A lack of distinct cell identities in single-cell measurements: revisiting Waddington's landscape

Breanne Sparta[1,2.*], Timothy Hamilton[1,3*], Serena Hughes[1,3], and Eric J. Deeds[1,2.3]

[1]Institute for Quantitative and Computational Biosciences, University of California, Los Angeles,
[2]Department for Integrated Biology and Physiology, University of California, Los Angeles,
[3]Bioinformatics Interdepartmental Program, University of California, Los Angeles
*These authors contributed equally to this work
E-mail: deeds@ucla.edu

Abstract

A prevailing interpretation of Waddington's landscape is that distinct cell types with distinct physiologies are produced and stabilized by dynamical attractors in gene expression space. This notion is often applied in the analysis of single-cell omics data, where cells are clustered into groups that represent cell types, prior to downstream analyses like differential gene expression. Until the advent of single-cell measurement technologies, however, it has been impossible to characterize the heterogeneity of cells in the neighborhoods of these attractor states. In this work, we apply graph theory to characterize the distribution of cells in epigenetic space, using data from various tissues and organisms as well as various single-cell omics technologies. Rather than finding distinct clusters of cells that map cell types to specific regions of epigenetic space, we found that cells of very distinct types and lineages occupy the same region of space. Further, we found that the density distribution of cells is approximately power-law, with most cells existing in low-density regions, very far from other cells. This highly heterogeneous density distribution is unexpected, as it is not consistent with the distributions we would expect to see in the neighborhood of an attractor. We found these two observations are universal in single-cell data on epigenetic state of multicellular organisms, regardless of the tissue, organism, measurement technique employed, or the approach used to select the subset of genes on which the analysis was performed. The fact that currently-available single-cell data is inconsistent with the predictions of Waddington's landscape poses a challenge both for the robust analysis of these data and for our overall understanding of epigenesis in development.

Introduction

Genome-wide, single cell technologies resolve levels of biological organization that are obscured in other measurement modalities[1]. For example, single-cell RNA-sequencing (scRNA-seq) provides transcriptome-wide information about mRNA levels in tens of thousands to millions of individual cells[2]. This technique is often applied to characterize changes in cell-type specific gene expression during development or in response to external tissue perturbations. However, since scRNA-seq tools are inherently destructive, *a priori* knowledge of the cell type is typically lost. Without access to the corresponding morphology or cell surface markers, the majority of studies aim to identify physiologically relevant cell types based on similarities in gene expression patterns alone[3]. It is thus a common practice to apply clustering algorithms that group together cells that are in some sense "close" to one another in gene expression space[4]. This is extremely intuitive: cells with similar mRNA levels should have similar protein levels, similar levels of the functions carried out by those proteins, and thus should be similar phenotypically.

The idea that groups of physiologically similar cells should cluster together in gene expression space also directly follows from the classic picture of "Waddington's epigenetic landscape." This landscape has shaped the dominant paradigm for understanding the molecular basis of development for the past 80 years[5]. While Waddington's original proposal for this epigenetic landscape was abstract, this landscape has generally been interpreted as corresponding to gene expression space[6]. In this picture, cells move through a landscape of epigenetic constraints, descending through valleys that guide the production of terminally differentiated cell fates (Fig. 1A). Here, gene regulatory networks underly each cell fate, and their regulatory interactions generate dynamical attractors in gene expression space that ensure the production of functional cell types. These attractors structurally explain how the expression states of cell types can be *stable* (Fig 1B), and propose a molecular basis for "canalization" in development. Canalization ensures the developmental process is robust, such that cell lineages and terminally differentiated cell types do not switch fates despite internal gene expression noise or small environmental perturbations. Thus, Waddington's landscape hypothesizes a molecular basis of cellular differentiation that rests upon the idea of cell types as attractor states, and predicts that phenotypically similar cells should have similar molecular compositions.[7]

While clustering is nearly universally applied to identify cell types in scRNA-seq studies, these studies have also revealed high levels of heterogeneity in the gene expression states of cells; indeed, "heterogeneity" is an extremely common term in the titles of papers reporting scRNA-seq results[8–10].To date, however, there has been no attempt to directly characterize the structure of cell types in the underlying epigenetic space, nor to quantify the heterogeneity of cells in the neighborhood of cell-type attractors. In part, this is due to the fact that most analysis and clustering of scRNA-seq data is carried out after the application of a number of linear and non-linear transformations of the data, including normalization, log-transformation, PCA, and oftentimes non-linear dimensionality reduction tools like t-SNE and UMAP. These transformations dramatically alter the variance structure of the data[11]. They also generate large levels of distortion in local neighborhoods, so that cells that are neighbors in the original data set are not neighbors after these transformations are applied[12]. While these transformations aim to resolve issues of dimensionality and measurement noise in the analysis of scRNA-seq data, there is a distinct lack of empirical and theoretical evidence to support the use of these transformations in creating the appropriate basis for identifying physiologically similar groups of cells.

In this work, we began by characterizing the relationships between cells in the original, high-dimensional, genome-wide epigenetic data provided by a variety of single-cell technologies. To do this, we developed a straightforward approach that we term "ϵ networks," where we consider the types of cells that are within a certain distance cutoff (ϵ) of each cell in the dataset (Fig. 1C). Our findings revealed something surprising: none of the data that we analyzed is consistent with the structure predicted by Waddington's landscape. For instance, we first characterized a classic data set for human peripheral blood mononuclear cells (PBMCs) where a cell type identity was assigned orthogonally to each cell type using FACS before sequencing was carried out[13]. Rather than finding distinct clusters of cells that map cell types to specific regions of gene expression space, we find that cells of very distinct types and lineages occupy the same region of the space. This is true even when using popular approaches, like the Highly Variable Genes method, or even a supervised approach, to identify genes that should be useful in separating biologically distinct groups of cells[14].

In most scRNA-seq data sets we of course do not know the cell type of each cell ahead of time. Distinct groups of cells should nonetheless yield specific patterns in the sizes of the clusters within our ϵ networks. Interestingly, none of the scRNA-seq datasets we analyzed showed any evidence of distinct cell-type groups, including data from complex and mostly post-mitotic tissues like the brain and even whole organisms like *C. elegans* and hydra[15,16]. scRNA-seq data is known to be noisy, however, largely due to the low capture probability of individual transcripts within the cell[17]. It could be that this noise simply washes away the signal of differences between cell type groups. We found the same lack of distinct cell-type groups, however, in data obtained using completely different experimental modalities, including MEFISH, a microscopy-based technique that measures transcript abundance using a combinatorial set of fluorescently-labeled probes and is thought to have much lower levels of noise[18]. We also saw a complete lack of evidence for distinct cell type groups in other forms of single-cell epigenetic data, including measurements of protein levels and chromatin accessibility (scATAC-seq)[19].

Our ϵ network approach also allowed us to characterize how the cells themselves are distributed in gene expression space. Waddington's landscape suggests that most cells should be clustered near the "center" of their cell-type attractors, with fewer and fewer cells the further away one looks (i.e. few cells should be on or near the "tops" of the hills in between valleys on the landscape, Figs. 1A and B). This should lead to a scenario where most cells are in regions of relatively high density, while few cells are in regions of low density. Instead, we found that the densities are distributed as an approximate power law, where the vast majority of cells are found in very low-density regions while a few cells are found in very high-density regions[20]. As with our observations above, we found this "fractal density" distribution in all of the data we analyzed, including a large number of 10x scRNA-seq datasets, data on gene expression collected using MERFISH, data on protein levels, and chromatin accessibility measured by scATAC-seq.

