron classes revealed 128 distinguishable molecular states, mostly agreeing but in some cases grouping anatomical classes together or revealing additional subtypes.[39] Other cases of lineage-specific priming establishing L/R asymmetries in the gene expression of related neuron pairs in *C. elegans* have been described.[40] It is likely that differences between related cells in different anatomical positions will be even more pronounced in more complex organisms across all their developmental time points; accordingly we need to develop new ideas to represent spatial information within a consensus ontogeny. Fortunately, apart from cells that extensively migrate or circulate, lineage history and anatomical position should be correlated such that a nomenclature grounded in lineage should actually aid in cataloging spatial information.

To make the tree as informative as possible, we should ultimately adorn it with the key TFs, signaling pathways, etc. responsible for nudging cells in different directions. Increasing evidence argues for a TF "selector code" in cell type specification, where cell types can be defined by unique combinations of TFs that are continuously expressed. For example, the vast majority of 118 *C elegans* neuron classes are characterized by a distinct TF expression pattern.[41] In *Drosophila*, manipulations of the TF code have been shown to enable the complete morphological and transcriptional conversion between neuronal cell types.[42] We propose TF combinations as the most useful additional cell type descriptors rather than the currently used "most differentially expressed" genes or cell-surface markers.

Resolution and nomenclature

In addition to relating cell types in a reference tree to one another in a meaningful way, we also need a system of nomenclature. What is the right level of resolution at which to apply labels? Single-cell measurements bring the temptation of iterative sub-classifications, driven by biological or technical variation where no two cells are identical, to the point of becoming meaningless. For species, resolution is defined by whether individuals are able to interbreed to produce fertile offspring. But for conventional "atlases" of cells, there is no equivalently crisp rule. One can take a systematic approach (e.g., defining cell types as subsets of cells that can be reliably classified via machine learning as belonging to that type as opposed to other types in an atlas[43,44]), but these remain conflated with the quality of the data (i.e., the number of cell types would grow with higher quality scRNA-seq data).

We suggest a nomenclature grounded, at its highest level, in the conventional name of cells within the branch segment (e.g., hepatocytes), with some arbitrary label appended if multiple branch segments share the same name (e.g., hepatocyte-A). More granular labels could be applied to cells based on their sample-of-origin or progression along the branch segment. For example, hepatocyte-A-E14-65 might be a particular kind of hepatocyte that was derived from an E14 mouse and had progressed 65% of the way along the branch segment in which it resides. As we learn more about the key TFs defining the progression across different branch segments, we can optionally incorporate these into the nomenclature describing an individual cell (e.g., hepatocyte-A-E14-65-TF.XYZ). In addition (or alternatively), terms could be added that summarize where any given cell falls with respect to the principal components of heterogeneity within its branch segment. In summary, the nomenclature would attempt to convey maximal information about a cell's past, present, and future with a minimal number of terms.

As noted, we are not suggesting to completely replace currently used cell type names by an entirely new naming scheme, as this would both be impractical and hurt our ability to relate scientific findings to the past. However, given the inexactness of contemporary nomenclature, we strongly argue for these terms to be pinned to specific branch segments, such that we end up with a systematic nomenclature whose construction is data-driven and fully describable, and moreover inclusive of all cell types that arise during the natural life cycle of that organism. This approach will also allow us to differentiate between progenitor populations, establish an objective definition of what constitutes a novel cell type, and systematically identify the molecular characteristics that are most predictive of a cell belonging to a particular type.

