Genome Biology

METHOD

Open Access

ESCHR: a hyperparameter-randomized ensemble approach for robust clustering across diverse datasets

Sarah M. Goggin[1] and Eli R. Zunder[1,2*]

*Correspondence:
ezunder@virginia.edu

[1] Neuroscience Graduate Program, School of Medicine, University of Virginia, Charlottesville, VA 22902, USA
[2] Department of Biomedical Engineering, School of Engineering, University of Virginia, Charlottesville, VA 22902, USA

Abstract

Clustering is widely used for single-cell analysis, but current methods are limited in accuracy, robustness, ease of use, and interpretability. To address these limitations, we developed an ensemble clustering method that outperforms other methods at hard clustering without the need for hyperparameter tuning. It also performs soft clustering to characterize continuum-like regions and quantify clustering uncertainty, demonstrated here by mapping the connectivity and intermediate transitions between MNIST handwritten digits and between hypothalamic tanycyte subpopulations. This hyperparameter-randomized ensemble approach improves the accuracy, robustness, ease of use, and interpretability of single-cell clustering, and may prove useful in other fields as well.

Keywords: Single-cell RNA-seq, Mass cytometry, Consensus clustering, Soft clustering

Background

Clustering is widely used for exploratory data analysis across diverse fields, where it is applied to identify dataset grouping structures in an unsupervised manner. In particular, clustering has become a workhorse tool for single-cell analysis, enabling the identification and characterization of cell populations that share similar molecular profiles within heterogeneous biological samples [1]. The output of clustering analysis is often used for direct comparison of biological samples, to identify changes in the abundance or molecular state of specific cell populations. Furthermore, clustering output is frequently carried forward into additional downstream analyses such as cell type classification or trajectory analysis [2–4]. Therefore, the accuracy and reproducibility of clustering partitions is important for the quality of single-cell analysis. This importance has motivated the development of hundreds [5] of clustering methods with a variety of algorithmic strategies, but there are still important shortcomings in all of these methods which reduce their effectiveness.

An ideal clustering method for single-cell analysis would satisfy the following requirements:

1) Operate without the need for human input such as hyperparameter tuning. The vast majority of existing methods require selection and optimization of hyperparameters, which can significantly impact clustering quality [6–9]. Manual hyperparameter tuning is time-consuming and relies subjectively on human intuition about which groupings appear correct [10]. Automated methods have been proposed to overcome this limitation, but many are computationally inefficient, and all are biased by the criteria used for optimization [9, 11–13].
2) Perform well across diverse single-cell datasets from different tissues and across multiple measurement modalities such as single-cell/single-nucleus RNA sequencing (scRNA-seq and snRNA-seq), single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq), flow cytometry, mass cytometry, and multiplexed imaging analysis such as high-content fluorescence imaging, imaging mass cytometry (IMC), multiplexed ion beam imaging (MIBI), and multiplexed error-robust fluorescence in situ hybridization (MERSCOPE). Generalizability is a concern in existing methods; many clustering methods perform well on gold-standard single-cell datasets, but do not generalize well to datasets from other tissue types or from other single-cell analysis modalities which may have different or more complex distributions or structural properties [7–10, 14].
3) Produce stable and consistent partitions that are robust to random sampling and minor perturbations. Existing methods do not reliably produce robust partitions when applied to complex, high-dimensional single-cell datasets. Meaningfully different results can be produced with different hyperparameter combinations [8], slight perturbations of a dataset [10, 14], or even when an identical dataset and hyperparameters are run multiple times due to randomization steps in most clustering algorithms (Additional File 1: Fig. S1a, b).
4) Capture and describe the wide variety of discrete and continuous grouping structures present in single-cell datasets [15, 16]. Most existing methods implement hard clustering, which assumes a data structure with discrete, well-separated groups, but is unable to characterize overlap or continuity between groups. Alternative computational methods for trajectory inference can better capture specific types of continuum-like processes such as cell differentiation in single-cell datasets, but these methods make a different set of assumptions about data structure that can be equally restrictive.
5) Quantify uncertainty at the levels of individual data points and clusters. There are many scenarios where clustering can provide useful information, but a single optimal solution to the clustering task either does not exist or cannot be determined [17]. In many cases, there is additionally no known ground truth that could define what a correct solution might look like. Therefore, measures of uncertainty are crucial to assess the reliability and aid interpretability of clustering results before using them as inputs for downstream analytical methods or for purposes such as hypothesis development or orthogonal validation of results.

6) Scale to analyze large single-cell datasets with millions of cells. While many of the most commonly used methods are scalable, several that have been developed to address these key challenges for clustering have done so at the expense of scalability. Methods that improve on these other challenges can only be realistically impactful if they can produce results for the large dataset sizes that are becoming increasingly commonplace.

Recently developed clustering methods have made progress towards some of these goals. Ensemble and consensus methods represent a promising approach to improve clustering robustness by combining information from multiple diverse partitions [18–25]. Fuzzy and soft clustering methods allow data points to belong to multiple clusters, and can therefore be used to provide a more complete description of both continuous and discrete data structures [26, 27]. There are several methods that provide measures of stability or uncertainty at the cluster level [9, 11, 24, 28], but cell-level measures of uncertainty are rarely provided in single-cell methods [29, 30]. Additionally, deep learning methods have shown promise in generating informative lower-dimensional representations of diverse types of high-dimensional biological data [31]. However, none of these approaches have been able to incorporate all of the six key features described above.