Our findings thus demonstrate that none of the single-cell data on epigenetic state that we analyzed is consistent with the predictions of Waddington's landscape. As mentioned above, however, analyses of these data are not conducted on the raw space, but rather in data that has been subjected to significant sets of non-linear transformations and dimensionality reduction[21,22]. While the dominant interpretation of Waddington's landscape in the literature assumes that the cell-type attractors should exist in the raw epigenetic spaces[7], it could be that these transformations recover the expected attractor-like structure. We thus applied our analysis to scRNA-seq datasets after each step of the common analysis pipeline. With the exception of the FACS-sorted PBMC data, application of this approach to

ten other large datasets did not generate clear cell-type groups, suggesting that current "best practices" for data transformation within the field do not robustly generate data consistent with the Waddington's landscape picture[11,23].

Our findings have broad implications for both the analysis of single-cell genomics data and our overall understanding of the molecular basis of development. scRNA-seq analysis pipelines entail many (ultimately arbitrary) parameter choices, some of which have a large impact on the results of the analysis (see Fig. S5 in the Supplementary Information for a simple example for human PBMCs)[13]. In some cases, analysis of the data does not yield the expected cell-type structure, even after considerable effort[11,23]. Our work suggests that one source of this problem is the underlying structure of the data itself: the transformations employed in these pipelines were never explicitly designed to recover meaningful cell type groups from data where the cell types are distributed in highly overlapping regions of gene expression space with fractal densities. Future work may be able to discover novel transformation pipelines that are more robust and effective at operationally separating cells into meaningful cell types. Beyond that, however, our work shows that available single-cell data on epigenetic state are completely inconsistent with the predictions of Waddington's landscape. Revising this well-established paradigm in light of our findings will represent a major empirical and theoretical challenge for the emerging era of single-cell biology.

Results

A lack of distinct cell groups in single-cell data. As mentioned above, our initial goal in this work was to characterize the variation of cells around the cell-type attractors in gene expression space. The first step in characterizing that variation is to identify these attractors themselves (Fig. 1A). In scRNA-seq studies this is universally done by clustering cells in the dataset after the application of a series of non-linear transformations and dimensionality reduction steps[23]. These transformations dramatically alter the structure of the data; cells that are originally "close" to each other in the raw data are almost universally moved to be far apart after these transformations[12,24]. Transforming the data could thus alter the attractor structure of the landscape. Moreover, several common transformations (particularly "Counts Per Million" or CPM normalization and log-transformation) significantly alter the shapes of the UMI count distributions[24]. If we wish to honestly characterize transcriptional variation, then transformations that alter the variance of the underlying distributions should clearly be avoided.

We thus developed a straightforward approach, which we call an "ϵ network," that allows us to characterize the relationships between cells in the original, high-dimensional data set. To construct an ϵ network, we first choose a cell and consider a ball of radius ϵ around that cell (Fig. 1C, panel 1). In two dimensions, this gives rise to a circle, but in higher dimensions this represents a hypersphere centered on the cell in question. We then draw a similar hypersphere around every cell in the dataset (Fig. 1C, panel 2). If two cells are closer to each other than this radius ϵ (i.e., they lie within each other's hyperspheres and are thus closer to each other than the ϵ cutoff), then we connect those cells in the network (Fig. 1C, panel 3). A more detailed schematic of the construction of ϵ networks is available in the supplement (Fig. S2.1). For any given value of ϵ, we can thus convert our original high-dimensional dataset into a standard undirected graph. This allows us to use the powerful tools of graph

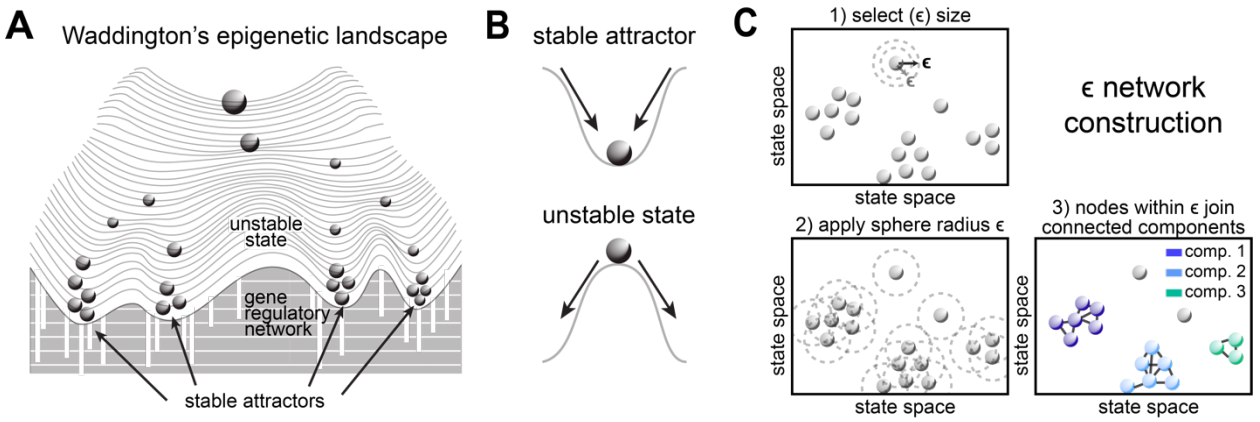

Figure 1. Waddington's Epigenetic Landscape draws predictions from dynamical systems theory. **A.** A schematic of Waddington's landscape, showing how the progression of cells through development is canalized by the underlying gene regulatory networks. The position of a cell on this landscape is interpreted as corresponding to its position in gene expression space, where cells that are constrained to a specific region are expected to be of similar cell types. **B.** Schematic of the dynamical systems that construct the hills and valleys of Waddington's Landscape. The valleys in Waddington's Landscape represent stable attractors, where these attractors function to buffer a cell from perturbations. The hills in Waddington's landscape correspond to unstable states, where any perturbation results in increasing divergence from the original state. **C.** Schematic showing how our ϵ networks are constructed. First, we consider a hypersphere of radius ϵ around any given cell in the dataset (in this 2-D schematic, this corresponds to a circle around each cell, panel 1). We consider a similar hypersphere for every cell in the dataset (panel 2). For each cell, an edge is drawn between the cell and any neighbor that is within the ϵ radius. This generates a network that connects cells based on a pairwise distance criteria (panel 3). Note that, at any given ϵ, the network naturally forms a set of clusters of cells that are all connected to each other, called "components" in graph theory. The largest of these is the giant component.

theory to analyze networks at different values of ϵ, giving us insight into how the cells are distributed in gene expression space.

At any given value of ϵ, the network will naturally partition into different groups of cells that are all connected to each other, giving rise to a set of clusters or "components" (Fig. 1C, panel 3). If the cells are arranged in gene expression space according to the Waddington paradigm, we should see groups of similar cells in distinct regions of the space (see Fig. 2A for a schematic of three distinct cell types). As ϵ increases, cells of the same type should group together first; cells of different types should join together into the same cluster or component only after the radius ϵ exceeds the distance between the groups (Fig. 2A). One way to track this behavior is to look at the size of the largest cluster in the graph (called the "giant component" in graph theory) as a function of increasing ϵ. We should see that the giant component first includes only one of the cell types in the system (Fig. 2B). At larger ϵ radii, we should see the other two groups of cells join all at once. This should give rise to a characteristic step-like behavior in the giant component, where cell types corresponding to distinct attractors are added to the giant component in large groups (Fig. 2B).

To test this idea, we first considered a classic scRNA-seq data set from human Peripheral Blood Monocytes (PBMCs)[13]. In this particular case, the cells were pre-sorted using FACS to separate them into distinct cell-type groups using extremely well-established cell surface markers. Since this was done before sequencing, we know the "true" cell-type label for each cell in the dataset[25], which allows us to not only monitor the size of the giant component, but also its composition. We first considered a mixture of three terminally-differentiated cell types: B cells, Natural Killer (NK) cells, and monocytes. As can be seen from Fig. 2C, the size of the giant component (gray line) shows no step-like behavior whatsoever. Instead, the transition in the giant component is smooth, and involves all the different

cells of the different types joining the giant component more-or-less at the same time (Fig. 2C). We see the same behavior when adding various subsets of T cells from the PBMC data set into the data, suggesting that all the PBMCs purified in this experiment occupy essentially the same region of gene expression space (Fig. S2.2B).