Grounding the concept in a specific example

To better ground the concept of a consensus ontogeny, let us consider how one could derive a particular branch, the murine hematopoiesis lineage. A CRISPR-based lineage-tracing system can be incorporated into hematopoietic stem cells by activating the cassette with a lineage-specific promoter in transgenic mice, or in this specific case by *ex vivo* transduction and transplantation. Because these methods can now perform continuous recording for weeks with multiple events per cell division and capture scRNA-seq profiles from the same cells,[45] it is plausible that we will soon be able to record every cell division within a single individual. To obtain a consensus ontogeny, this would need

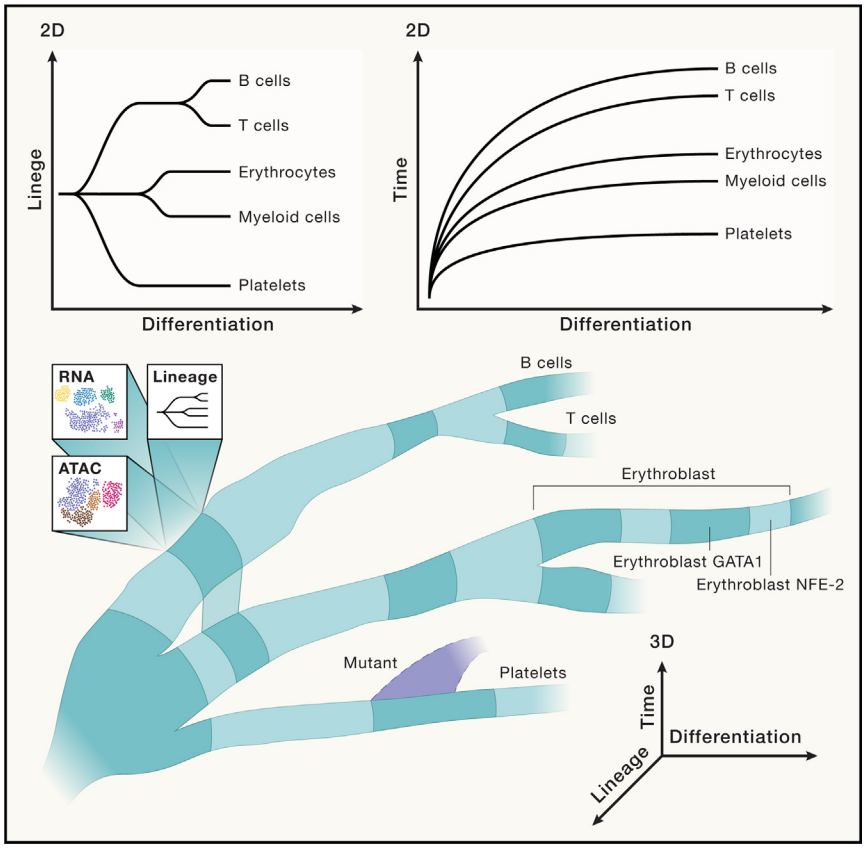

Figure 4. A schematic of a consensus ontogeny for fetal hematopoiesis

The overall tree structure represents the shared developmental trajectories[49] of cells belonging to a certain cell lineage (e.g., erythrocytes, platelets etc.), but does not depict each individual cell division as it occurs in an individual organism. The consensus ontogeny could be visualized as a 3D tree-shaped structure (bottom), where the coordinates represent cell lineage (while allowing for divergence and convergence events between branches), differentiation state along a certain cell lineage branch, and biological time. Branches could be automatically segmented into "cell types" along the differentiation axis, e.g., based on maximum information gain, and annotated according to key distinctive features. Although the proposed visualization is based primarily on averaged developmental lineage information, associated molecular data is used for the branch segmentation and can be visualized for each single cell for arbitrary cell types. An exemplary branch of a mutant mouse which manifests in both delayed and arrested platelet development is shown (dotted line).

to be repeated over many time points and individuals, because measuring molecular profiles of the cells for whom we establish lineage is only possible by terminating the experiment at a single time point. Given the pace of technology, we do not anticipate the required number of cells, time points, and individuals will be limiting. To computationally construct a consensus ontogeny, one would start from the purely lineage-derived tree for each individual using phylogenetic algorithms and then bundle these in a way that captures the patterns that are invariant across individuals. Toward a similar end, algorithms have recently been described that reconstruct time-scaled phylogenies with estimates of population sizes and commitment biases of progenitor states from lineage barcodes.[46]