To address this need for a single method that performs robustly across diverse datasets with no hyperparameter tuning and transparently communicates uncertainty, we developed a clustering algorithm that applies *E*n*S*emble *C*lustering with *H*yperparameter *R*andomization (ESCHR). This algorithm requires no human input due to hyperparameter randomization, which explores a wide range of data subspaces that contribute to the final consensus clustering step. Our implementation of ESCHR in Python (https://github.com/zunderlab/eschr) [32] can be used as a self-contained framework for clustering, or it can be integrated into commonly used single-cell analysis pipelines such as the scverse ecosystem [33]. To evaluate this new method, we performed extensive benchmarking tests, which demonstrated that ESCHR outperforms both general clustering methods and clustering methods specifically developed for single-cell analysis [24, 25, 34–41] in terms of accuracy on synthetic datasets with a known "ground truth" and in terms of robustness on real single-cell datasets encompassing diverse tissues (bone marrow, pancreas, developing and adult brain), organisms (mouse, human), cell numbers (from hundreds to millions), and measurement techniques (single-cell RNA sequencing, mass cytometry, flow cytometry).

After benchmarking for accuracy and robustness, we applied ESCHR clustering to two complex real-world datasets—first to the MNIST dataset [42], a commonly used example for machine learning image analysis, and then in the single cell context to investigate the relationships between tanycyte populations in the hypothalamus, which have been previously shown to display spatial and molecular-level continuity between subtypes [43–47]. In both of these exploratory analyses, the soft cluster assignments and uncertainty scoring from ESCHR were used to identify regions of low confidence cluster assignments corresponding to transitional overlap between clusters and map the key feature transitions that define these regions.

Results

Overview of ESCHR clustering

To develop a robust and scalable clustering method for the analysis of single-cell datasets, we employed an ensemble and consensus approach, which has been shown to improve robustness across many domains of machine learning [21, 48–54]. This approach consists of two main steps: (1) generate a set of base partitions, referred to as the ensemble, and (2) use this ensemble to generate a final consensus partition. The graph-based Leiden community detection method [55] was selected as a base algorithm to generate the clustering ensemble, because it is widely used for single-cell analysis, and is efficiently implemented to be scalable for large datasets [3].

A key element of successful consensus approaches is generating sufficient diversity in the ensemble [21, 49, 50, 56]. To generate this diversity, ESCHR randomizes four hyperparameters for each base partition: subsampling percentage, number of nearest neighbors, distance metric, and Leiden resolution. Within a given base partition, ESCHR first selects a subsampling percentage by random sampling from a Gaussian distribution with μ scaled to dataset size (within 30–90%) and then extracts the specified subset of data from the full dataset. Next, ESCHR randomly selects values for the number of nearest neighbors (15–150) and the distance metric (euclidean or cosine) and uses these to build a k-nearest neighbors (kNN) graph for the extracted subset of data. Finally, ESCHR performs Leiden community detection on this kNN graph using a randomly selected value for the required resolution-determining hyperparameter (0.25–1.75). The ranges for randomization of these hyperparameters were optimized empirically (Additional File 1: Fig. S2a–f and Methods). This subsampling and randomization scheme is used to produce diversity among each of the different base partitions (Fig. 1a). This diversity provides many different views of the dataset, and the full ensemble of these views provides a more comprehensive picture of the dataset grouping structure (Additional File 1: Fig. S3), which is less likely to be influenced by the stochastic variations present in any single view, including the full unsampled dataset. In addition to generating ensemble diversity, this hyperparameter randomization approach is what enables ESCHR to operate without the need for hyperparameter tuning at this first stage of the algorithm.

After generating a diverse ensemble of base partitions, ESCHR applies a bipartite graph clustering approach to obtain the final consensus partition. First, the base partitions are assembled into a bipartite graph, where cells are represented by one set of vertices, base clusters are represented as a second set of vertices, and each cell is connected by an edge to each of the base clusters it was assigned to throughout the ensemble (Fig. 1b). Next, ESCHR applies bipartite community detection to obtain the final consensus partition (Fig. 1b) [57]. Bipartite community detection is applied here instead of more common consensus approaches that suffer from information loss [58]. To remain hyperparameter-free without the need for human intervention in this consensus stage of the algorithm, ESCHR performs internal hyperparameter selection to determine the optimal resolution for the final consensus clustering step by selecting the medoid from a range of resolutions (Additional File 1: Fig. S4). After obtaining the final consensus partition, ESCHR converts the ensemble bipartite graph to a final weighted bipartite graph by collapsing all base partition cluster nodes assigned to the same consensus cluster into a single node. Cells are then connected to these consensus cluster nodes by edges