It is important to note that, for this first analysis, we considered all the genes in the genome. This includes ribosomal and metabolic genes that we expect should not vary significantly between cells of different types, and as such it is possible that the cells in this case are overlapping simply because of ubiquitously expressed genes. To test this hypothesis, we considered several "feature selection" approaches that are designed to identify genes that vary in biologically informative ways across the data set. The most popular approach involves finding "Highly Variable Genes" (HVGs) whose variance is higher than one would expect given the average expression level[17]. Interestingly, HVGs did not separate these lymphocytes into different groups (Fig. 2D). We also considered an alternative, more statistically grounded feature selection method that we recently developed (Differentially Distributed Genes or DDGs), and found that they similarly could not separate the groups of cells (Fig. 2D)[26].

Interestingly, even when we employed *supervised* feature selection approaches, which take advantage of the

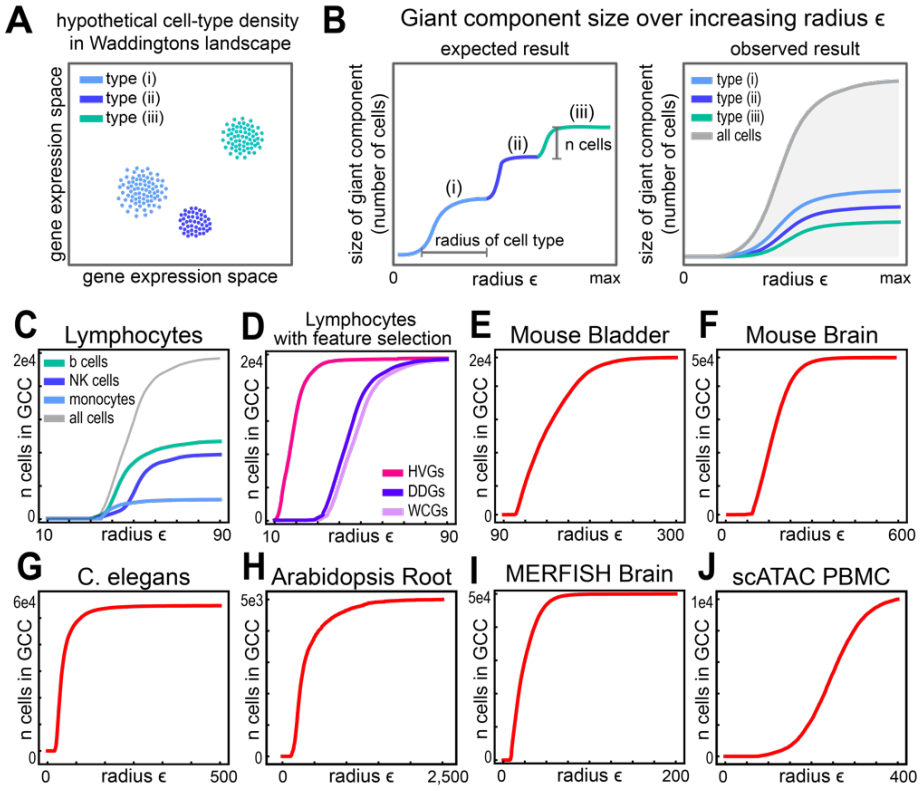

Figure 2. Giant component analysis of ϵ networks for data from various single cell modalities. **A.** Schematic showing the expectation of distinct cell groups from Waddington's Landscape. Each different colored group of cells corresponds to one of the valleys in the landscape, and occupies a distinct region of gene expression space (Fig. 1A). **B.** Schematic showing how the size of the giant component changes as a function of ϵ. On the left, we have the expected behavior we would see if cells were distributed according to the predictions of Waddington's landscape. The line is colored by which cell type causes the increase in the size of the giant component as more cells are included when the radius increases. Waddington's landscape predicts a stepwise pattern as homogenous, relatively dense groups of cells are added to the giant component. In contrast, we see a continuous curve for single-cell data; cells of different types occupy more-or-less the same region of gene expression space, and join the giant component together in a heterogeneous group. **C.** Data from the Lymphocytes on the 10X platform using all the genes as features. The cells were FACS purified before sequencing thus have orthogonal labels with cell types. **D.** Data from the Lymphocytes on the 10X platform using Highly Variable Genes (HVGs), Differentially Distributed Genes (DDGs), and genes determined to be different on average between the cell types using the Wilcoxon rank-sum test (WCGs). All three feature selection methods do not show the predicted step-like behavior. **E-J,** Size of the giant component vs epsilon for additional single cell datasets. Unless listed in the title, all datasets were collected using the 10X platform. All datasets do not show the step-like behavior expected from Waddington's Landscape

fact that we know the cell types *a priori*, we could not separate the cells into distinct attractor-like groups. For instance, we used the Wilcoxon rank-sum test to find genes that are differentially expressed in one of the cell types compared to the other two (WCGs)[27]. Even these genes could not provide the expected step-like behavior (Fig. 2D). Indeed, a more detailed supervised analysis could not find a single gene that could be used to separate all three groups of cells with reasonable cutoffs (Figs. S2.3, S2.4), and sets of genes selected using a supervised approach did not result in well-separated groups (Fig. S2.3G). So, no matter how we analyzed this data, we found that the populations of cells are basically overlapping in gene expression space, with, for instance, many B cells more similar to monocytes and NK cells than they are to other B cells.

We then extended our analysis to a number of other published scRNA-seq data sets. We first analyzed data from the mouse bladder, as an example of a tissue with moderate complexity, and found no evidence of step-like behavior (Fig. 2E)[28]. Similarly, we saw a smooth transition in a much more complex tissue, the mouse brain (Fig. 2F). This is a particularly interesting case; this dataset consists largely of neurons, which are fully differentiated, post-mitotic cells with diverse physiological roles in the brain. Despite this diversity, and the lack of stem cells that might "tie together" different differentiated cell populations, we still found no evidence of cells occupying distinct regions of gene expression space. We then considered scRNA-seq data from the developing *C. elegans* embryo[15] and saw exactly the same behavior (Fig. 2G). This is particularly intriguing because of the deterministic nature of *C. elegans* development, where the distinct cellular fates of various lineages are extremely well-characterized[29]. In this data set, cell-type annotations have been made by applying the standard analysis pipeline, which includes feature selection, normalization, and dimensionality reduction, and then referencing high-quality microscopy data to guide the clustering results. Yet, as with the lymphocyte case, we see that these various cell types are added to the giant component all together, rather than in distinct groups as expected (Fig. S4.4B). This behavior persists even when we just look at transcription factor genes in *C.* elegans, which are thought to drive fate transitions within these lineages and thus should show some evidence of the expected attractor structure (Fig. 1A and S4.2E). We saw the same behavior for every metazoan data set we considered, including data from whole-organism data from *Hydra vulgaris* (Fig. S2.2D). We even saw this lack of separation in scRNA-seq data for the model plant *Arabidopsis thaliana*, suggesting that this is a feature of all complex multicellular organisms (Fig. 2H)[30].

All of the datasets referenced above were obtained using the popular 10X genomics platform, which is known to be highly "noisy" and "sparse" due to the low capture probability of each individual mRNA in the cells being sequenced[31]. It is possible that low capture probabilities destroys the separation between cell types in gene expression space. To test this, we considered scRNA-seq techniques that have much higher capture probabilities. For instance, the BD Rhapsody platform captures 50-75% of mRNAs in each cell, compared to ~5% for 10X, although it can only provide data for ~1,000 genes in the genome[32]. We applied our analysis to BD Rhapsody data for PMBCs, and again found no evidence of distinct cellular groups (Fig. S2.2I,J)[32].