One possible visual representation of the consensus ontogeny would have cell lineage (e.g., erythrocyte-A), differentiation state (90% progression along branch) and time (E14) as three axes (Figure 4). We would use the molecular information collected alongside developmental lineage to both characterize differentiation states and automatically determine borders of branch segments along this axis, with each such segment comprising a cell type e.g., based on objective criteria such as maximum information gain. An algorithm commonly used to build decision trees based on information entropy is C4.5, which works for both non-binary choices and continuous categories and on multi-omics data.[47] Computational approaches to build decision trees for cell type classification from a combination of lineage and single-cell genomics data have been reported.[48] They also allow the automatic definition of the most informative features (mostly TFs) at each decision point, which we could use for automatically naming individual branch segments (e.g., erythroblast-A-GATA1, erythroblast-B-NFE2). Heterogeneity within branch segments could be summarized via principal components analysis of gene expression profiles.

The consensus ontogeny could be displayed in 3D on an interactive website and allow researchers to explore segments of interest, whose molecular states for all contained cells could then be visualized, e.g., as manifolds. Providing pre-computed differentially expressed genes compared to other specific stages, lineages, or all other lineages could be easily incorporated. Starting from constructing sub-structures for individual organs or physiological systems, we can work toward assembling a full consensus ontogeny for an organism, similar to how the reference genome started from individual contigs. Since we are not planning on representing the explicit cell divisions of any given individual, we can combine branches derived in different experiments/individuals as long as there is overlap. Mutants or disease states could be readily incorporated into this framework if it is possible to apply lineage-tracing approaches. They might diverge from the wild-type tree in any of the three major axes, such as aberrations in time spent to reach a certain state, the actual differentiation state reached, or even giving rise to entirely new lineages. In particular for mutants (e.g., KO mice), we foresee "tree thinking"[30] as being highly useful for studying the origins of the underlying defect(s). When it is not possible to apply lineage-tracing approaches to pathological states, cell states with only molecular information could still be mapped onto the lineage tree by using inferred trajectories. Similarly, we could devise strategies to compare individual branches,

sub-structures, or entire trees across organisms, incorporating the learned lineage information, but largely building on computational approaches that are currently being developed for molecular states.[24,50,51,52,53]

What about *H. sapiens*?

A primary weakness of this proposal is our inability to apply genome-editing-based lineage-tracing methods to derive a reference ontogeny for humans. There are at least four avenues to ameliorate this: (1) Generation of consensus ontogenies of closely related organisms (e.g., rhesus macaque); (2) *in vitro* human "stembryo" models[54]; (3) lineage tracing based on somatic mutations in chromosomal or mitochondrial DNA[55,56]; and (4) lineage inference based on molecular profiles from human tissues. It is likely that a combination of these approaches will be required to arrive at an approximate human representation. Furthermore, the overwhelming majority of human biology is shared by closely related models. We predict that there will be few instances in which a mouse or macaque ontogeny/nomenclature is insufficient for human data, just as there are hardly any human-specific genes.

OPPORTUNITIES

Approaches of information containment and organization do not merely provide platforms for more efficient data accumulation, but ideally also serve to distill new structural relations and patterns. Besides organizing the vast amount of existing cellular data, unified species-specific cell type trees have the potential to bring to light some key tenets of developmental biology.

In its simplest form, a consensus ontogeny, to which any single-cell dataset could be mapped, would enable biologists to be more precise about which cell populations they are referring to. This would both allow us to synthesize the vast amount of research that is done in different subfields all over the world, as well as to communicate effectively with future scientists. However, beyond the obvious advantage of a common, stable nomenclature, a reference tree would provide biological insights on its own. First, it would lay the foundation for the nomination and validation of factors that shape specific cell type transitions. Such insights could facilitate and expedite the efficient and faithful derivation of those cell types in the laboratory, whether for basic science or therapeutic purposes. Second, it would support the systematic "placement" of *in vitro* systems (e.g., stembryos, organoids) in relation to wild-type development. Third, it would facilitate characterization of the statistical properties of inter- and intra-individual phenotypic variation, whether disease-related or not. For diseases in which developmental processes are directly (e.g., cleft lip) or indirectly (e.g., fewer nephrons → hypertension) involved, the moment of causality may trace back to statistical deviations from the normal distribution of particular branch segments, e.g., too few progenitors of a given type, or arrested development. Such insights could provide a whole new understanding of how disease phenotypes arise, including Mendelian disorders as well as endophenotypes that shape common human disease risk. Fourth, it would facilitate trees, branches and cell types to be aligned across species, e.g., to understand the origins of cell types, evolutionary innovations, etc., in a systematic manner.