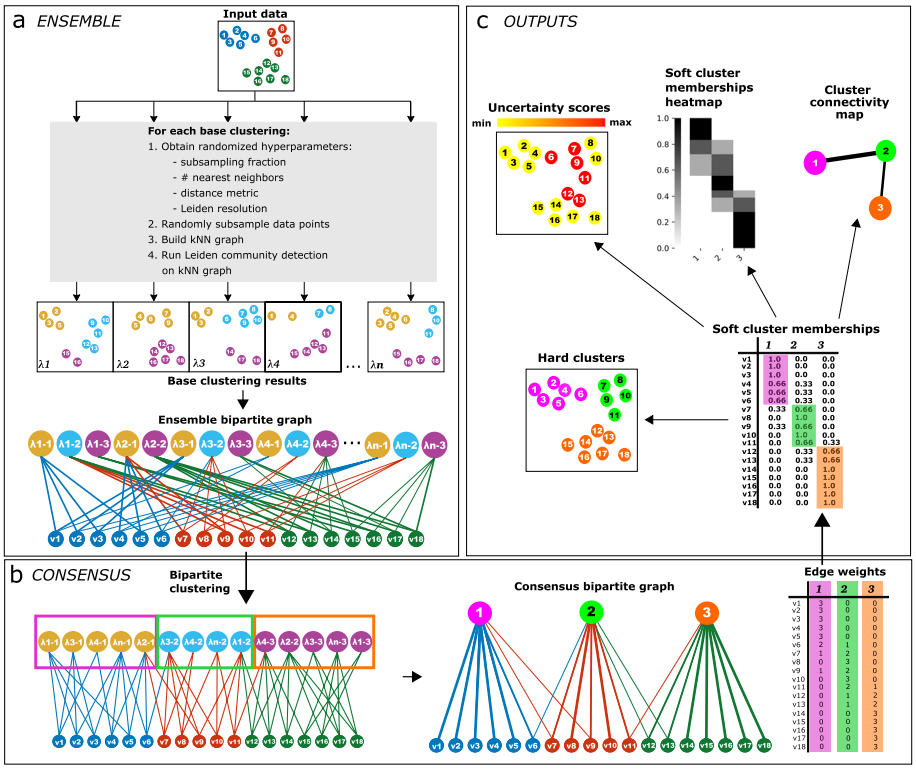

Fig. 1 ESCHR framework overview. **a** Starting from a preprocessed input dataset, ESCHR performs ensemble clustering using randomized hyperparameters to obtain a set of base partitions. This set of base partitions is represented using a bipartite graph where one type of node consists of all data points and one type of node consists of all clusters from all base partitions and edges exist between data points and each base cluster they were assigned to throughout the ensemble. **b** Leiden bipartite clustering is performed on the ensemble bipartite graph. Base clusters are collapsed into their assigned consensus clusters obtained through the bipartite clustering and edge weights are summed such that each data point now has a weighted edge to each consensus cluster representing the number of base clusters it had been assigned to that were then collapsed into that consensus cluster. **c** Soft cluster memberships are obtained by scaling edge weights between 0 and 1, and can then be visualized directly in heatmap form and used to generate hard cluster assignments, per-data point uncertainty scores, and cluster connectivity maps

with weights representing the number of times each cell was assigned to any of the base partition clusters that were collapsed into a given consensus cluster (Fig. 1b). These raw membership values are then normalized to obtain proportional soft cluster memberships, and hard cluster labels are assigned as the consensus cluster in which a cell has the highest proportional membership (Fig. 1c).

While many analysis strategies for single-cell datasets require hard clustering labels, these by definition cannot convey whether a cell is at the borderline between multiple clusters or located firmly in the center of a single cluster. Hard clusters also do not provide any insight into potential continuity between clusters. Using the soft cluster memberships derived from the weighted consensus bipartite graph, ESCHR provides several additional outputs beyond hard cluster assignments that enable a more comprehensive characterization of the grouping structures within a dataset. Firstly, soft cluster memberships can be directly visualized in heatmap form to identify areas of cluster overlap at the single-cell level (Fig. 1c). Importantly, these soft membership heatmap visualizations can serve as complements or even alternatives to the widely used but also widely

misinterpreted [59] stochastic embedding methods (i.e. UMAP [60], t-SNE [61]) for visualizing the complex relational structures within single-cell datasets. ESCHR also produces an Uncertainty Score for every object, derived from its soft cluster membership, which quantifies regions of higher and lower certainty in hard cluster assignment (Fig. 1c). Finally, ESCHR produces a cluster-level map of the continuity structure within a dataset by using the soft cluster memberships to calculate a corrected-for-chance measure of the connectivity between each pair of hard clusters (Fig. 1c and Methods).

ESCHR soft clustering and uncertainty scores capture diverse structural characteristics and quantify uncertainty in cluster assignments

We first sought to examine how ESCHR uncertainty scores and soft clustering could enable effective and informative analysis for datasets containing complex combinations of continuity and discreteness, and how these results compared to a wide range of alternative clustering methods used for single-cell analysis or general purpose clustering (Additional File 2: Table S1 and Methods). For this analysis, we generated a synthetic scRNA-seq dataset containing 1000 cells and 1000 features using the DynToy package [62]. This dataset is generated by sampling "cells" from a complex trajectory model, with library size and transcript distributions per cell modeled on a real scRNA-seq dataset. Specifically, "cells" are sampled from prototypical "cell states", where each cell has a varying probability of belonging to multiple neighboring states, and the ground truth hard cluster labels are assigned as the state in which the cell has the highest percent membership. This process generates a dataset which is similar to real single-cell data but provides known ground truth grouping structure and known ground truth continuity structure (Fig. 2a–b, Additional File 1: Fig. S7a), which is not generally available for real datasets (Additional File 1: Supplementary Note 1).

We first compared the ESCHR hard clustering results (Fig. 2c, Additional File 1: Fig. S7b) and uncertainty scores (Fig. 2d) with the true hard cluster labels and the true membership percentage for those labels. While ESCHR successfully captures all of the ground truth cell states, it also adds two additional clusters (ESCHR clusters 9 and 6) between true clusters M2 and M3 and between M1 and M7. However, the ground truth membership percentages show that these regions are highly transitional, with low percentages for the maximum membership (Fig. 2b). ESCHR uncertainty scores correspond closely to this observed ground truth continuity in Fig. 2b, indicating that the uncertainty scores can identify regions of uncertainty in cluster assignment due to ground truth continuity and cluster overlap. In addition to quantifying this level of uncertainty per "cell," ESCHR also provides information at the cluster level about which clusters overlap, and to what extent, through direct visualization of the soft cluster memberships. This reveals an overlap structure that corresponds to the ground truth patterns of transitional membership between groups, such as between ESCHR clusters 7, 9, and 1 (corresponding to true labels M2 and M3) and ESCHR clusters 1, 8, and 2 (corresponding to true labels M3, M6, and M5) (Fig. 2e).