While scRNA-seq methods are extremely popular, alternative approaches have been developed that allow for the quantification of mRNA levels in single cells that do not rely on the "capture-and-sequence" paradigm of scRNA-seq. Perhaps foremost among these is the MERFISH technique, which builds on single-molecule FISH approaches and uses a combinatorial set of labeled probes to identify individual mRNA molecules from a subset of genes in the genome in a microscopy image[18] .This approach does not suffer from the issue of low capture probability like scRNA-seq, and is generally

thought to produce less "noisy" data. We analyzed a recent data set published by the Vizgen corporation, consisting of a mouse brain slice with ~50,000 cells with measurements for around ~600 genes in each cell (Fig. 2I)[33]. We saw exactly the same lack of distinct cell types for this data as we do for the scRNA-seq mouse brain data (Fig. 2J). We also analyzed another published MERFISH dataset of around 15,000 mouse brain cells following a simulated Traumatic Brain Injury (TBI). This data contains measurements for 160 distinct genes per cell, of which 80 genes were specifically selected as marker genes for cell types[34]. We found no evidence of distinct cell type groups in either the entire TBI dataset or a dataset in which we just considered these 80 marker genes (Figs. S4.3B and S4.3C). Together, these findings strongly suggest that our observations are not a consequence of either low capture probabilities or noise. Since these datasets have much lower dimensionality than typical scRNA-seq datasets (~100s of genes rather than ~20,000 as is typical for scRNA-seq), these findings also demonstrate that the lack of distinct cell-type groups is not simply an artefact of the high dimensionality of the data.

It could also be that the observation of a lack of separation between cell groups is an artefact of the choice of the Euclidean distance (i.e. the l_2 norm) to describe the relationship between cells. Although this distance is essentially universally employed in scRNA-seq analysis, it could be that another notion of distance would be more useful in describing differences between cell-type groups[23]. To test this, we used the "Manhattan distance" (the l_1 norm) to construct our ϵ networks, and found the same exact results for both the raw data and feature-selected subsets of the Lymphocyte data and MERFISH brain data (Fig. S2.5). This suggests that the finding described here is not limited to just the Euclidean distance, but persists with other definitions of distance as well.

In addition to feature selection, scRNA-seq data is often subjected to normalization and z-score transformation. Normalization is often achieved by converting the raw counts to "counts per million," which eliminates the variation in the total number of UMI counts (also referred to as "read depth") among cells in the sample[23]. Since some of this variation in read depth could be technical in nature, this step is thought to control for non-biological differences between cells. Interestingly, however, CPM transformations did not result in resolution of the three cell type groups in Lymphocytes (Fig. S2.6A), nor did it generate separate groups of cells in the MERFISH brain data (Fig. S2.7). It is also often argued that different genes in the dataset will have intrinsically different scales of variation: some genes might be expressed at a high level (say, 100s of copies in many cells) and other genes might be expressed at a low level (at most 1 copy found in very few cells). This would lead to different genes having a vastly different impact on the measured distance, with those that vary more contributing more to the distance than those that vary less. One simple way to deal with this problem is to perform a z-score transformation of the data, which ensures that all of the genes vary on the same scale across the dataset[23].As with the CPM case, however, z-score transformation failed to generate separate groups of cells in both the Lymphocyte and MERFISH brain data (Figs. S2.6A and S2.7).

Finally, we also looked at other data on the epigenetic state of cells, including scATAC-seq data, which provides genome-wide information on chromatin accessibility[35]. scATAC-seq data from PBMCs yields essentially identical results to that obtained from scRNA-seq (Fig. 2J), as does PBMC data from the BD Rhapsody platform for protein levels (Fig. S2.2J). Indeed, we could not find any published single-cell dataset on epigenetic state that was consistent with well-separated attractors for individual cell types, as predicted by Waddington's landscape (Fig. 1A and 2B).

The density distribution of cells is inconsistent with attractors in epigenetic space. In addition to being separated from one another, a key feature of the Waddington's picture is that cell types should correspond to attractors in the epigenetic landscape [7,36–38]. A key property of an attractor is the fact that, if a cell is perturbed away from the attractor by noise or an environmental perturbation, the natural dynamics of the gene regulatory network will induce the cell to move back towards the attractor[39,40]. This is usually represented using a "potential well" picture, which explains how cells near an attractor will tend to return to that attractor if they are perturbed while cells in unstable regions (say, at the top of a hill between two attractors) will tend to move away (Fig. 1B). In the presence of noise, the structure of these potential wells suggests that the probability of finding a cell in a certain region of the landscape should vary with the height of the landscape, leading to a large number of cells near the attractors at the bottom of the valleys and few cells on the hills in between them (Fig. 1A). Indeed, this principle has been used to attempt to infer the shape of the landscape from scRNA-seq data[7]. Waddington's landscape thus leads us to expect that most cells will be in regions of similar density, with similar numbers of neighbors in the ϵ networks we constructed.

Inspection of our ϵ networks indicated, however, that this was likely not the case. For one, the transitions in the giant component that we observe are gradual; in other words, the giant component initially begins forming at small distances, but as the size of this cluster grows, it takes larger and larger distances to bring more cells into the giant component (Fig. 2C-J). This suggests that some cells are much farther away from each other than others. Also, we visualized our ϵ networks and found clear differences in density, with some cells having many neighbors and many cells having comparatively few (Fig. 3A). This suggested that the distribution of local densities might give further insight into the structure of the data.

To investigate this phenomenon in greater detail, we calculated the "density distribution" for these single-cell datasets. For any given value of ϵ, we can count how many neighbors each cell has (Fig. 1C), and then plot a histogram of this data. In graph theory, this histogram is known as the "degree distribution" of the network, and is often instructive about the topology of the network itself[41,42]. In our case, this also corresponds naturally to the distribution of local densities, since each ϵ-ball has the same volume (i.e. if one cell has 100 neighbors, then clearly the local density around that cell is much higher than a cell that has only 1 neighbor). If most cells are near the center of an attractor (the bottom of the well in Fig. 1B), they should have more-or-less the same number of neighbors, which should give rise to a roughly binomial (i.e. approximately Gaussian or Gaussian-like) distribution of neighborhood sizes (Fig. 3B).

We first considered the 10X PBMC scRNA-seq data described above and found an approximately power-law distribution of local densities (Fig. 3C). Although the density distribution shown in Fig. 3C is only for one particular value of ϵ, we find the same scaling behavior across a wide range of ϵ values, suggesting that this is not simply an artifact of choosing precisely the right radius for the neighborhoods in question (Fig. S2.2A). In this case, we find a small number of cells that are in extremely dense regions of gene expression space, with thousands of neighbors; most cells, however, are found in very low density regions, with either no neighbors or just one or two (Fig. 3C). The ϵ networks we observe are thus similar to classical "scale-free" networks, though we should note that our analysis here is insufficient to determine if these distributions are truly scale-free or simply similar to a power law[43]. Regardless, this highly heterogeneous density distribution is completely inconsistent with the distributions we would naturally expect to see in the neighborhood of a stable attractor (Fig. 3B).