Omnis cellula e cellula

Since around 2015, single-cell technologies that measure diverse analytes in individual cells have blossomed. Computational efforts have been focused on clustering cells in each individual dataset by their similarity in one dimension (e.g., mRNA), in order to partition the data into digestible subsets (cell types) for further analysis. More recently, this has been expanded to aligning datasets generated by different methods or labs as well as integrating additional molecular phenotypes. This standard workflow, although essential to our learning curve, presents a challenge when it comes to cell type definitions and annotations. The biggest issue, in our view, is that the resulting corpus is heavily biased toward the systems in which the data is being generated, rather than being anchored in a "real world" distribution. As a field, we need to move from simply performing dimensionality reduction of our burgeoning data in isolation, and rather toward building a shared understanding of the underlying reality of multicellular biological systems. We argue that wild-type development, a reproducible process that subsumes a large part of the diversity and dynamics that we are interested in, provides an excellent scaffold for this goal.

Given the speed and breadth of innovation in this space, we believe a future where we can rival the precision of *C. elegans* cell type descriptors is attainable in more complex organisms. The human reference genome revolutionized biology in the 2000s by providing both a centralized coordinate system for reporting and comparing results across studies and a basis for understanding the regulatory mechanisms underlying gene expression. In its current state, the reference genome is a linear composite of merged haplotypes from less than two dozen people. It contained gaps, biases, and errors and did not accurately reflect global human genomic variation, leading to calls for a more sophisticated and complete human pangenome with a graph-based representation of genomic diversity.[57] Yet despite these shortcomings, that first draft has been invaluable for the acceleration of biological discoveries. We believe that in analogy to the human reference genome, a consensus ontogeny of cell types, both for key model organisms as well as for the human species, will provide an essential shared framework for our study of cell fate specification in development and disease (Figure 5). As it becomes more sophisticated, we can derive related learnings about the origins of natural and disease-related phenotypic variation, as well as start to understand the building blocks and fundamental regulatory logic of how cell types and states come to be.

Like Linnaeus, who did not know about evolutionary relationships when he proposed his species classification system yet was faced with information overload and a need to install an infrastructure in order to accumulate, process, and retrieve the bits of factual information,[4] we need to start somewhere and fill in and structure the pages we already have while leaving room for large amounts of missing data. Although this step may seem daunting given how much we don't know at this moment, it is useful to keep in mind how well we have been served by Linnaean taxonomy, even if new species are still in the process of being discovered and named to this day. A unified