We next evaluated the results from multiple different clustering methods and found that there was wide disagreement between the results of these different methods (Fig. 2f, Additional File 1: Fig. S7c). Seurat, Scanpy, and Phenograph, which are all based on either Leiden or Louvain as their base clustering method, all identify approximately

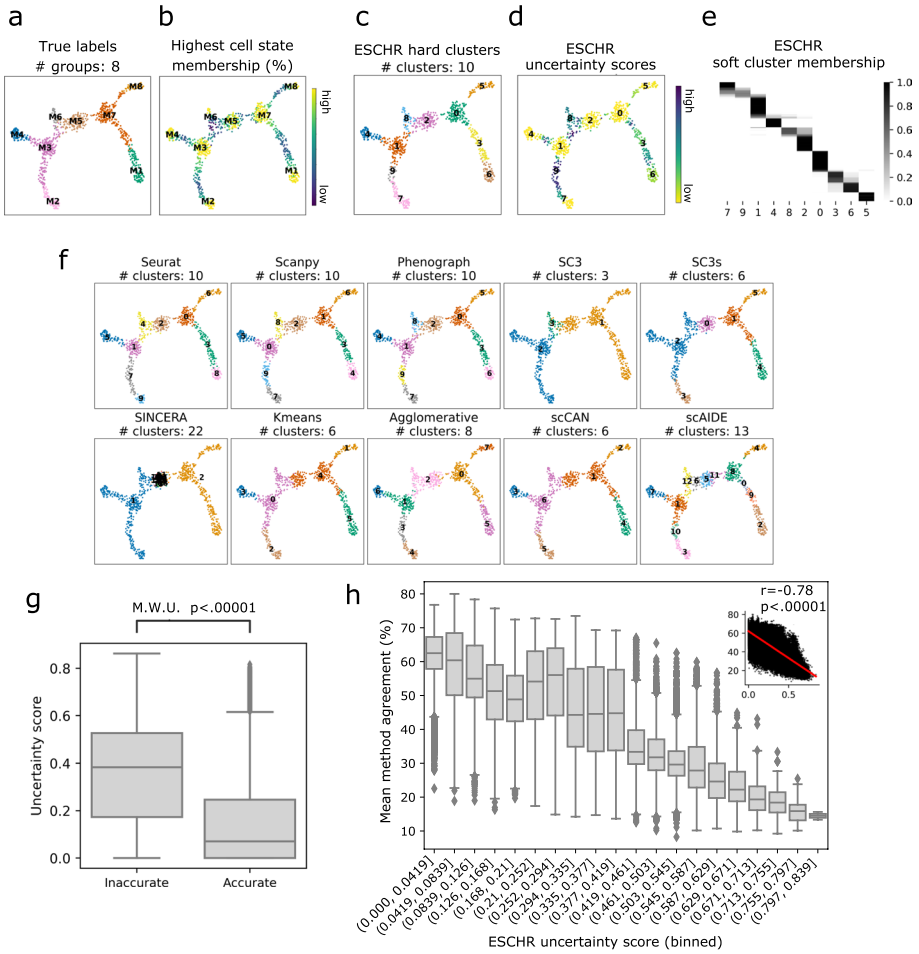

Fig. 2 Visualization of ESCHR clustering and uncertainty scores compared to other clustering methods. UMAP visualizations of **a** ground truth cluster labels, **b** ground truth cell state membership, **c** ESCHR hard clusters, and **d** ESCHR uncertainty scores. **e** Heatmap visualization of ESCHR soft cluster memberships. **f** UMAP visualizations of cluster assignments from selected comparison methods. Points are colored by cluster ID. **g** Box and whisker plot comparing uncertainty scores of data points from ESCHR hard clustering that were accurately assigned versus not accurately assigned. The box shows the quartiles of the dataset, whiskers extend to 1.5*IQR, plotted points are outliers. Two-sided Mann–Whitney *U* test was used for statistical analysis. $N = 126{,}545$, 750,955 for inaccurate and accurate groups respectively. **h** Comparison of ESCHR uncertainty scores versus method agreement per each individual data point. Primary box and whisker plot *x*-axis is binned ESCHR uncertainty scores and *y*-axis is the average method agreement across all pairs of methods; inset scatterplot shows raw data (i.e., not binned) with a red line of best fit and Pearson correlation statistic

the same clusters as ESCHR, but importantly each of these methods has selected different boundaries between these clusters. While the results from the remaining methods exhibit more diversity, it is notable that none have placed cluster boundaries within the regions of ground truth high single state membership but rather have over-clustered transitional regions or under-clustered by grouping multiple true clusters together. The regions of disagreement between the different clustering methods highlight areas that are challenging for and perhaps not well suited to the discreteness assumptions of traditional hard clustering. High ESCHR uncertainty scores and overlapping soft cluster memberships correspond to regions of disagreement between other clustering methods,

providing further evidence that these metrics can help identify regions that are challenging for traditional clustering methods due to continuous data structures such as overlap between ground truth clusters.

To assess whether ESCHR uncertainty scores were similarly informative across diverse datasets, we generated an additional 4 simulated datasets using DynToy and 16 additional structurally diverse synthetic datasets which consist of randomly generated Gaussian distributions varying in number of objects (5000 or 10,000), number of features (20, 40, 50, 60), number of clusters (3, 8, 15, 20), cluster sizes, cluster standard deviations, cluster overlap, and feature anisotropy (Additional File 1: Figs. S5–S6, Additional File 2: Tables S2–S3). To quantitatively evaluate the utility of ESCHR uncertainty scores across our full set of 21 structurally diverse synthetic datasets with ground truth cluster labels, we first compared ESCHR uncertainty scores to the accuracy of assignment compared to ground truth labels per data point across all datasets, and found that ESCHR uncertainty scores were significantly higher in inaccurately assigned cells (Fig. 2g). We then quantified the level of agreement between clustering assignments from all the different clustering algorithms we tested (in Fig. 2f) and used this as an alternative external indicator for per data point uncertainty and difficulty of clustering (Methods). This analysis revealed that higher ESCHR uncertainty scores were significantly negatively correlated with method agreement (Fig. 2h). Taken together, these comparisons demonstrate that ESCHR uncertainty scores identify meaningful uncertainty and that when used in combination with the soft clustering results, they enable more in-depth interpretation of dataset structure than other methods which produce only hard cluster assignments. Furthermore, ESCHR is able to provide these high-quality insights for datasets with diverse structural characteristics without the need for human intervention such as hyperparameter tuning.