As mentioned above, the 10X platform is known to be noisy, so it could be that the scale-free density distribution we observe is a reflection of the platform and not the underlying biology. To test this, we considered an artificial dataset in which a set of 92 cDNA standards were introduced to the 10X platform at defined concentrations that span orders of magnitude[13]. In this "ERCC control" data, we do not observe any evidence of scale-free behavior, suggesting that power-law densities are not simply an artifact of the scRNA-seq technique (Fig. 3D). Indeed, we observed scale-free density distributions for all the 10X data sets discussed above, including mouse bladder (Fig. 3E), mouse brain (Fig. 3F), *C. elegans* (Fig. 3G), *A. thaliana, H. vulgaris*, etc. (Figs. S3.1 C,D). We also observe scale-free density distributions in the BD Rhapsody PBMC data (Fig. 3H), further suggesting that this observation is not an artefact of the low capture probability entailed by the 10X platform.

To test whether this observation was specific to the scRNA-seq paradigm, we also analyzed several available MERFISH datasets. Interestingly, the Vizgen data on the mouse brain showed striking power-law behavior across four orders of magnitude in local densities (Fig. 3I). We saw similar behavior for the TBI data (considering either the entire dataset or just marker genes, Figs. S3.1 E,F) and MERFISH data from cultured MCF10A cells (Fig. S3.1G)[44]. As with our analysis of the giant component, these findings strongly suggest that these power-law

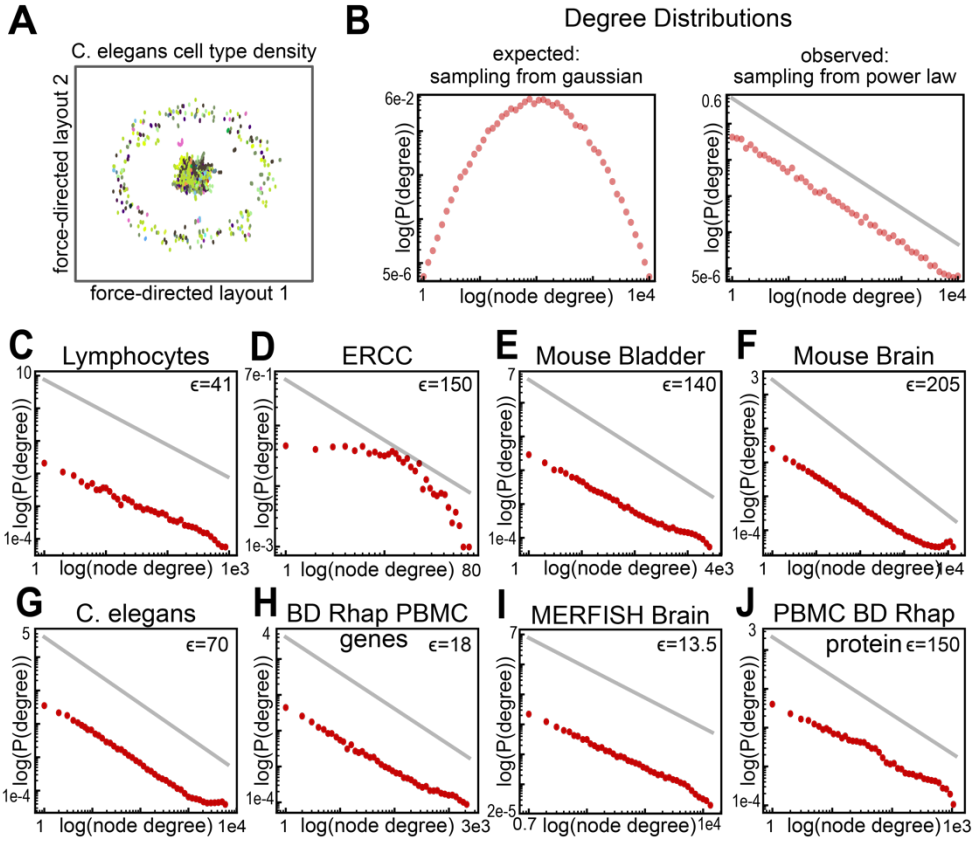

Figure 3. Approximately scale-free density distributions in a variety of single-cell datasets. **A.** Force Directed layout of the epsilon network (ϵ = 70) for data from *C. elegans* embryos collected on 10X platform. For clarity, all "orphan" cells with no neighbors were removed. Visualization of the network reveals heterogeneous density in the giant component with smaller clusters of cells elsewhere in the space. Note that the cells are colored according to the annotated cell type in this data set, and there is clearly no separation of these cell types in the underlying space. **B.** The left panel shows a schematic of the expected degree distribution from attractors as described in the Waddington's Landscape paradigm. The right panel is a schematic to what the degree distribution would look like if the data were sampled from a power law distribution. **C.** Degree Distribution for the epsilon network of FACS-purified Lymphocytes collected on the 10X Platform. **D.** Degree Distribution of the ERCC control data, which does not show scale-free density. **E-I.** Degree Distributions of various single cell gene expression measurements at specified epsilons, all of which show scale-free like behavior. Unless listed, the platform used for collection is 10X. **J.** Degree distribution for an epsilon network built using protein expression data collected on the PBMC Rhapsody Platform. This dataset also shows approximately power-law behavior across almost three orders of magnitude in density. The value of ϵ is shown for each degree distribution as an inset. In all panels showing approximately scale-free behavior, the gray line represents a reference power-law with an exponent of -1.

densities are not a simple consequence of either the high levels of noise nor the inherently high dimensionality of scRNA-seq data. The fact that this observation holds across quite different subsets of genes (chosen in each data set for completely different reasons) also suggests that power-law densities are a general feature of the distribution of cells in gene expression space regardless of the subset of genes considered[15,28,30,32,34,44,45]. Interestingly, we also see this scaling behavior in other epigenetic datasets, including the BD Rhapsody protein data (Fig. 3J) and scATAC-seq transcription factor data (Fig. S31.H), and if we use the l_1 norm instead of the standard l_2 (Fig. S3.3). As with the lack of separation between cell types, approximately power-law density distributions are thus a universal feature of the single-cell data we analyzed.

Popular non-linear transformations cannot reliably separate groups of cells. In the scRNA-seq field, analyses like cell-type clustering are essentially never performed on the raw data, but rather only on data that has been subjected to a set of non-linear transformation and dimensionality reduction steps[23]. A typical pipeline for scRNA-seq data would start with the raw UMI counts, normalize the data so that the total number of counts in each cell is equivalent (e.g. “Counts per Million” or CPM normalization), perform a log transformation (log (CPM +1)), identify a set of HVGs, perform PCA (choosing a number of components based on visual inspection of a scree plot or according to other criteria), and then use non-linear techniques like t-SNE or UMAP to visualize the data in two or three dimensions [23]. Clustering is usually carried out using the popular Louvain or Leiden algorithms after either the PCA or t-SNE/UMAP step[23]. Given that many groups report successful clustering of their data using this broadly similar set of approaches, we considered whether these transformations could reliably separate cells into distinct groups and, if so, which parts of this pipeline seem critical for that separation.

As above, we first considered the case of the FACS-separated PBMC data. Interestingly, we found that most combinations of transformations in this pipeline did not result in distinct cell groups (including CPM normalization on its own, PCA on its own, log transformation, or selection of HVGs, Figs. S4.1). Combining all of these transformations with PCA did, however, generate the step-like behavior we originally expected to see (Fig. 4A). This is encouraging, suggesting that, while there are not distinct attractors in the raw count data, these popular transformations could recover them. Application of this pipeline to other datasets, however, produced either no separation between cell groups or very little step-like behavior. For instance, in the case of the mouse bladder, the transformed data gives essentially the same pattern as the raw data, regardless of the feature selection technique employed (Fig. 4B). The mouse bladder is certainly not constituted of a single cell type, so the fact that the standard pipeline cannot separate cell groups suggests that this pipeline is not a universally reliable approach. This may underly the fact that some studies can successfully recover clear cell-type groups, while others fail to do so[4].