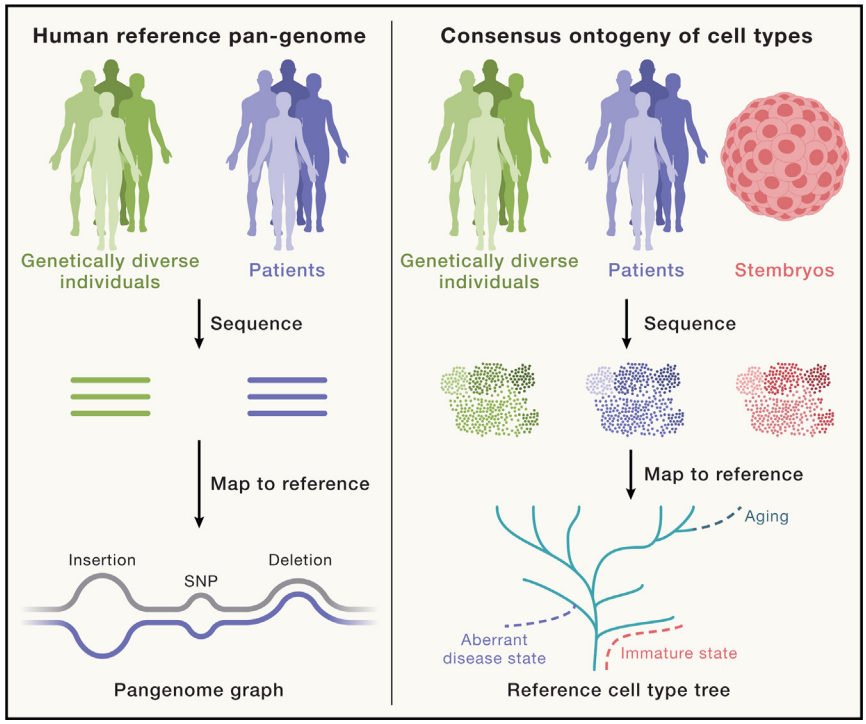

Figure 5. Graph-based references enable representation of intra- and inter-individual differences in both genotype and phenotype

Left: The human pan-genome aims to provide a more accurate and diverse representation of global genomic variation, including in repetitive regions of the genome, and to improve gene-disease association studies across populations.[57] Starting with a conventional reference genome, variants are added as additional branches which depart from the reference sequence but later rejoin it. Each branch can be associated with an allele frequency and the graph structure can be updated as new haplotypes are discovered.

Right: In analogy, a consensus ontogeny would enable comprehensive representation of intra- and inter-individual variation in the context of health and disease and would also allow accurate mapping of *in vitro* stembryo and organoid lineages.

reference tree would be a stable scaffold to which additional information, both about health and disease states, could be accrued over the decades.

There are undoubtedly logistical challenges associated with the creation of a consensus ontogeny and universal nomenclature, e.g., which species, who will build it, who will decide on the nomenclature, who will maintain it, etc. However, we don't think the challenges are insurmountable. Single-cell data is already being widely shared in an open-source manner by the community, and tools are being developed for optimal integration of all available datasets of given systems,[58] so there is no reason to believe that it will be different once we start accommodating lineage information. Though there is currently no specific consortium or expert panel in place that is focusing on this question, we believe it is highly worthy of pursuing an endpoint that would ensure that we have a precise language to communicate about biology. Initial drafts will be incomplete but can be updated as data quality improves. This is exactly the strength of the lineage-based approach, in that the overall coarse tree structure will be stable over time, even as resolution increases and additional molecular phenotypes are added. C57BL/6 mice did, do, and will develop the same way in 1983, 2023 and 2063.

Although in our view a consensus ontogeny is the end goal, this does not mean that developmental atlases based on purely molecular measurements and trajectories, which are attainable in the present, aren't useful or worthy of constructing.[59,60] Since such methodologies and large datasets are already available whereas further refinement of lineage-tracing technology is required for constructing dense cell type trees, we instead advocate for using the molecular information we have at hand in the present. At the same time, however, we should start a concerted effort to construct data-driven, consensus ontogenies of cell types that span the development of vertebrate model organisms, however coarse and incomplete they may be initially.[22,23,24,61]

In *The Order of Things*, Foucault laid out the case that each historical era brings with it a different conception of what it is to know and this, in turn, is grounded in each epoch's experience of order.[1] These unifying sets of rules for forming knowledge are *epistemes* that "define the conditions of possibility of all knowledge, whether expressed in a theory or silently invested in a practice." Order in the Renaissance episteme was based on subjective resemblance, exemplified by the Aristotelian taxonomy (e.g., animals that live on water vs. animals that live on land). This was followed by the episteme of the Classical era, where phenomena were broken down into their constituent elements and systematically differentiated from others, focusing on classifications of external features as exemplified by Linneaen taxonomy, with the world a place of differences rather than similarities. In contrast, the Modern episteme relies on history, not order, and relates "discontinuous but analogous elements in such a way that they are then able to establish causal relations and structural constants between them". The basic shift here is moving from organizing things in a table according to identity and differences (Classical) to relating them to one another according to functional analogies and temporal succession (Modern). In a sense, this is what shifting the goal from a periodic table to a consensus ontogeny would achieve, by not only naming and ordering cells based on differences along a visible axis (molecular state) but also by relating those to their invisible relationships across time (lineage).

