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], Gunalan Natesan[1] and Eric J. Deeds[1,2,3]

[1]Institute for Quantitative and Computational Biosciences, University of California, Los Angeles,
[2]Department for Integrative 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

The prevailing interpretation of Waddington's landscape is that attractors in gene expression space produce and stabilize distinct cell types. This notion is often applied in single-cell omics data, where groups of cells are clustered for analysis. Here 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. We found that cell types exist in the same regions of epigenetic space, with highly heterogeneous density distributions that are inconsistent with expected densities near an attractor. The lack of attractor structure could not be explained by technical noise, scale variance among genes, nor the subset of genes that were used; nor could it be rescued by any standard set of transformations. These findings pose a challenge for the robust analysis of single-cell data and open the possibility for alternative explanations of canalization during development.

Introduction

Genome-wide single cell omics 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 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 perturbations. However, since scRNA-seq tools are inherently destructive, *a priori* knowledge of the cell type of each cell 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 "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). Gene regulatory networks underly each cell fate, and their 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. Waddington's landscape is thus a generative model for cell-types, and through the molecular mechanism of attractor structures, it 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[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].

In this work, we systematically characterized the relationships between cells in epigenetic space, using data provided by various single-cell technologies. We began our search for groups of transcriptionally similar cells in the original, high-dimensional, genome-wide data, before considering different common transformations. To do this, we developed and validated an approach that we term "ϵ networks," where we consider the 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. 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 specific subsets of genes that should be useful in separating biologically distinct groups of cells[13]. Further, we found no evidence to support the hypothesis that the absence of attractor structures can be explained by technical noise. Regardless of the measurement technology, normalization technique applied, dimensionality reduction method used, or combination thereof, we were unable to identify clear cell-type groups in more than ten large single-cell RNA-seq datasets, including the 10x version 3 chemistry with improved mRNA capture. The absence of these groups remained consistent even in completely different experimental modalities, including MERFISH, a microscopy-based technique that measures transcript abundance using a combinatorial set of fluorescently-labeled probes, measurements of protein levels, and measurements of chromatin accessibility (scATAC-seq)[14,15].

Our ϵ network approach also allowed us to characterize how the cells themselves are distributed in gene expression space. Waddington's landscape predicts 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[16]. As with our observations above, we found this "fractal-like 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 data on chromatin accessibility.

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 (Fig. S1)[17]. In some cases, analysis of the data does not yield the expected separation of cell-type marker genes, even after considerable effort[11,18]. 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 heterogeneous 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. More importantly, however, our analysis of available single-cell data on epigenetic state did not find any evidence to support the hypothesis that cell types are maintained by dynamical attractors in gene expression space. The extensive heterogeneity we observe presents major challenges both for our empirical and theoretical approaches to single-cell biology.

Results

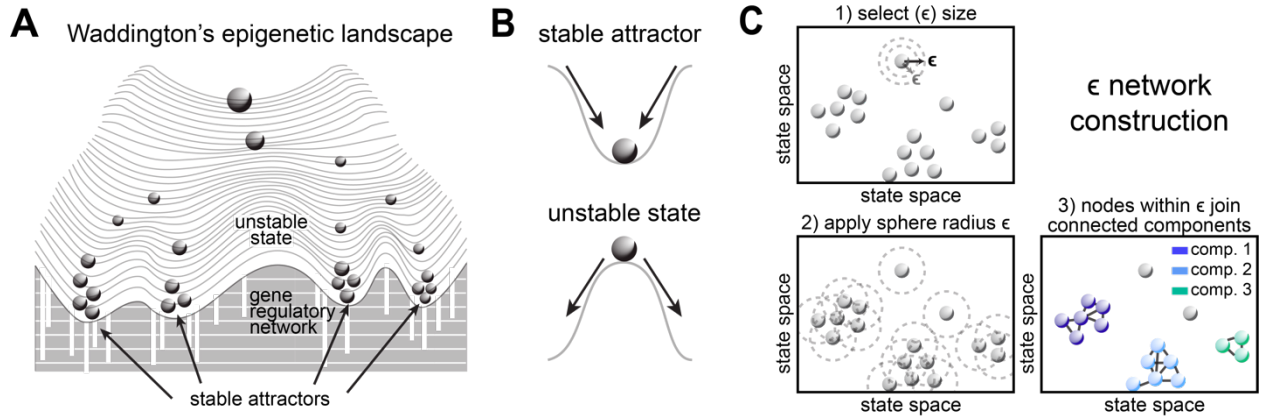

Figure 1. Waddington's Epigenetic Landscape draws predictions from dynamical systems theory. **A.** Schematic of Waddington's landscape, where the epigenetic state of developing cells is *canalized* by interactions in gene regulatory networks. Cell types are thought to represent different "basins of attraction" in this landscape, and cells of the same type should have similar gene expression patterns. **B.** Schematic of the states that construct the hills and valleys of Waddington's Landscape. The valleys represent stable attractors, and function to buffer a cell from perturbations. The hills correspond to unstable states, where any perturbation results in increasing divergence from the original state. **C.** Schematic of ϵ network construction. First, we consider a hypersphere of radius ϵ around any given cell in the dataset (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 (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.

A lack of distinct cell groups in single-cell data. 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). We developed a straightforward approach, which we term "ϵ networks", which iterates through a range of distance thresholds and measures how cells are distributed relative to one another in epigenetic space. For any given dataset, we can calculate the distance between any two cells using their corresponding feature vectors (i.e.

gene expression values). To start, 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 notion of "divergence" in the definition of the Lyapunov exponent, the standard definition of an attractor, and a host of other definitions and theorems)[19,20].