Most of the other data sets we considered fell in between the behavior seen for PBMC lymphocytes and the mouse bladder (Figs. S4.2A-E, S4.3A-F). For instance, for the mouse kidney, we do see some distinct jumps, but not the characteristic step-like behavior we expect for well-separated cell type groups (Fig. 4C). Interestingly, this intermediate behavior is also seen in the Vizgen MERFISH brain data we analyzed (Figs. S4.3C). We also see a similar behavior for the *C. elegans* embryo (Fig. 4D). In most data sets, we lack orthogonal cell-type annotations, so we cannot tell if these individual “jumps” correspond to single cell types, or groups of cell types, joining the giant component all at once (as is the case for the lymphocytes, Fig. 4A). To test this possibility, we focused on the *C.*

elegans data, where we have operationally standard annotations for the cell type of each cell in the data set[15]. At each value of ϵ, the graph will contain multiple components or clusters (Fig. 1B). The moderate step-like behavior we observe here could be due to the fact that cell type groups are coalescing into separate clusters, joining one another, and then adding to the giant component all at once. To test this, we considered the composition of all of the clusters in the graph as a function of ϵ, not just the giant component. The heat map in Fig. 4E considers clusters that are 90% homogeneous for a certain cell type (in other words, they are mostly composed of just one cell type). At a given ϵ we count the number of clusters that have *at least* the percentage of cells of that type in the cluster. So, for the row at 50%, the heat map indicates the number of components that are 90% pure for that cell type, and contain at least 50% of the cells of that type in the data set.

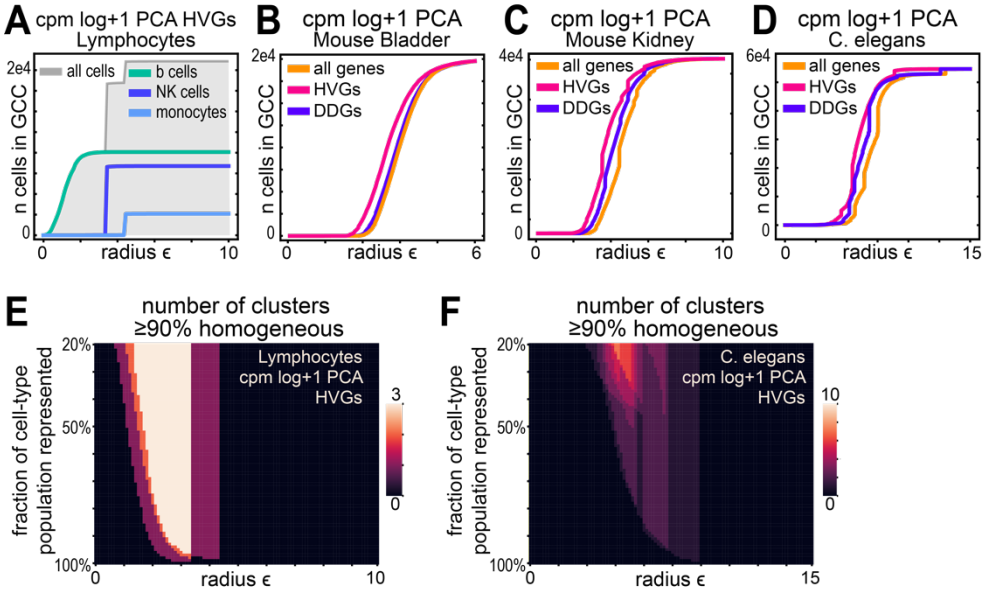

Figure 4. Analysis of common non-linear transformations used in scRNA-seq analysis indicates that they cannot separate individual cell types in general. **A.** Size of giant component vs ϵ for FACS sorted Lympocytes after Counts Per Million (CPM) normalization, log CPM + 1transformation, selection of Highly Variable Genes (HVGs) and dimensionality reduction using PCA. The number of Principle Compnents in this and all other graphs in this figure was chosen using visual inspection to find the "elbow" in the scree plot of explained variance vs. component number, as is commonly done in scRNA-seq. In this case, we actually do see the step-like behavior we expect (Fig. 2B). **B-D.** Size of the giant component vs. ϵ for the indicated datasets. In all cases, the data was subjected to CPM normalization, log CPM + 1 transformation, feature selection and PCA. For Feature selection, we used either all the genes, just the HVGs, or the subset of Differentially Distributed Genes (DDGs). The dramatic stepwise behavior seen in the Lymphocyte data is not seen in any of these datasets, particularly the mouse bladder. The other two cases do display some discrete jumps in the size of the giant component. **E.** To determine if the discrete jumps in the size of the giant component correspond to the addition of large, homogeneous clusters of cells, we generated a heat map to track the number of components in the graph that satisfy a set of criteria. To be counted in these heat maps, a component has to be 90% homogeneous for a given cell type (i.e. 90% of the cells in the component have to belong to just a single cell type). The x-axis of the heat map is the radius ϵ. For each ϵ, we considered all of the components in the graph. For those that were 90% homogeneous, we then calculated what percentage of the cells of that type were included in that component. If a cluster contains more than the indicated percentage on the y-axis, it is counted as satisfying the criteria. The heatmap shows the number of components that satisfy the criteria. For the Lymphocyte data, we see a range of ϵ's where we have 3 components that are more than 90% homogeneous and have collected nearly 100% of the cells of that type. This indicates that the three cell types are distributed in different regions of the space, as expected. **F.** A heat map as in panel E, but for data from *C. elegans* embryos. Cell type annotations were taken from the authors of the study, and unannotated cells were removed from the data prior to analysis. We see almost no cases where clusters that are 90% homogeneous contain a large fraction of the cells of that type. This indicates that the standard transformations do not generate well-separated groups of cells corresponding to distinct cell identities in the *C. elegans* data.

In the transformed *C. elegans* data, we see very few components that are both highly homogeneous and contain most of the cells of a given type (Fig. 4F). For instance, we never see more than 3 clusters that contain more than 50% of the cells of any given annotated cell type, regardless of

the ϵ value we consider. This is in contrast to the PBMC data, where we do indeed see three distinct clusters that mostly consist of a single cell type and that collect most of the cells of that type in the dataset (Fig. 4E). Given that there are 36 annotated cell types in the *C. elegans* data (note that we removed any "unannotated" cell from the data set prior to running this analysis), this suggests that, even though we see discrete "jumps" in the sizes of the giant component for many data sets, this does not correspond to clusters of single cell types first joining together with one another, and then with the giant component. Instead, we see small isolated "islands" of cells of similar type, but those islands are generally closer to cells of a different cell type than to other islands of cells of the same type. Thus, while the transformations applied here do seem to generate small groups of "similar cells," they certainly do not result in the expected attractor structure predicted by Waddington's landscape.

Discussion

For over 80 years, Waddington's landscape has been the dominant picture used to explain canalization in ontogeny[5,7,36,38,46,47]. Since the 1960s, the near-universal interpretation of this landscape has been that individual cell types should correspond to attractors in gene expression space. Indeed, the landscape picture itself was influential in developing the language used in modern dynamical systems theory (for instance, the notion of an attractor's "basin of attraction" was directly inspired by Waddington's ideas)[5,7,36,38,45,47]. Mathematical models of development and differentiation universally cast the process as a set of bifurcations, whereby the number of attractors in gene expression space change during development[5–7,36,38,48]. These concepts have also been deployed extensively in the study of cancer, stem cells and stem cell reprogramming, and other areas [5,36,38,48]. Despite its widespread acceptance, however, it has only been recently that single-cell measurements have allowed us to directly test this picture.