ACKNOWLEDGMENTS

We thank M. Elowitz, A. Schier, C. Trapnell, and R. Waterston, as well as C. Qiu and other members of the Shendure Lab for helpful conversations on this broad topic. We also thank the reviewers for their constructive comments on the original version of this manuscript.

DECLARATION OF INTERESTS

S.D. is currently also affiliated with Gordian Biotechnology. The affiliation started after the work was submitted and the work isn't being done in association with the affiliation.

REFERENCES

1. Moore, F.C.T., and Foucault, M. (1971). The order of things: An archaeology of the human sciences. Man *6*, 421.
2. Clutton-Brock, J. (1995). Aristotle, The Scale of Nature, and Modern Attitudes to Animals. Soc. Res *62*, 421–440.
3. von Linné, C. (1767). Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis (Vindobonae, Typis Ioannis Thomae von Trattner) https://doi.org/10.5962/bhl.title.156783.
4. Müller-Wille, S., and Charmantier, I. (2012). Natural history and information overload: The case of Linnaeus. Stud. Hist. Philos. Biol. Biomed. Sci. *43*, 4–15.
5. Virchow, R. (1860). Cellular Pathology: As Based Upon Physiological and Pathological Histology. Twenty Lectures Delivered in the Pathological Institute of Berlin During the Months of February, March and April, 1858 (London John Churchill MDCCCLX).
6. Ramón, S., and Cajal, Y. (1904). Textura del Sistema Nervioso del Hombre y de los Vertebrados (Imprenta y Librería de Nicolás Moya, Madrid).
7. Zeng, H. (2022). What is a cell type and how to define it? Cell *185*, 2739–2755.
8. Clevers, H., Rafelski, S., Elowitz, M., Klein, A., Shendure, J., Trapnell, C., Lein, E., Lundberg, E., Uhlen, M., Martinez-Arias, A., et al. (2017). What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? Cell Systems *4*, 255–259.
9. Bard, J., Rhee, S.Y., and Ashburner, M. (2005). An ontology for cell types. Genome Biology *6*, R21. https://doi.org/10.1186/gb-2005-6-2-r21.
10. Morris, S.A. (2019). The evolving concept of cell identity in the single cell era. Development *146*. https://doi.org/10.1242/dev.169748.
11. Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., et al. (2016). The origin and evolution of cell types. Nat. Rev. Genet. *17*, 744–757.
12. Xia, B., and Yanai, I. (2019). A periodic table of cell types. Development *146*. https://doi.org/10.1242/dev.169854.
13. Miller, J.A., Gouwens, N.W., Tasic, B., Collman, F., van Velthoven, C.T., Bakken, T.E., Hawrylycz, M.J., Zeng, H., Lein, E.S., and Bernard, A. (2020). Common cell type nomenclature for the mammalian brain. Elife *9*. https://doi.org/10.7554/eLife.59928.
14. Pasquini, G., Rojo Arias, J.E., Schäfer, P., and Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. Comput. Struct. Biotechnol. J. *19*, 961–969.
15. Wang, J., Zhang, K., Xu, L., and Wang, E. (2011). Quantifying the Waddington landscape and biological paths for development and differentiation. Proc. Natl. Acad. Sci. USA *108*, 8257–8262.
16. Michielsen, L., Lotfollahi, M., Strobl, D., Sikkema, L., Reinders, M.J.T., Theis, F.J., and Mahfouz, A. (2022). Single-cell reference mapping to construct and extend cell type hierarchies. bioRxiv. https://doi.org/10.1101/2022.07.07.499109.
17. Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol *100*, 64–119.