To construct an "ϵ network", we take an n-dimensional hypersphere of radius ϵ, where each dimension represents a gene, and we apply this hypersphere to each cell in the data (Fig. 1C, panels 1 and 2). If two cells are closer to each other than this radius ϵ (i.e., they lie within each other's hyperspheres), then we connect those cells in the network (Fig. 1C, panel 3 and detailed in 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 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). 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. S2.1iii). To track this behavior, we measured the size of the largest cluster in the graph (called the "giant component") 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, other groups of cells are expected to join all at once. This should give rise to a characteristic step-like behavior in the giant component (Fig. 2B).

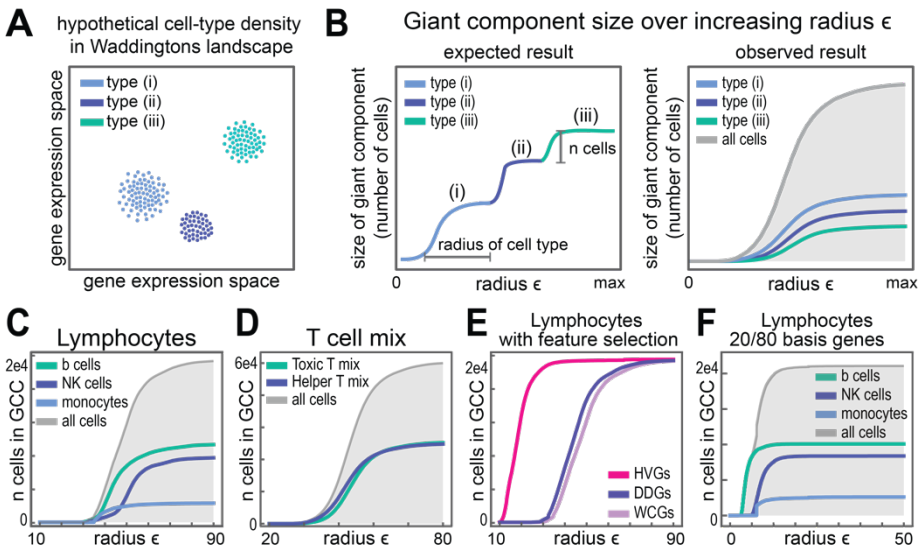

Figure 2. ϵ network analysis of FACS-separated lymphocytes. **A)** Schematic showing the expected distribution of cells in gene expression space, where each group exists in a distinct valley of Waddington's Landscape and thus occupies a distinct region. **B)** Schematic showing how the size of the giant component changes as a function of ϵ. The expected behavior, illustrated on the left, demonstrates homogenous, dense groups of cells being sequentially added to the giant component in a step-wise fashion. On the right is an illustration of the observed behavior, where cells of different types occupy more-or-less the same region of gene expression space and join the giant component together. **C-F)** Size of the giant component as a function of ϵ for Lymphocytes sorted by FACS and then sequenced on the 10x genomics scRNA-seq platform. Giant component analysis using the raw UMI counts of all detected genes for **C)** cell types from distinct groups and **D)** cell types from different branches of T cells. In **E)** & **F)** b cells, NK cells, and monocytes, were analyzed using a subset of feature selected genes. **E)** uses either 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). **F)** uses genes determined using a supervised, distance-based approach (20/80 basis genes).

To test this idea, we first considered a classic scRNA-seq data set from human Peripheral Blood Monocytes (PBMCs)[17]. In this data, cells were sorted by FACS using cell surface markers to separate them into distinct cell-type groups prior to sequencing. We thus know the "true" cell-type label for each cell in the dataset[21], which allows us to not only monitor the size of the giant component, but also its composition. We first analyzed a subset of these data comprised of three cell types: B cells, Natural Killer (NK) cells, and monocytes. These terminally-differentiated cell types represent distinct developmental lineages of PBMCs, and should thus be easily separable in gene expression space. Yet, as can be seen from Fig. 2C, the size of the giant component (gray line) over increasing radius ϵ's shows no step-like behavior whatsoever. Instead, the transition in the giant component is smooth, as cells of the different types join the giant component more-or-less at the same time (Fig. 2C). We see the same behavior when evaluating various subsets of T cells from the PBMC data set into the data, suggesting that all the T cells purified in this experiment occupy the same region of gene expression space (Fig. 2D).

We then investigated whether our ϵ network approach is sensitive enough to identify separation between cell types when there is noise in the data or when there are only a few genes that differ between cell-type groups. To do this, we tested several simple gaussian mixture models (Fig. S2.2). We found that in all cases, the epsilon network approach identified separation in gaussian clouds, and yielded the expected staircase behavior (Fig S2.2 B,E,H), suggesting that our approach can find separate cell type groups in the presence of significant noise even when only a few genes differ between them.

In scRNA-seq analysis, cell-type clustering is typically performed on only a subset of "feature selected" genes that are expected to vary significantly between cells of different types. We returned to the PBMC data to test whether including ubiquitously expressed genes in our original analysis had caused overlap between cells of different types. We first considered the most popular approach of finding "Highly Variable Genes" (HVGs) whose variance is higher than one would expect given the average expression level[22]. Interestingly, HVGs did not separate these lymphocytes into different groups (Fig. 2E). 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. 2E)[23].

Interestingly, even when we employed *supervised* feature selection approaches, which take advantage of the 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)[24]. Even these genes could not provide the expected step-like behavior (Fig. 2E). Indeed, a more detailed supervised analysis, which used a per-gene distance criteria, could not find a single gene that could be used to separate all three groups of cells with reasonable cutoffs (Figs. 2F, S2.3), and sets of genes selected using this approach did not result in well-separated groups (Fig. 2F). Regardless of which set of genes we used, we found that the populations of cells remain 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 (Fig. S2.5).