ESCHR outperforms other methods across measures of accuracy and robustness

To systematically evaluate the performance of ESCHR vs. other clustering methods on real datasets as well as synthetic ones, we performed systematic benchmarking of ESCHR against other clustering algorithms (Additional File 2: Table S1) using a collection of 45 published real datasets in addition to the 21 synthetic datasets described above. This collection of 45 published datasets vary widely in size (300–2,000,000 cells), source tissue (e.g. blood, bone marrow, brain), measurement type (sc/nRNA-seq, mass cytometry, flow cytometry, non-single-cell datasets), and data structure (varying degrees of discreteness and continuity) (Additional File 2: Table S4). For our evaluation criteria, we selected two extrinsic evaluation metrics, Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI), to assess two aspects of the clustering results: (1) accuracy and (2) robustness. Extrinsic evaluation metrics measure the distance of a clustering result to some external set of labels, and our two selected metrics ARI and AMI represent different approaches to this problem, with divergent biases. ARI tends to yield higher scores in cases of similarly sized clusters and similar numbers of clusters within and between the partitions being compared, while AMI is biased towards purity and yields higher scores when there are shared pure clusters between the two partitions (Methods) [63]. Using ARI and AMI together should therefore provide a more complete comparison of clustering performance [12, 13].

When we applied these extrinsic metrics ARI and AMI to assess clustering accuracy for our collection of synthetic datasets, ESCHR outperformed all other clustering algorithms across both metrics, and this superior performance was statistically significant for all cases (Fig. 3a, Additional File 1: Table S5, Additional File 1: Fig. S8b–c). We also applied ARI and AMI to benchmark clustering accuracy in non-synthetic real datasets, although it is important to note that a priori known class labels do not generally exist for real-world single-cell datasets, and the various proxies accepted as ground truth labels should be interpreted with skepticism (discussed further in Additional File 1: Supplementary Note 1). Keeping these caveats in mind, ESCHR still clustered real datasets more accurately by ARI and AMI than all methods, significantly so in all comparisons except for scCAN and Agglomerative clustering by ARI and only scCAN by AMI (Fig. 3b, Additional File 1: Table S6, Additional File 1: Fig. S8b–c). Many of the ground truth labels that are widely accepted for real single-cell datasets are based on a hierarchical framework of clustering or manual labeling, which could explain why agglomerative clustering performs better relative to the other methods for this particular comparison.

After benchmarking for accuracy, we next used ARI and AMI to evaluate clustering robustness, by comparing results from repeated runs with different random subsamples of a given dataset (Methods). Due to its ensemble and consensus clustering approach, we expected ESCHR to perform well in these tests of robustness, and indeed it demonstrated superior performance to all other clustering algorithms on both synthetic and real data across both ARI and AMI metrics (Fig. 3d–e, Additional File 1: Fig. S8d–e). These results were significant for all comparisons except against scCAN on the real datasets by ARI (Additional File 1: Tables S5–S6). To gain insight into the generalizability of ESCHR versus the other methods for specific dataset types, we calculated the mean rank of each clustering algorithm across all metrics for major subcategories of our collection of datasets: cluster number for synthetic datasets (for which we have reliable ground truth cluster numbers), data modality for real datasets, and sample number across all datasets. Different clustering algorithms perform better or worse for different subsets,

(See figure on next page.)

Fig. 3 Systematic analysis of ESCHR clustering performance compared to competing methods on synthetic and real datasets. **a**–**d** Box and whisker plots comparing accuracy (**a** and **b**) and robustness (**c** and **d**) of results from ESCHR and all comparison methods across all synthetic (**a** and **c**) and real (**b** and **d**) benchmark datasets as measured by ARI (left) and AMI (right). Boxes show the quartiles of the dataset, and whiskers extend to 1.5*IQR. Data points used in the creation of box and whisker plots and shown in overlaid scatterplots are the means across 5 replicates for each dataset. Two-sided Wilcoxon signed-rank test with Bonferroni correction was used for statistical analysis comparing ESCHR to each method. $N = 21$ for comparisons using synthetic datasets and $N = 45$ for comparisons using real datasets. **e** Mean rank across all metrics shown in box-and-whisker plots for different cluster numbers of the synthetic datasets. Error bars show 1 standard deviation. **f** Mean rank across all metrics shown in box-and-whisker plots for different modalities of the real datasets. Points represent means across all replicates of all datasets in a given category and error bars show 1 standard deviation. **g** Mean rank across all metrics shown in box-and-whisker plots for different sample number bins for all real and synthetic datasets. Points represent means across all replicates of all datasets in a given category and error bars show 1 standard deviation. **h** Box plots of difference from the true cluster number for each synthetic dataset for each method. Values below zero reflect calculated cluster numbers being lower than true cluster numbers and higher than zero indicates more clusters than the true cluster number. SINCERA is shown separately due to the scale of values being 2 orders of magnitude different from all other methods. **i** Scalability comparison between ESCHR and other methods on synthetic datasets with an increasing number of data points. *X*-axis is log scaled but labels show the unscaled values for easier interpretation. Each dot represents 5 replicates and error bars show 1 standard deviation