18. Packer, J.S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., et al. (2019). A lineage-resolved molecular atlas of embryogenesis at single-cell resolution. Science *365*. https://doi.org/10.1126/science.aax1971.
19. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*. https://doi.org/10.1126/science.aaw3381.
20. McKenna, A., and Gagnon, J.A. (2019). Recording development with single cell dynamic lineage tracing. Development *146*. https://doi.org/10.1242/dev.169730.
21. Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498.
22. Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., and Klein, A.M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science *360*. https://doi.org/10.1126/science.aar5780.
23. Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science *360*, 981–987.
24. Qiu, C., Cao, J., Martin, B.K., Li, T., Welsh, I.C., Srivatsan, S., Huang, X., Calderon, D., Noble, W.S., et al. (2022). Systematic reconstruction of cellular trajectories across mouse embryogenesis. Nat. Genet. *54*, 328–341.
25. Calderon, D., Blecher-Gonen, R., Huang, X., Secchia, S., Kentro, J., Daza, R.M., Martin, B., Dulja, A., Schaub, C., Trapnell, C., et al. (2022). The continuum of embryonic development at single-cell resolution. Science *377*, eabn5800.
26. Farzadfard, F., and Lu, T.K. (2018). Emerging applications for DNA writers and molecular recorders. Science *361*, 870–875.
27. . Preprint at Chen, W., Choi, J., Nathans, J.F., Agarwal, V., Martin, B., Nichols, E., Leith, A., Lee, C., and Shendure, J. (2021). Multiplex genomic recording of enhancer and signal transduction activity in mammalian cells bioRxiv. https://www.biorxiv.org/content/10.1101/2021.11.05.467434v1.
28. Lu, Z., Zhang, M., Lee, J., Sziraki, A., Anderson, S., Ge, S., Nelson, P.T., Zhou, W., and Cao, J. (2022). A comprehensive view of cell-type-specific temporal dynamics in human and mouse brains. Preprint at bioRxiv. https://doi.org/10.1101/2022.10.01.509820.
29. Cao, J., Zhou, W., Steemers, F., Trapnell, C., and Shendure, J. (2020). Sci-fate characterizes the dynamics of gene expression in single cells. Nature Biotechnology *38*, 980–988. https://doi.org/10.1038/s41587-020-0480-9.
30. Stadler, T., Pybus, O.G., and Stumpf, M.P.H. (2021). Phylodynamics for cell biologists. Science *371*. https://doi.org/10.1126/science.aah6266.
31. Masatoshi, N. (1989). Molecular Evolutionary Genetics. By Nei MasatoshiNew York. In Genetics Research, *54* (Columbia University Press), p. 243. https://doi.org/10.1017/s0016672300028767.
32. Rosenberg, N.A., and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. *3*, 380–390.
33. Waddington, C.H. (1957). The Strategy of the Genes: A Discussion of some Aspects of theoretical Biology (George Allen & Unwin).
34. Wagner, D.E., and Klein, A.M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet *21*, 410–427.
35. Cepko, C. (2014). Intrinsically different retinal progenitor cells produce specific types of progeny. Nat. Rev. Neurosci. *15*, 615–627.
36. Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. Nat. Biotechnol. *36*, 469–473.
37. Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat. Biotechnol. *36*, 442–450.
38. . Preprint at Minkina, A., Cao, J., and Shendure, J. (2022). Tethering distinct molecular profiles of single cells by their lineage histories to investigate sources of cell state heterogeneity bioRxiv. https://doi.org/10.1101/2022.05.12.491602.