The lymphocyte data used in our initial analysis was 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[25]. It is possible that low capture probabilities destroy 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[26]. We applied our analysis to BD Rhapsody data for PMBCs, and again found no evidence of distinct cellular groups (Fig. 3A)[26]. We also considered data from Parse Biosciences, a plate-based platform that uses combinatorial barcoding, as well as data from the 10x version 3 chemistry with improved sensitivity[27,28]. Even when we applied feature selection methods to these data, we were unable to find evidence of distinct cellular groups (Fig. 3B,C).

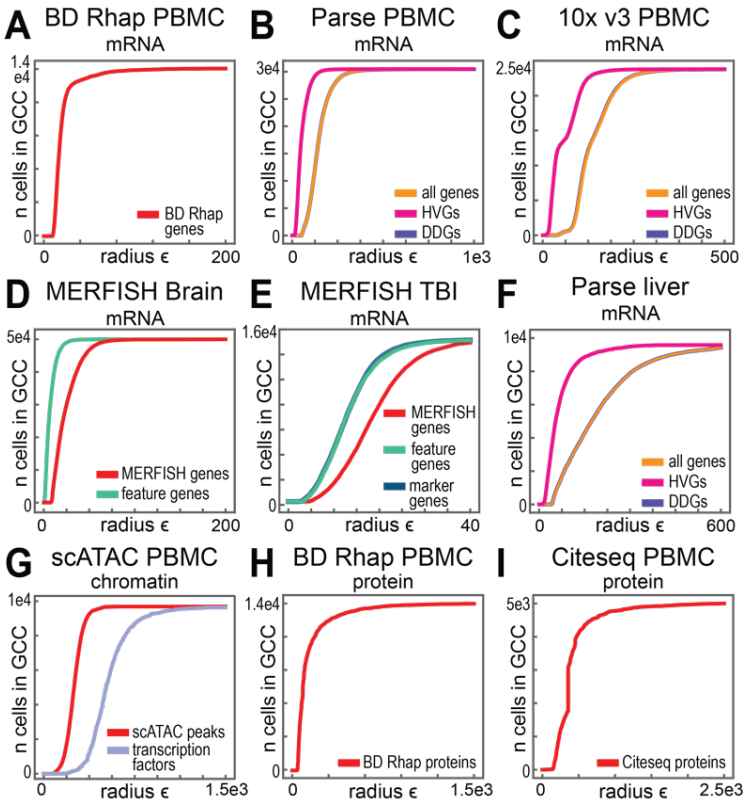

Figure 3. ϵ network analysis of data from different single-cell measurement platforms. **A-I)** Size of the giant component as a function of ϵ for **A)** targeted immune panel genes sequenced in Peripheral Blood Mononuclear Cells (PBMC) on the BD Rhapsody platform. **B)** all genes, HVGs, or DDGs in PBMCs sequenced using the Parse Biosciences platform. **C)** all genes, HVGs, or DDGs in PBMCs sequenced on the 10x genomics platform with version 3 chemistry. **D)** MERFISH or 'feature genes' detected in Brain tissue using the MERFISH platform. **E)** MERFISH, 'feature genes', or marker genes detected in TBI Brain tissue using the MERFISH platform. **F)** all genes, HVGs, or DDGs in mouse liver sequenced using the Parse Biosciences platform. **G)** all peaks, or transcription factor peaks of chromatin accessibility detected in PBMCs sequenced using scATAC. **H)** proteins detected in PBMCs using the BD Rhapsody platform. **I)** proteins detected in PBMCs using the Citeseq platform.

Alternative approaches have also been developed that allow for the quantification of mRNA levels in single cells that do not rely on the "capture-and-sequence" paradigm. For instance, the MERFISH technique uses a combinatorial set of labeled probes to identify individual mRNA molecules from a subset of genes in the genome in a microscopy image[14]. We analyzed a 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. 3D)[29]. 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[30]. 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 (Fig. 3E). To further test whether appropriate subsets of "cell-type determining genes" could separate groups of cells, we identified genes in the two MERFISH datasets with

multimodal distributions (Fig. S3.1A), reasoning that these genes might comprise different sets of genes with distinct expression levels in different subsets of cells. However, using these sets of 23 "feature genes", we again found no evidence of distinct cell type groups (Fig. 3D,E). 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.