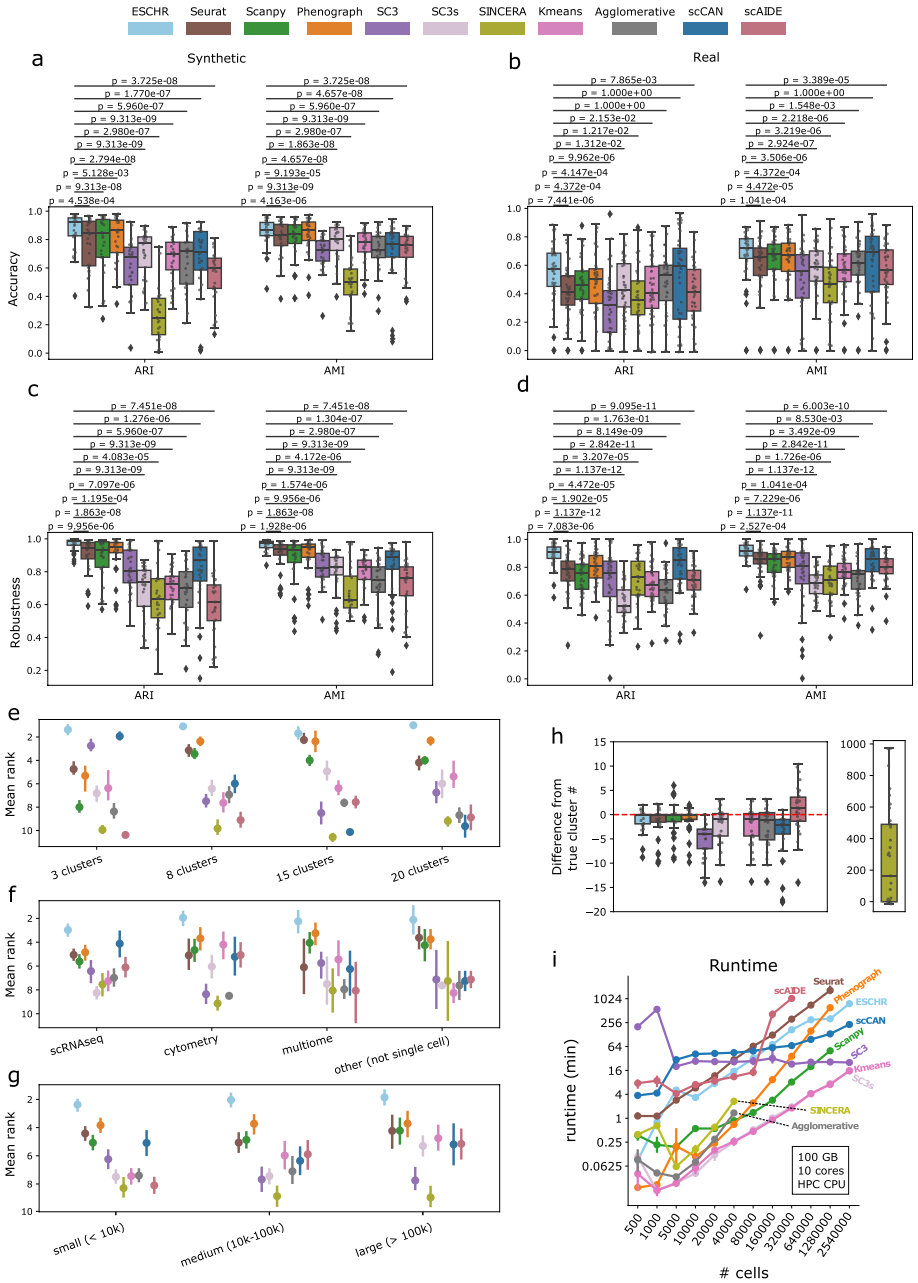

Fig. 3 (See legend on previous page.)

but ESCHR is consistently ranked first or tied for first across these subcategories of both synthetic (Fig. 3e, g) and real datasets (Fig. 3f, g), indicating that its performance is more generalizable to diverse datasets than the other tested clustering algorithms.

We next evaluated the scalability of each method over a range of dataset sizes. While ESCHR generally takes the longest, this does not present a practical limitation for typical usage, as it is able to successfully complete analyses on millions of data points and the runtime scales linearly (Fig. 3g). This analysis also revealed that several of the alternative clustering algorithms we tested could not successfully run to completion for larger datasets. The dataset size limit for ESCHR is effectively the size limit of its underlying base

clustering method, the Leiden algorithm implemented in Python [55]. While it is true that our method does have longer runtimes than some of the commonly used methods we compare to here, we believe it is worth the wait due to the demonstrated superior accuracy and robustness of our results, and perhaps even more importantly due to the additional insights afforded by the uncertainty scores and soft cluster membership information highlighted in Fig. 2. Additionally, the manual guess-and-check hyperparameter tuning that is required to achieve desired results with other methods can be very time-consuming (not to mention highly subjective), and so it is possible that in practical usage ESCHR could potentially end up providing useful results more quickly than other methods. When taken together, these quantitative evaluations demonstrate that ESCHR performs favorably compared to the other methods tested here and achieves our desired goals of providing accurate and robust results, being generalizable to a broad range of diverse datasets, and being scalable to large datasets.

ESCHR soft clustering and uncertainty scores provide increased interpretability in exploratory data analysis of the MNIST dataset

To illustrate how ESCHR can identify regions of continuity and provide insight into cluster overlap and dataset structure, we selected the MNIST dataset for further analysis. This dataset, consisting of 70,000 handwritten digits with ground truth labels, is often used for machine learning demonstrations because the images can be visualized for intuitive interpretation [42]. Other clustering algorithms set to default hyperparameters do not recapitulate the ground truth labels with high accuracy (Additional File 1: Fig. S9a), explained in part by the real variation that exists within the ground truth sets. For example, there are two common variations of the handwritten digit 1, and most of the clustering algorithms capture this difference. Of all the clustering algorithms tested, ESCHR clusters the MNIST dataset with the highest robustness and accuracy (Additional File 1: Fig. S9b), but it consistently splits the 1 and 9 digits into separate subsets (Fig. 4a–b), and in some cases, it splits the digit 4 as well (Additional File 1: Fig. S9a). ESCHR usually produces highly consistent results from run to run thanks to its consensus clustering step, but this inconsistency around the digits 4 and 9 is suggestive of a high degree of continuity within and between these two classes (Additional File 1: Fig. S9c), which is highlighted by elevated ESCHR uncertainty scores in this region (Fig. 4c). The soft cluster membership heatmap also draws attention to the visual similarities between digits 3, 5, and 8, as well as the two types of handwritten 1 digits (Fig. 4d). These subset-level differences and connections between related digits motivated further investigation of the ESCHR outputs for the MNIST dataset.