The predictions of Waddington's landscape are clear: cell types should correspond to discrete attractors in gene expression space (Fig. 1A) [5–7,36]. Our analysis demonstrates that available single-cell measurements are completely inconsistent with these predictions. For one, rather than occupying distinct regions of gene expression space, cells of very distinct types and lineages occupy the same region of that space (Fig. 2). This is true not only on a genome-wide scale, but also for subsets of genes taken from "feature selection" approaches that are meant to distinguish sells from one another. Even if we know the cell types in the data *a priori*, we cannot find a set of genes that can reliably distinguish three different types of cells from one another (Fig. S2.3). We found this not just for scRNA-seq data taken from the 10X platform, which is known to generate noisy data, but also for more targeted technologies that provide higher-quality data for an ostensibly biologically informative subset of genes (e.g. the MERFISH data from mouse brains in Fig. 2J and Figs. S4.3B,C). This lack of separation is also present in chromatin accessibility data and data on protein levels (Figs. 2J, S2.2H,J). Moreover, the other clear prediction of Waddington's landscape, that the cells should be found in attractor states no matter how well separated they might be, also does not conform to our findings. Instead of seeing cells clustered around a "typical" gene expression state corresponding to the center of an attractor, we find that cells are extremely heterogenous, leading to an approximately power-law or "fractal" density distribution (Fig. 3).

Since its inception, the analysis of single-cell genomics data, and particularly scRNA-seq, has relied on a series of nonlinear transformations and dimensionality reduction steps that are applied before

any attempt is made to cluster cells into cell types[4,23]. To a careful observer, this fact alone suggests that the data itself is inconsistent with Waddington's landscape; for instance, mathematical models of differentiation and development do not posit dynamics in log(CPM + 1)–HVG–PCA space, but rather directly in the space of mRNA and protein levels, since these are the natural variables of the system[7]. Nonetheless, these transformations have clearly been operationally useful, since no one can deny the sheer volume of papers that have derived meaningful biological insights from this data. Here, we found that application of the "standard pipeline" of scRNA-seq analysis sometimes does separate cells into discrete, homogenous groups (Fig. 4A). In other cases, however, this approach simply does not "work;" in fact, in most cases, these transformations result in groups of physiologically distinct cells that occupy more-or-less the same region of the space (Fig. 4). In scRNA-seq data analysis, and indeed in single-cell genomics more broadly, one finds that nearly every paper employs a different set of transformations, approaches and parameters. Our own analysis indicates that the results of the pipeline depend heavily on the values of those parameters (Fig S5). This suggests that the results of these pipelines are not robust and can be unreliable; in some studies, even with considerable effort, one cannot separate the data into reasonable, distinct groups of cells[11,49]. Our findings imply that the ultimate source of this heterogeneity and difficulty in analysis is the fact that the raw data itself does not display the structure that 60 years of appeal to the Waddington picture suggested it should.

Our findings have wide-ranging implications both for the practical study of single-cell genomics data and the conceptual frameworks underlying our understanding of multicellular biology. Interestingly, the landscape originally conceptualized by Waddington made reference to an abstract epigenetic space quite different from the modern interpretation, in part because the proposal itself was made long before the advent of modern molecular biology[5,46]. One hypothesis that emerges from our work is that Waddington's explanation for canalization is fundamentally correct, but it is just that the epigenetic space is not a space of mRNA levels, protein levels, or chromatin accessibility. In this scenario, cell types correspond to attractors in this (heretofore undescribed) epigenetic landscape, and there is some non-linear projection from those attractors into the spaces that current single-cell technologies allow us to access experimentally. This idea is supported by the fact that analysis pipelines can generate clusters of cells that seem to correspond to our expectations for discrete cell types, suggesting it is possible to "invert" this non-linear map to find the image of these attractors in the available data (Fig. 4A)[4,11]. If this hypothesis holds true, it may be possible to find a more principled approach to inverting the projection that performs more reliably than the set of transformations that form the current standard of practice in the field[4,23]. Understanding the structure of the raw data is a clear first step in any attempt to develop a more principled approach to single-cell analysis.

One other alternative is that there is a separation of cells in gene expression space as Waddington's landscape suggests, but in a subspace with only a few genes that characterize the differences between cells; this would likely include the canonical "marker genes" for different cell types. In our analysis, even using supervised labels for the Lymphocyte data, we could not find such a subspace. That being said, this finding is based on potentially noisy 10X data, and so it could be that, with a more reliable measurement technique, we could find this subspace if we knew the true cell type for each cell in the data set *a priori*. More extensive datasets in which true cell type labels are known and epigenetic state is probed by alternative approaches will clearly be key to exploring this hypothesis further. If this turns out to be the case, the current approach in the field, which uses tools like HVG analysis to find 1000s of genes that can be used to differentiate cell types, would have to be replaced

by a supervised or unsupervised approach to finding this small subset of "cell type determining" genes.

It is possible, however, that canalization in development takes a very different form from that originally envisioned by Waddington. For instance, single-cell measurements of physiological responses of cells ranging from breast tissue to the immune system reveal incredible diversity in those responses[8–10,49]. In other words, the biological responses of individual cells that correspond to our classical notion of a "cell type" is itself quite heterogeneous, suggesting that categorizing cells into discrete groups in the first place may mask critical aspects of their physiology. Emerging functional data at the single-cell level thus suggest that the "final step" in development may not be a discrete set of cell types, as Waddington's landscape posits (Fig. 1A), but rather a more continuous spectrum of states and phenotypes that is not well-approximated by the attractor picture. Regardless of whether or not there are cell-type attractors to be found in some as of yet uncharacterized epigenetic space, it is clear that careful analysis of available data, and a willingness to test even well-established paradigms against that data, is critical to the future of single-cell biology.

Methods

Datasets The vast majority of data we analyzed was taken directly from freely-available repositories on the internet. Table S1 in the Supplmentary Information summarizes each of these datasets, including the name we have given to each data set, number of cells, number of features measured, and a link to the corresponding resource where we obtained the data. CSV files for all of the datasets we used, specific to each transformation and analysis, are also provided for each dataset as additional supplementary material.

The MERFISH data that we analyzed is the only data that is not publicly accessible in an appropriate format online. The MERFISH data for the mouse Traumatic Brain Injury dataset was obtained using the protocols and analysis described in [34]. The count X cell matrix for this dataset was kindly provided by Zach Hemminger and Roy Wollman, and is available as supplementary information. Similarly, the Vizgen corporation has made image data from their MERFISH experiments on mouse brain slices freely available on the web[33]. Data for one of these slices was obtained through image analysis as described in, and again kindly provided by Zach Hemminger and Roy Wollman. As with the TBI data, this data is also provided as additional supplementary information.

ϵ network construction, analysis and visualization For every dataset, the raw data consists of a matrix where the rows correspond to each individual cell and the columns to the features measured in the single-cell experiment (note that the data is sometimes represented with the columns as cells and the rows as features). For scRNA-seq data, these features are all the genes in the genome, and the entries in the matrix are the number of UMI counts for that gene in that cell. All of the other data considered here (MERFISH, BD Rhapsody mRNA and protein, etc.) ultimately consists of a similar matrix, just with either fewer genes measured (in the case of MERFISH, for example) or a different type of measurement for each entry (for instance, protein levels rather than UMI counts in the BD Rhapsody protein data).

For any given dataset, we can easily calculate the distance between any two cells using their corresponding feature vectors. Here, we used the simple Euclidean distance (i.e. the l_2 norm) to define the

distance, since this is definition of distance used in the vast majority of scRNA-seq and single-cell genomics studies . This is also the natural notion of distance used in the analysis of dynamical systems (see, e.g., the Hartman-Grobman theorem on stability analysis through linearization about an equilibrium point, and a host of other definitions and theorems)[39,40]. If we calculate the distance between every pair of cells, this gives us a symmetric distance matrix with a number of elements equal to the number of cells in the dataset squared.