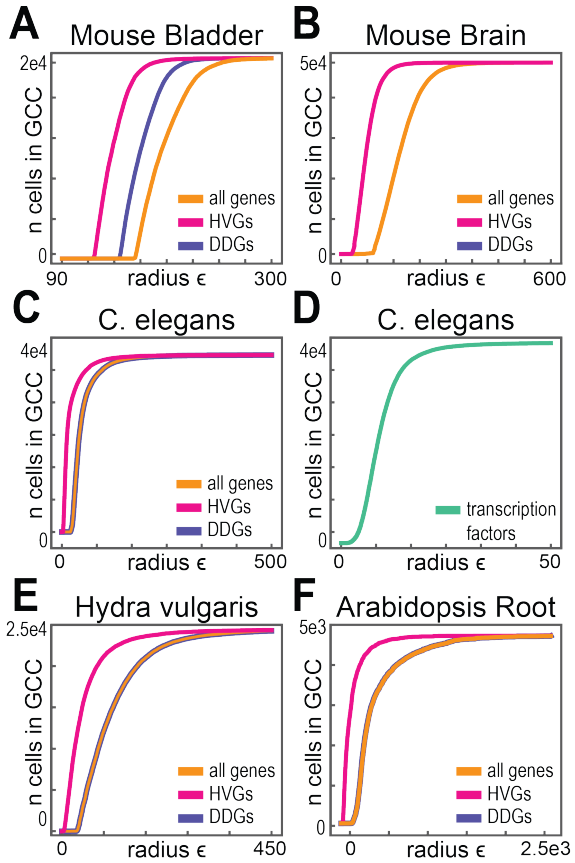

Figure 4. ϵ network analysis of scRNA-seq data from different tissues and organisms. **A-F)** Size of the giant component as a function of ϵ for all genes, HVGs, DDGs, or transcription factors detected using the 10x genomics scRNA-seq platform in **A)** mouse bladder **B)** mouse brain, **C,D)** *C. elegans* embryo **E)** *Hydra vulgaris,* and **F)** *Arabidopsis* root.

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[31]. scATAC-seq data from PBMCs yields essentially identical results to that obtained from scRNA-seq (Fig. 3G). Because scATAC-seq assays DNA accessibility, rather than mRNA levels, we inspected whether evaluating just the subset of genes that encode transcription factors would recover distinct groups of cells. Once again, considering only transcription factors did not produce distinct regions of cells (Fig. 3G). We further considered whether single-cell data on protein levels could identify these distinct groups. Looking at PBMC data on protein levels from the BD-rhapsody or the cite-seq platform did not recover these distinct groups (Fig. 3H,I). Thus, regardless of the single-cell platform, epigenetic modality measured, or subset of genes considered, 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).

We next extended our analysis to characterize scRNA-seq data from samples with distinct developmental and phylogenetic origins. In each case we constructed ϵ networks for the original, high dimensional data and different subsets of genes. 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 in any of the feature selected data (Fig. 4A)[32]. Similarly, regardless of how we selected the feature space, we saw a smooth transition in the much more complex tissue, the mouse brain (Fig. 4B). 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. We then considered scRNA-seq data from the developing *C. elegans* embryo[33] and saw exactly the same behavior (Fig. 4C). 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[34]. 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 4D). We saw the same behavior for every data set we considered, including whole-organism data from *Hydra vulgaris* (Fig. 4E) and even data for the model plant *Arabidopsis thaliana*, suggesting that this is a feature of all complex multicellular organisms (Fig. 4F)[35].

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[18]. A typical pipeline for scRNA-seq data would start with the raw UMI counts, perform "counts per million" (CPM) or an equivalent normalization, perform a log transformation (e.g. log (CPM +1)), identify a set of HVGs, perform PCA, and then use non-linear techniques like t-SNE or UMAP to visualize the data in two or three dimensions [18]. Clustering is usually carried out using the Louvain or Leiden community-detection algorithms after either the PCA or t-SNE/UMAP step[18]. 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 Lymphocyte 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. S5.1). Combining all of these transformations did, however, generate the step-like behavior we originally expected to see, regardless of the set of genes the pipeline of transformations was applied to (Fig. 5A). Interestingly, we observed that the CPM transformation applied to the set of HVGs (but not other gene sets) was also able to achieve some separation (Fig S5.1C).

This is encouraging, suggesting that, while there are not distinct attractors in the raw count data, either applying this popular pipeline of transformations, or just HVG selection and CPM normalization, might be able to recover the attractor structure. Application of these transformations to other datasets, however, produced either no separation between cell groups or very little step-like behavior (Fig. S5.2, S5.3, S5.4). For instance, when applying the standard analysis pipeline to the mouse bladder, the transformed data gives essentially the same pattern as the raw data, regardless of the feature selection technique employed (Fig. 5B and S5.2A). This suggests that the standard pipeline is not a universally reliable approach.

Most of the other data sets we considered fell in between the behavior seen for PBMC lymphocytes and the mouse bladder (Figs. S5.2, S5.3, S5.4). For instance, for the *C. elegans* embryo, we do see some distinct jumps, but not the characteristic step-like behavior we expect for well-separated cell type groups (Fig. 5D and S5.2C). 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. 5A). 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[33]. At each value of ϵ, the graph will contain multiple components or clusters (Fig. 1B).