To further investigate the continuity and overlap structure that was indicated by the uncertainty scores and soft cluster membership heatmap, cluster connectivity mapping was applied to identify significant overlap beyond what would be expected by random chance for the ESCHR clusters (Fig. 4e) (Methods). This revealed significant overlap between clusters "3"– "5"– "8," "1a"– "1b," and "4"– "9a"– "9b." To explore the nature of the continuity structure underlying the significantly overlapping clusters "1a" and "1b," we devised a simple rank ordering scheme based on the soft membership values for the datapoints in these two clusters and then used this ordering score to examine both the continuous progression of soft membership values across the rank-ordered

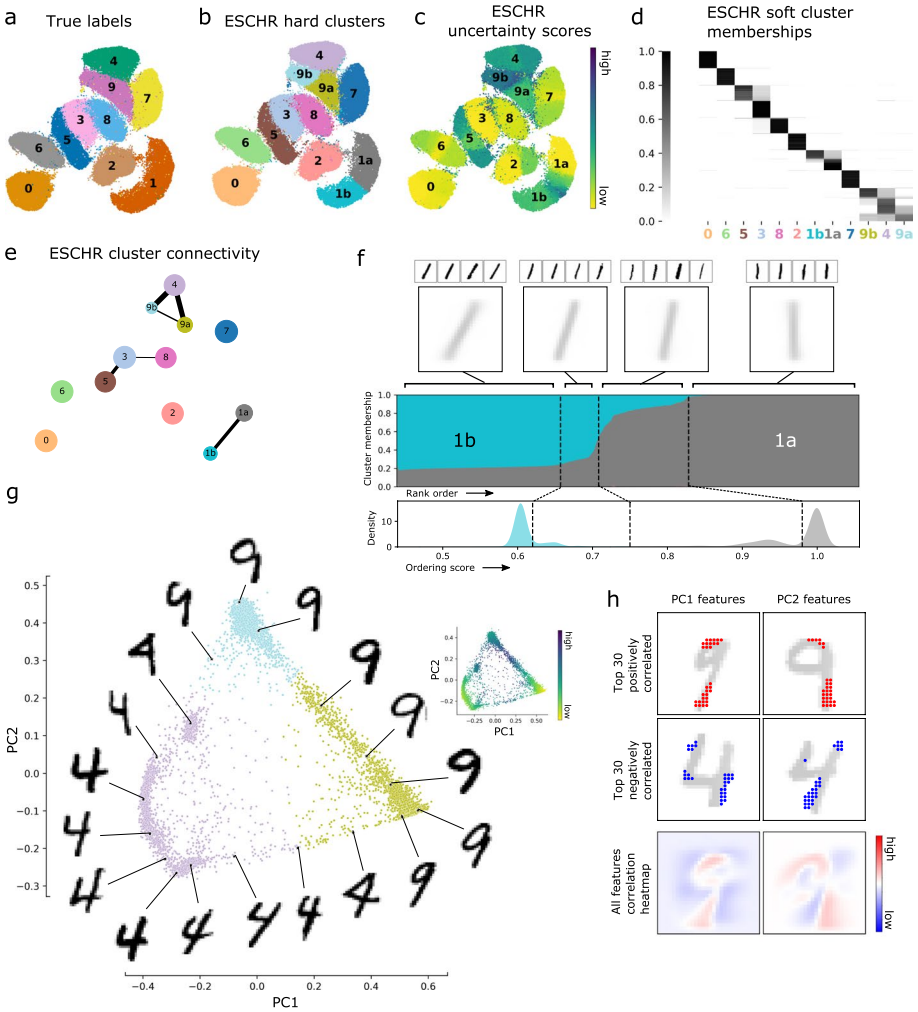

Fig. 4 ESCHR-guided exploration of the benchmarking dataset MNIST. **a** UMAP visualization with points colored by true class labels. **b** UMAP visualization with points colored by ESCHR hard cluster labels. **c** UMAP visualization with points colored by ESCHR uncertainty score. **d** Heatmap visualization of ESCHR soft cluster memberships. **e** Nodes represent ESCHR hard clusters and are located on the centroid of the UMAP coordinates for all data points assigned to that hard cluster. Node size is scaled to the number of data points in a given cluster. Edges exist between nodes that were determined to have significant connectivity by ESCHR cluster connectivity analysis, and edge thickness is scaled to the connectivity score. **f** Stacked bar plot showing the soft membership of datapoints in clusters 1b and 1a, ordered by increasing ESCHR soft cluster membership (SCM) rank ordering score (middle); kernel density estimation across the ordering score (bottom); dashed lines indicate boundaries between ordering score density peaks to separate "core" and "transitional" datapoints (middle and bottom); smaller images show individual representative images and larger images show summed pixel intensities for all datapoints contained within each dashed partition (top). **g** Visualization of data points from ESCHR clusters 4, 9a, and 9b projected onto the first two principal components resulting from PCA performed on the soft membership matrix of these three clusters. The primary scatterplot shows points colored by their ESCHR hard cluster assignment, and the inset scatterplot shows points colored by the ESCHR uncertainty score. Images are real examples from the MNIST dataset. **h** Scatterplot points in the first two rows of plots show the pixel locations of the 30 features with the largest positive (first row, red) and 30 largest negative (second row, blue) Pearson correlation to each of the PCs. Example digit images are underlaid in light gray to aid interpretation. The final row contains heatmaps with each pixel colored according to its Pearson correlation with PC1 (left) or PC2 (right), with bright red indicating a large positive correlation and dark blue indicating a large negative correlation

datapoints and their density along this ordering score (Methods). This revealed that each cluster had a high-density peak of "core" datapoints with a secondary smaller "transitional" peak (Fig. 4f, bottom). Individual representative MNIST digit images (Fig. 4f, top row) and summed pixel intensities (Fig. 4f, second row) from the images within each of these regions indicate that the core "1b" images are heavily slanted whereas the core "1a" images are vertically straight, with the images from the lower density transitional peaks falling in between these extremes. The two high-density peaks consisting of images with distinctly different styles of 1 s explain why ESCHR and many of the other clustering methods tested identified two clusters corresponding to this single digit (Additional File 1: Fig. S9a), while the high degree of pixel overlap between the two styles and the presence of images with intermediate slantedness explain the high degree of continuity and significant overlap detected by ESCHR.