39. Taylor, S.R., Santpere, G., Weinreb, A., Barrett, A., Reilly, M.B., Xu, C., Varol, E., Oikonomou, P., Glenwinkel, L., et al. (2021). Molecular topography of an entire nervous system. Cell *184*, 4329–4347.

40. Charest, J., Daniele, T., Wang, J., Bykov, A., Mandlbauer, A., Asparuhova, M., Röhsner, J., Gutiérrez-Pérez, P., and Cochella, L. (2020). Combinatorial action of temporally segregated transcription factors. Dev. Cell *55*, 483–499.

41. Hobert, O., Glenwinkel, L., and White, J. (2016). Revisiting Neuronal Cell Type Classification in Caenorhabditis elegans. Curr. Biol *26*, R1197–R1203.

42. . Preprint at Özel, M.N., Gibbs, C.S., Holguera, I., Soliman, M., Bonneau, R., and Desplan, C. (2022). Coordinated control of neuronal differentiation and wiring by a sustained code of transcription factors bioRxiv. https://doi.org/10.1101/2022.05.01.490216.

43. Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol *20*, 194.

44. Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*. https://doi.org/10.1126/science.aba7721.

45. Choi, J., Chen, W., Minkina, A., Chardon, F.M., Suiter, C.C., Regalado, S.G., Domcke, S., Hamazaki, N., Lee, C., Martin, B., et al. (2022). A time-resolved, multi-symbol molecular recorder via sequential genome editing. Nature *608*, 98–107.

46. Fang, W., Bell, C.M., Sapirstein, A., Asami, S., Leeper, K., Zack, D.J., Ji, H., and Kalhor, R. (2022). A general framework for analyzing progenitor state dynamics via retrospective lineage barcoding. Cell *185*, 4604–4620. Quantitative fate mapping.

47. Ross Quinlan, J. (2014). C4.5 (Programs for Machine Learning (Elsevier)).

48. Veleslavov, I.C., and Stumpf, M.P.H. (2020). Decision tree models and cell fate choice. Preprint at bioRxiv. https://doi.org/10.1101/2020.12.19.423629.

49. Pucella, J.N., Upadhaya, S., and Reizis, B. (2020). The Source and Dynamics of Adult Hematopoiesis: Insights from Lineage Tracing. Annu. Rev. Cell Dev. Biol. *36*, 529–550.

50. Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., et al. (2022). Mapping single-cell data to reference atlases by transfer learning. Nat. Biotechnol. *40*, 121–130.

51. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol *36*, 411–420.

52. Wang, J., Sun, H., Jiang, M., Li, J., Zhang, P., Chen, H., Mei, Y., Fei, L., Lai, S., et al. (2021). Tracing cell-type evolution by cross-species comparison of cell atlases. Cell Rep. *34*, 108803.

53. Yuan, M., Yang, X., Lin, J., Cao, X., Chen, F., Zhang, X., Li, Z., Zheng, G., Wang, X., et al. (2020). Alignment of Cell Lineage Trees Elucidates Genetic Programs for the Development and Evolution of Cell Types. iScience *23*. https://doi.org/10.1016/j.isci.2020.101273.

54. Moris, N., Anlas, K., van den Brink, S.C., Alemany, A., Schröder, J., Ghimire, S., Balayo, T., van Oudenaarden, A., and Arias, A.M. (2020). An in vitro model of early anteroposterior organization during human development. Nature *582*, 410–415. https://doi.org/10.1038/s41586-020-2383-9.

55. Bizzotto, S., and Walsh, C.A. (2022). Genetic mosaicism in the human brain: from lineage tracing to neuropsychiatric disorders. Nat. Rev. Neurosci *23*, 275–286.

56. Ludwig, L.S., Lareau, C.A., Ulirsch, J.C., Christian, E., Muus, C., Li, L.H., Pelka, K., Ge, W., Oren, Y., Brack, A., et al. (2019). Lineage tracing in humans enabled by mitochondrial Mutations and single-cell genomics. Cell *176*, 1325–1339. https://doi.org/10.1016/j.cell.2019.01.022.

57. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446.

58. Sikkema, L., Strobl, D., Zappia, L., Madissoon, E., Markov, N.S., Zaragosi, L., Ansari, M., Arguel, M., Apperloo, L., et al. An integrated cell atlas of the human lung in health and disease. Preprint at bioRxiv. https://www.biorxiv.org/content/10.1101/2022.03.10.483747v1

59. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554.

60. Bergen, V., Soldatov, R.A., Kharchenko, P.V., and Theis, F.J. (2021). RNA velocity-current challenges and future perspectives. Mol. Syst. Biol. *17*, e10282.

61. Mittnenzweig, M., Mayshar, Y., Cheng, S., Ben-Yair, R., Hadas, R., Rais, Y., Chomsky, E., Reines, N., Uzonyi, A., Lumerman, L., et al. (2021). A single-embryo, single-cell time-resolved model for mouse gastrulation. Cell *184*, 2825–2842.