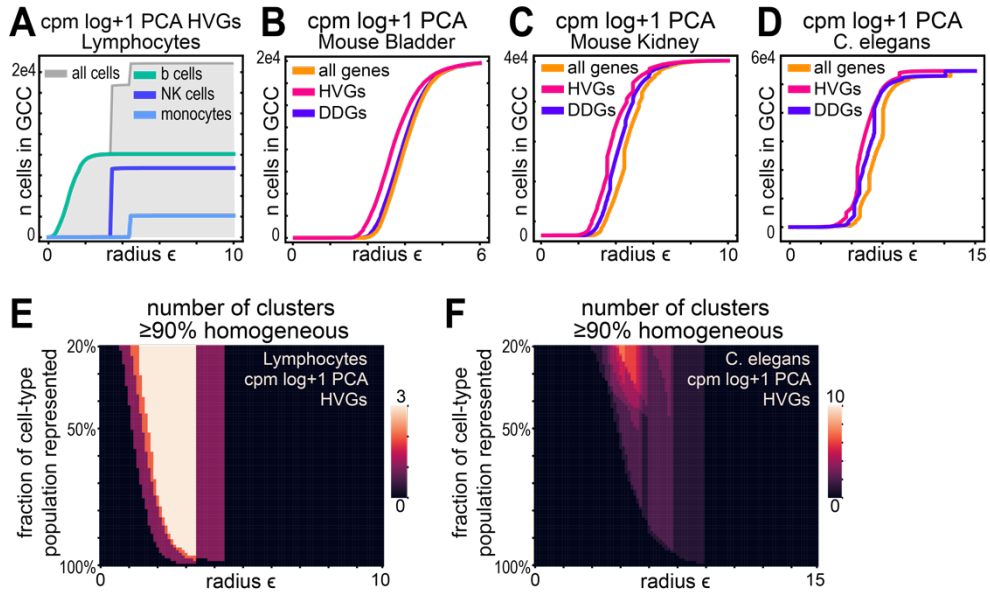

Figure 5. Common non-linear transformations do not recover cell-type groups. **A.** Size of giant component vs ϵ for FACS sorted Lympocytes after Counts Per Million (CPM) normalization, log CPM + 1transformation, selection of Highly Variable Genes (HVGs) and dimensionality reduction using PCA. **B-D.** Size of the giant component vs. ϵ for the indicated datasets, with the same set of transformations in (A), applied to all genes, HVGs, or Differentially Distributed Genes (DDGs). **E,F** Heatmap characterizing the size and composition of components in the ϵ networks for E) the FACS sorted lymphocytes and F) the *C. elegans* data. The x-axis indicates increasing radius ϵ, the y-axis represents the fraction of particular cell types represented in each cluster, relative to the total number of that particular cell type in the data. The heatmap is colored by the number of components meeting a 90% homogeneity criteria. For the Lymphocyte data, we see a range of ϵ's where 3 components are more than 90% homogeneous and have collected nearly 100% of the cells of that type. In *C. elegans* we see very few cases where clusters that are 90% homogeneous contain a large fraction of the cells of that type.

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 quantified the composition of all of the clusters in the graph as a function of ϵ, not just the giant component.

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. 5F, S5.5 L,M,N, and S5.6 A,B). For instance, we never see more than 5 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. 5E, S5.5 E,F,G,H). Given that there are 36 annotated cell types in the *C. elegans* data, 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 (Fig S5.7). 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.

Interestingly, in a few cases, such as the 10x v3 PBMC data, we observe that CPM transformation of HVGs alone (i.e. without log transformation or PCA) produces distinct jumps in the ϵ network analysis (Fig. S5.4, and S5.8). To distinguish between the possibility that the two groups of cells clusters correspond to distinct PBMC lineages, or the possibility that CPM-HVG transformation artificially introduces separation in the data[24], we took advantage of the well-established marker genes in this tissue and evaluated marker gene expression across the two groups. Analysis of gene expression in the two

clusters suggested that, in the v3 PBMC data, the two groups produced by CPM-HVG transformation appear to correspond more to "read-depth" than any pattern of marker gene expression (Fig S5.8). Indeed, CPM transformation of a selection of 34 random genes also produced two distinct clusters in this data (Fig S5.8D), raising the possibility that the generation of apparent clusters can occur spuriously when non-linear transforms are applied.

Since the standard pipeline or CPM-HVG transformation does not consistently separate groups of cells, we evaluated whether other common transformations of the data could produce the expected separation. It is also often argued that due to differences in relative expression levels, different genes in the dataset will have intrinsically different scales of variation. This would lead to different genes having a vastly different impact on the distance. One simple way to deal with this problem is to perform a z-score transformation of the data[18]. However, z-score transformation failed to generate separate groups of cells for any of the feature-selected single-cell data we tested (Fig. S5.9). Further, neither the combination of CPM then z-score normalization, or vice-versa, succeeded in separating the lymphocyte groups into distinct groups, regardless of which set of genes we tested (Fig. S5.10).

In addition to accounting for "sampling differences", cell-size, and gene weights, we also tested a number of alternative feature spaces and clustering techniques. We tested cell-cycle regression methods[36], different metrics of distance, as well as non-single linkage clustering algorithms (Figs. S5.11, S5.12, S5.13, S5.14, S5.15, and S5.16). However, across all of the single-cell data we tested, none of these various scaling, measurement, or clustering techniques produced the expected separation of cells into distinct groups.

The density distribution of cells is inconsistent with attractors in epigenetic space. In addition to being separated from one another, Waddington's picture predicts that cell types should correspond to attractors in the epigenetic landscape [7,37–39]. 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[19,20]. 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 and S6.1). 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.

To investigate this phenomenon we calculated the "density distribution" for various 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[40,41]. 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. 6B).

We first considered the 10X PBMC scRNA-seq data described above and found an approximately power-law distribution of local densities (Fig. 6C). Although the density distribution shown in Fig. 6C 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. S6.2A). 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. 6C). 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[42]. 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. 6B).