Once we calculate this distance matrix, we can use that matrix to generate an epsilon network. To do so, we first define the cutoff ϵ to be some value. Then, for each cell i in the dataset, we go through the row in the distance matrix and consider every other cell j. If the distance between these cells is less than the cutoff ($d_{i,j} < \epsilon$), then we add cell j to the list of cells that are connected (or adjacent) to cell i. Doing this for every cell in the dataset generates a standard adjacency list representation of the ϵ network.

We analyzed the resulting ϵ network using a standard set of algorithms on graphs. For instance, we used a Depth-First-Search (DFS) to determine all of the components in each of our ϵ networks. The largest such component is the giant component, and the number of cells in that largest component is the size of the giant component. Similarly, the "degree" of any cell i for a given value of ϵ is just the size of the adjacency list for that node in the graph [42]. Degree distributions were calculated as a histogram across these individual degrees. Note that the histograms in Fig. 3 and in the Supplementary Information use a standard logarithmic binning approach for approximately scale-free distributions.

Distance matrices were calculated either using the scanpy package in Python [50]or using custom-built C++ software (particularly for larger datasets). Analysis of the giant component size and composition as a function of ϵ was performed using custom-built C++ software. Degree distributions and force-directed layouts were calculated using the NetworkX package in Python, and all plots were generated using the matplotlib package in Python[51,52]. All software used in this work is available from the authors upon request.

Acknowledgements

The authors thank Alexander Hoffmann, Tom Kolokotrones, Jukka Keranen, Pavak Shah, Nina Gilshteyn, and other members of the Deeds lab for many helpful discussions and comments. This work was supported by an NIH IRACDA postdoctoral fellowship 2K12GM106996-06 to BS and NIH R01-GM143378 to EJD.

References:

1. Tang, X., Huang, Y., Lei, J., Luo, H. & Zhu, X. The single-cell sequencing: new developments and medical applications. *Cell Biosci.* **9**, 53 (2019).

2. Nguyen, Q. H., Pervolarakis, N., Nee, K. & Kessenbrock, K. Experimental Considerations for Single-Cell RNA Sequencing Approaches. *Front. Cell Dev. Biol.* **6**, (2018).

3. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).

4. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).

5. Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: A Landscape Takes Shape. *Cell* **128**, 635–638 (2007).

6. Matsushita, Y. & Kaneko, K. Homeorhesis in Waddington's landscape by epigenetic feedback regulation. *Phys. Rev. Res.* **2**, 023083 (2020).

7. Wang, J., Zhang, K., Xu, L. & Wang, E. Quantifying the Waddington landscape and biological paths for development and differentiation. *Proc. Natl. Acad. Sci.* **108**, 8257–8262 (2011).

8. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).

9. SoRelle, E. D. *et al.* Single-cell RNA-seq reveals transcriptomic heterogeneity mediated by host–pathogen dynamics in lymphoblastoid cell lines. *eLife* **10**, e62586 (2021).

10. Muhl, L. *et al.* Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. *Nat. Commun.* **11**, 3953 (2020).

11. Chari, T., Banerjee, J. & Pachter, L. The Specious Art of Single-Cell Genomics. 2021.08.25.457696 (2021) doi:10.1101/2021.08.25.457696.

12. Cooley, S. M., Hamilton, T., Deeds, E. J. & Ray, J. C. J. A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. *bioRxiv* 689851 (2019) doi:10.1101/689851.

13. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

14. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).

15. Packer, J. S. *et al.* A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).

16. Siebert, S. *et al.* Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. *Science* (2019) doi:10.1126/science.aav9314.

17. Kim, J. K., Kolodziejczyk, A. A., Ilicic, T., Teichmann, S. A. & Marioni, J. C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6**, 8687 (2015).

18. Moffitt, J. R. & Zhuang, X. Chapter One - RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). in *Methods in Enzymology* (eds. Filonov, G. S. & Jaffrey, S. R.) vol. 572 1–49 (Academic Press, 2016).

19. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).

20. *Network Science by Albert-László Barabási*.

21. Tangherloni, A., Ricciuti, F., Besozzi, D., Liò, P. & Cvejic, A. Analysis of single-cell RNA sequencing data based on autoencoders. 727867 (2021) doi:10.1101/727867.

22. Wagner, F., Barkley, D. & Yanai, I. *Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis*. 655365 https://www.biorxiv.org/content/10.1101/655365v2 (2019) doi:10.1101/655365.

23. Lueken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

24. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).

25. Liao, X., Makris, M. & Luo, X. M. Fluorescence-activated Cell Sorting for Purification of Plasmacytoid Dendritic Cells from the Mouse Bone Marrow. *JoVE J. Vis. Exp.* e54641 (2016) doi:10.3791/54641.

26. Sparta, B., Hamilton, T., Aragones, S. D. & Deeds, E. J. Binomial models uncover biological variation during feature selection of droplet-based single-cell RNA sequencing. 2021.07.11.451989 (2021) doi:10.1101/2021.07.11.451989.

27. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **18**, 50–60 (1947).

28. Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222-3241.e26 (2021).

29. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).

30. Shulse, C. N. *et al.* High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types. *Cell Rep.* **27**, 2241-2247.e4 (2019).

31. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics Proteomics Bioinformatics* **19**, 253–266 (2021).

32. Mair, F. *et al.* A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level. *Cell Rep.* **31**, 107499 (2020).

33. Vizgen MERFISH Mouse Receptor Map. https://info.vizgen.com/mouse-brain-map.

34. Littman, R. *et al.* Joint cell segmentation and cell type annotation for spatial transcriptomics. *Mol. Syst. Biol.* **17**, e10108 (2021).

35. 10k Peripheral blood mononuclear cells (PBMCs) from a healthy donor (Next GEM v1.1). *10x Genomics* https://www.10xgenomics.com/resources/datasets/10-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-next-gem-v-1-1-1-1-standard-2-0-0.

36. Ferrell, J. E. Bistability, Bifurcations, and Waddington's Epigenetic Landscape. *Curr. Biol.* **22**, R458–R466 (2012).

37. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).

38. Shakiba, N. *et al.* How can Waddington-like landscapes facilitate insights beyond developmental biology? *Cell Syst.* **13**, 4–9 (2022).

39. Strogatz, S. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. (1994).

40. Garfinkel, A., Shevtsov, J. & Guo, Y. *Modeling Life*. (Springer International Publishing, 2017). doi:10.1007/978-3-319-59731-7.

41. Jeong, H., Néda, Z. & Barabási, A. L. Measuring preferential attachment in evolving networks. *Europhys. Lett. EPL* **61**, 567–572 (2003).

42. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).

43. Zhou, B., Meng, X. & Stanley, H. E. Power-law distribution of degree–degree distance: A better representation of the scale-free property of complex networks. *Proc. Natl. Acad. Sci.* **117**, 14812–14818 (2020).

44. Foreman, R. & Wollman, R. Mammalian gene expression variability is explained by underlying cell state. *Mol. Syst. Biol.* **16**, e9146 (2020).

45. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

46. Baedke, J. The epigenetic landscape in the course of time: Conrad Hal Waddington's methodological impact on the life sciences. *Stud. Hist. Philos. Biol. Biomed. Sci.* **44**, 756–773 (2013).

47. Nicoglou, A. Waddington's epigenetics or the pictorial meetings of development and genetics. *Hist. Philos. Life Sci.* **40**, 61 (2018).

48. Kaity, B., Sarkar, R., Chakrabarti, B. & Mitra, M. K. Reprogramming, oscillations and transdifferentiation in epigenetic landscapes. *Sci. Rep.* **8**, 7358 (2018).

49. Goldman, S. L. *et al.* The Impact of Heterogeneity on Single-Cell Sequencing. *Front. Genet.* **10**, (2019).

50. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

51. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

52. Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function using NetworkX. http://conference.scipy.org/proceedings/SciPy2008/paper_2/.