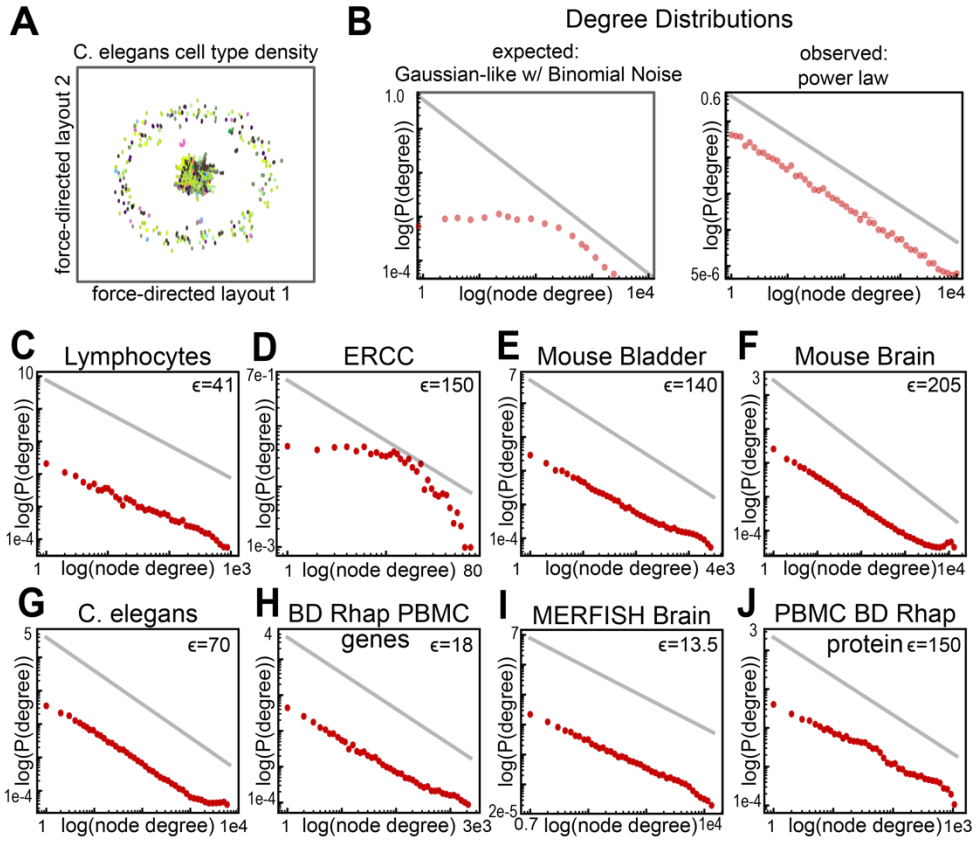

Figure 6. Density distributions of single-cell datasets indicate the absence of attractor structures. **A.** Force Directed layout of the ϵ-network (ϵ = 70) for data from *C. elegans* embryos collected on 10X platform (all "orphan" cells with no neighbors were removed for visual clarity). Cells are colored according to the annotated cell type. **B.** A schematic of the degree distribution expected to be produced by attractors (left) vs the degree distribution expected when sampling from a power law distribution (right). **C-J)** Degree Distribution for the ϵ-network of various datasets at the indicated ϵ: C) FACS- separated Lymphocytes D) ERCC control data E) Mouse Bladder, F) Mouse Brain, G) *C. elegans,* H) BD Rhapsody PBMC genes, I) MERFHISH Brain, J) BD Rhapsody PBMC protein. The gray line represents a power-law with an exponent of -1 for reference.

As mentioned above, the 10X platform is known to be noisy, so it could be that the heterogeneous 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[17]. 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. 6D). Interestingly, we observed highly heterogeneous density distributions for all the 10X data sets discussed above, including

mouse bladder (Fig. 6E), mouse brain (Fig. 6F), *C. elegans* (Fig. 6G), *H. vulgaris*, *A. thaliana,* etc. (Figs. S6.3 C,D). We also observe scale-free density distributions in the BD Rhapsody PBMC data (Fig. 6H), 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. 6I). We saw similar behavior for the TBI data (considering either the entire dataset or just marker genes, Figs. S6.3 E,F) and MERFISH data from cultured MCF10A cells (Fig. S6.3G)[43]. As with our analysis of the giant component, these findings strongly suggest that these power-law 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), as well as in feature-selected 10X data (Fig. S6.4-S6.8), 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[26,30,32,33,35,43,44]. Interestingly, we also see this scaling behavior in other epigenetic datasets, including the BD Rhapsody protein data (Fig. 6J) and scATAC-seq transcription factor data (Fig. S6.3H), and if we use the ℓ^1 norm or the Hamming distance instead of the standard ℓ^2 (Fig. S6.9, S6.10). Further, we applied various normalization and dimensionality reduction techniques to feature and non-feature selected data to test whether these transformations could recover locally dense regions of cells. While some transformations altered the curvature of the distributions, none of the approaches recovered gaussian-like density distributions (Fig. S6.11, S6.12). As with the lack of separation between cell types, non-gaussian density distributions are thus a universal feature of the single-cell data we analyzed.

Discussion

For over 80 years, Waddington's landscape has been the dominant picture used to explain canalization in ontogeny[5,7,37,39,45,46]. 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)[5,7,37,39,44,46]. 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, constraining and producing different phenotypic fates[5–7,37,39,47]. These concepts are deployed extensively in the study of cancer, stem cells and stem cell reprogramming, and other areas [5,37,39,47]. 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) [5–7,37]. 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), regardless of how we select the subset of genes on which to focus our analysis.

We found this not just for scRNA-seq data taken from the 10X platform, 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. 3D). This lack of separation is also present in chromatin accessibility data and data on protein levels (Figs. 3G, S5.3E). 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. 6 and Fig. 7A).

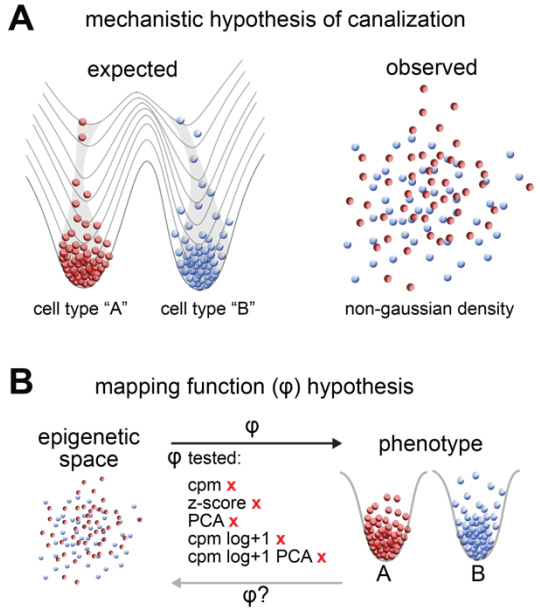

Figure 7. Current single-cell epigenetic data does not support our current conception of cell types. A) Schematic illustrating cell-type densities predicted from the mechanistic hypothesis of canalization depicted in Waddington's Landscape compared to the observed non-gaussian densities. B) Schematic illustrating the inability to find a function that maps transcriptome space back into a phenotypically separable cell-type space.

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[4,18]. 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. We tested the hypothesis that this "standard pipeline" has discovered a way to approximate a function that maps the epigenetic state to phenotype, and recovers physiologically distinct groups of cells (Fig. 7B). We found that application of the standard pipeline of scRNA-seq analysis sometimes does separate cells into discrete, homogenous groups (Fig. 5A). In the vast majority of cases, however, this approach simply does not "work:" these transformations generally project groups of physiologically distinct cells into essentially the same region of state space, with non-gaussian density distributions (Fig. 6 and Fig. S5.2, S5.3, S5.4). These findings suggest the results of the standard pipeline are not robust and can create unreliable cell-type partitions; in some studies, even with considerable effort, one cannot separate the data into reasonable, distinct groups of cells[11,48]. 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[5,45]. 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 space of merely mRNA levels. While we have considered protein levels and patterns of chromatin accessibility in this work, it is possible that further advances in the technologies that generate these data could reveal a more Waddington-like pattern in those data. In particular, advances in tools that allow single-cell characterization of protein levels, especially at the genome-wide level, could eventually allow us to determine if the patterns we observe here hold for a larger set of proteins and across a wider range of systems [49–51]. An alternative possibility could be that Waddington's landscape involves a completely different epigenetic state of the cell. In this scenario, cell types would 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. 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[4,18]. 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 primarily based on 10X data, which is limited in its ability to robustly detect transcripts with low copy number[3,25,52]. 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[8–10,48]. 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 Supplementary 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 [30]. The count-by-cell matrix for this dataset was kindly provided by Zach Hemminger and Roy Wollman. Similarly, the Vizgen corporation has made image data from their MERFISH experiments on mouse brain slices freely available on the web[29]. Data for one of these slices was obtained through image analysis as described in, and again kindly provided by Zach Hemminger and Roy Wollman. All the MERFISH count matrices are 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 focused on 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. We considered other types of metrics, including the l_1 norm and Hamming distance, as described in the text and supplementary information. Calculating the distance between every pair of cells results in 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[41]. 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[53] 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[54,55]. All software used in this work is available from the authors upon request.

Acknowledgements

The authors thank Alexander Hoffmann, Tom Kolokotrones, Roy Wollman, Jukka Keranen, Pavak Shah, Zach Hemminger, Nina Gilshteyn, and other members of the Deeds lab for many helpful discussions and comments. We also thank Zach Hemminger and Roy Wollman with their assistance in analyzing MERFISH data. 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., and Zhu, X. (2019). The single-cell sequencing: new developments and medical applications. Cell Biosci. *9*, 53. 10.1186/s13578-019-0314-y.

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

3. Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. *9*, 75. 10.1186/s13073-017-0467-4.

4. Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. *20*, 273–282. 10.1038/s41576-018-0088-9.

5. Goldberg, A.D., Allis, C.D., and Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. Cell *128*, 635–638. 10.1016/j.cell.2007.02.006.

6. Matsushita, Y., and Kaneko, K. (2020). Homeorhesis in Waddington's landscape by epigenetic feedback regulation. Phys. Rev. Res. *2*, 023083. 10.1103/PhysRevResearch.2.023083.

7. 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. *108*, 8257–8262. 10.1073/pnas.1017017108.

8. Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. *33*, 155–160. 10.1038/nbt.3102.

9. SoRelle, E.D., Dai, J., Bonglack, E.N., Heckenberg, E.M., Zhou, J.Y., Giamberardino, S.N., Bailey, J.A., Gregory, S.G., Chan, C., and Luftig, M.A. (2021). Single-cell RNA-seq reveals transcriptomic heterogeneity mediated by host–pathogen dynamics in lymphoblastoid cell lines. eLife *10*, e62586. 10.7554/eLife.62586.

10. Muhl, L., Genové, G., Leptidis, S., Liu, J., He, L., Mocci, G., Sun, Y., Gustafsson, S., Buyandelger, B., Chivukula, I.V., et al. (2020). Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. Nat. Commun. *11*, 3953. 10.1038/s41467-020-17740-1.

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

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

13. Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095. 10.1038/nmeth.2645.

14. Moffitt, J.R., and Zhuang, X. (2016). Chapter One - RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). In Methods in Enzymology Visualizing RNA Dynamics in the Cell., G. S. Filonov and S. R. Jaffrey, eds. (Academic Press), pp. 1–49. 10.1016/bs.mie.2016.03.020.

15. Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. *37*, 925–936. 10.1038/s41587-019-0206-z.

16. Network Science by Albert-László Barabási.

17. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. 10.1038/ncomms14049.

18. Lueken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. 10.15252/msb.20188746.

19. Garfinkel, A., Shevtsov, J., and Guo, Y. (2017). Modeling Life (Springer International Publishing) 10.1007/978-3-319-59731-7.

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

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

22. Kim, J.K., Kolodziejczyk, A.A., Ilicic, T., Teichmann, S.A., and Marioni, J.C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun. *6*, 8687. 10.1038/ncomms9687.

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

24. Mann, H.B., and Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Stat. *18*, 50–60. 10.1214/aoms/1177730491.

25. Wang, X., He, Y., Zhang, Q., Ren, X., and Zhang, Z. (2021). Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. Genomics Proteomics Bioinformatics *19*, 253–266. 10.1016/j.gpb.2020.02.005.

26. Mair, F., Erickson, J.R., Voillet, V., Simoni, Y., Bi, T., Tyznik, A.J., Martin, J., Gottardo, R., Newell, E.W., and Prlic, M. (2020). A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level. Cell Rep. *31*, 107499. 10.1016/j.celrep.2020.03.063.

27. Tran, V., Papalexi, E., Schroeder, S., Kim, G., Sapre, A., Pangallo, J., Sova, A., Matulich, P., Kenyon, L., Sayar, Z., et al. (2022). High sensitivity single cell RNA sequencing with split pool barcoding. 2022.08.27.505512. 10.1101/2022.08.27.505512.

28. 20k Human PBMCs, 3’ HT v3.1, Chromium X 10x Genomics. https://www.10xgenomics.com/resources/datasets/20-k-human-pbm-cs-3-ht-v-3-1-chromium-x-3-1-high-6-1-0.

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

30. Littman, R., Hemminger, Z., Foreman, R., Arneson, D., Zhang, G., Gómez-Pinilla, F., Yang, X., and Wollman, R. (2021). Joint cell segmentation and cell type annotation for spatial transcriptomics. Mol. Syst. Biol. *17*, e10108. 10.15252/msb.202010108.

31. 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.

32. Yao, Z., van Velthoven, C.T.J., Nguyen, T.N., Goldy, J., Sedeno-Cortes, A.E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., et al. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell *184*, 3222-3241.e26. 10.1016/j.cell.2021.04.021.

33. 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 C. elegans embryogenesis at single-cell resolution. Science *365*, eaax1971. 10.1126/science.aax1971.

34. 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. 10.1016/0012-1606(83)90201-4.

35. Shulse, C.N., Cole, B.J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., Turco, G.M., Zhu, Y., O'Malley, R.C., Brady, S.M., et al. (2019). High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types. Cell Rep. *27*, 2241-2247.e4. 10.1016/j.celrep.2019.04.054.

36. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. 10.1038/nbt.3192.

37. Ferrell, J.E. (2012). Bistability, Bifurcations, and Waddington's Epigenetic Landscape. Curr. Biol. *22*, R458–R466. 10.1016/j.cub.2012.03.045.

38. Moris, N., Pina, C., and Arias, A.M. (2016). Transition states and cell fate decisions in epigenetic landscapes. Nat. Rev. Genet. *17*, 693–703. 10.1038/nrg.2016.98.

39. Shakiba, N., Li, C., Garcia-Ojalvo, J., Cho, K.-H., Patil, K., Walczak, A., Liu, Y.-Y., Kuehn, S., Nie, Q., Klein, A., et al. (2022). How can Waddington-like landscapes facilitate insights beyond developmental biology? Cell Syst. *13*, 4–9. 10.1016/j.cels.2021.12.003.

40. Jeong, H., Néda, Z., and Barabási, A.L. (2003). Measuring preferential attachment in evolving networks. Europhys. Lett. EPL *61*, 567–572. 10.1209/epl/i2003-00166-9.

41. Clauset, A., Shalizi, C.R., and Newman, M.E.J. (2009). Power-law distributions in empirical data. SIAM Rev. *51*, 661–703. 10.1137/070710111.

42. Zhou, B., Meng, X., and Stanley, H.E. (2020). 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. 10.1073/pnas.1918901117.

43. Foreman, R., and Wollman, R. (2020). Mammalian gene expression variability is explained by underlying cell state. Mol. Syst. Biol. *16*, e9146. 10.15252/msb.20199146.

44. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. 10.1038/ncomms14049.

45. Baedke, J. (2013). 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. 10.1016/j.shpsc.2013.06.001.

46. Nicoglou, A. (2018). Waddington's epigenetics or the pictorial meetings of development and genetics. Hist. Philos. Life Sci. *40*, 61. 10.1007/s40656-018-0228-8.

47. Kaity, B., Sarkar, R., Chakrabarti, B., and Mitra, M.K. (2018). Reprogramming, oscillations and transdifferentiation in epigenetic landscapes. Sci. Rep. *8*, 7358. 10.1038/s41598-018-25556-9.

48. Goldman, S.L., MacKay, M., Afshinnekoo, E., Melnick, A.M., Wu, S., and Mason, C.E. (2019). The Impact of Heterogeneity on Single-Cell Sequencing. Front. Genet. *10*.

49. Bennett, H.M., Stephenson, W., Rose, C.M., and Darmanis, S. (2023). Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. Nat. Methods *20*, 363–374. 10.1038/s41592-023-01791-5.

50. Brunner, A.-D., Thielert, M., Vasilopoulou, C., Ammar, C., Coscia, F., Mund, A., Hoerning, O.B., Bache, N., Apalategui, A., Lubeck, M., et al. (2022). Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. Mol. Syst. Biol. *18*, e10798. 10.15252/msb.202110798.

51. Minakshi, P., Kumar, R., Ghosh, M., Saini, H.M., Ranjan, K., Brar, B., and Prasad, G. (2019). Chapter 14 - Single-Cell Proteomics: Technology and Applications. In Single-Cell Omics, D. Barh and V. Azevedo, eds. (Academic Press), pp. 283–318. 10.1016/B978-0-12-814919-5.00014-2.

52. Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. Nat. Protoc. *13*, 599–604. 10.1038/nprot.2017.149.

53. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. 10.1186/s13059-017-1382-0.

54. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. 10.1109/MCSE.2007.55.

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