We next examined the more complex relationship between subsets of the digits 4 and 9. Cluster connectivity mapping indicated that there was significant overlap among all three of the ESCHR clusters "4," "9a," and "9b" (Fig. 4e). Additionally in the soft membership heatmap, there appear to be some cells that are overlapping all three clusters, and some cells from clusters "9a" and "9b" that overlap separately with cluster "4" and not with each other (Fig. 4d). Unlike the simpler relationship between ESCHR clusters "1a" and "1b," which could be analyzed by linear one-dimensional reduction, the more complex relationship around digits 4 and 9 could not be adequately captured or described along a single dimension, so principal components analysis (PCA) was applied to the ESCHR soft cluster memberships corresponding to these three clusters in order to reduce these relationships into two dimensions (Methods). Representative images selected from throughout the resulting PC space reveal that the between-cluster continuity is indeed reflecting the existence of a continuous progression through different conformations of the two digits 4 and 9 (Fig. 4f). Specifically, we can see that there is a continuous progression through the 9's based on how slanted they are, with two areas of higher density at either extreme. This explains why a clustering algorithm would be likely to split this into two clusters, albeit with a high amount of uncertainty about precisely where to make the split. The images also illustrate how the more slanted closed 4's form a continuous transition primarily with cluster 9a and the more vertically oriented closed 4's form a continuous transition primarily with cluster 9b. This approach also allows us to identify features that are most correlated with the top two principal components. The top PC-correlated features lend further insight by identifying the specific pixels that are primarily capturing these changes in slantedness and upper loop closure (Fig. 4g). These analyses illustrate how structures within the MNIST dataset are not ideally suited for hard clustering assignment, but also how ESCHR is able to identify these structures and provide deeper insights than could be obtained by other hard clustering methods, or even beyond what is available from the ground truth class assignments.

ESCHR captures cell types and continuity in static adult tissue

To illustrate how ESCHR can provide additional interpretability and insight for single-cell datasets, we selected an integrated scRNA-seq dataset of hypothalamic tanycytes [46] for further analysis. Tanycytes are elongated ependymoglial cells that form the ventricular layer of the third ventricle and median eminence and have historically been

classified into four subtypes (α1, α2, β1, β2) based on the hypothalamic nuclei where they project to, their spatial localization along the third ventricle, and their morphological, structural, genetic, and functional properties (Fig. 5a) [64]. More recent studies have suggested that many of these properties may exhibit substantial continuity between and within each of these subtypes [43–47, 65]. However, individual tanycyte scRNA-seq studies and an integrated analysis of these datasets all reported discrete groupings of tanycytes defined by hard clustering approaches [44, 46, 66–68], with no insight into the robustness of these assignments and whether there is overlap or continuity between them.

Initial ESCHR analysis produced hard clustering outputs that match canonical tanycyte subtypes by their RNA expression profiles (Fig. 5b–d) [46]. Subtypes β1 (expressing Fizb, Penk, Rlbp1, and Ptn) and α2 (expressing Vcan, Nr2e1, Fabp5, and Slc7a11) are represented by multiple hard clusters, while the subtypes β2 (expressing Scn7a, Cal25a1, Meat, and Lrrtm3) and α1 (expressing Mafb, Necab2, Agt, Slc17a8, and Lyz2) each correspond to a single hard cluster, indicating that there is more transcriptional diversity within the β1 and α2 populations. On top of this, however, ESCHR uncertainty scores identify substantial heterogeneity within each hard cluster, including the β2 and α1 clusters (Fig. 5c), and the soft cluster memberships reveal additional levels of overlap and continuity between these canonical tanycyte subtypes (Fig. 5e). ESCHR cluster connectivity mapping (Methods) revealed significant overlap between the β1 clusters (2, 3, and 5) and each of the other three canonical subtypes (Fig. 5f). This result was somewhat unexpected, because transcriptional continuity was previously thought to exist only between spatially neighboring tanycyte subtypes [45, 65]. A more recent study provided evidence that β1 tanycytes exhibit some transcriptional continuity with both α1 and α2 tanycytes, but also indicated that β2 tanycytes were non-overlapping and transcriptionally distinct [43]. Our analysis with ESCHR soft clustering memberships and cluster connectivity provide additional corroboratory evidence for the transcriptional continuity between β1 and α1/α2 tanycytes, but also reveal a previously uncharacterized relationship of transcriptional continuity between β1 and β2 tanycytes.

To further investigate this previously uncharacterized transcriptional overlap between β1 and β2 tanycytes, specifically between ESCHR clusters 1 and 2, we selected the subset of cells comprising the transitional zone between clusters and rank-ordered these based on whether their soft cluster membership was closer to β1 (ESCHR cluster 1) or β2 (ESCHR cluster 2) (Fig. 5g and Methods). Using this rank ordering scheme, we identified genes with expression patterns that correlate with progression through the transition zone from β2 to β1 tanycytes, either decreasing across the transition like Igfbp5 (Fig. 5h, k), peaking during the transition like Tgfb2 (Fig. 5i, l), or increasing across the transition like Crym (Fig. 5j, m). We next sought to determine whether these gene expression patterns in the transitional zone between ESCHR clusters were also observed in the spatial distribution of β2 and β1 tanycytes along the median eminence and third ventricle where these subtypes are thought to reside (Fig. 5n). To investigate this possibility, we examined the in situ hybridization (ISH) database from the Allen Mouse Brain Atlas (ABA; http://mouse.brain-map.org) [69] and observed that the overlapping expression for these three genes did in fact manifest as progressive spatial overlap spanning the anatomical regions canonically associated with β2 and β1 populations (Fig. 5o–q). Altogether,

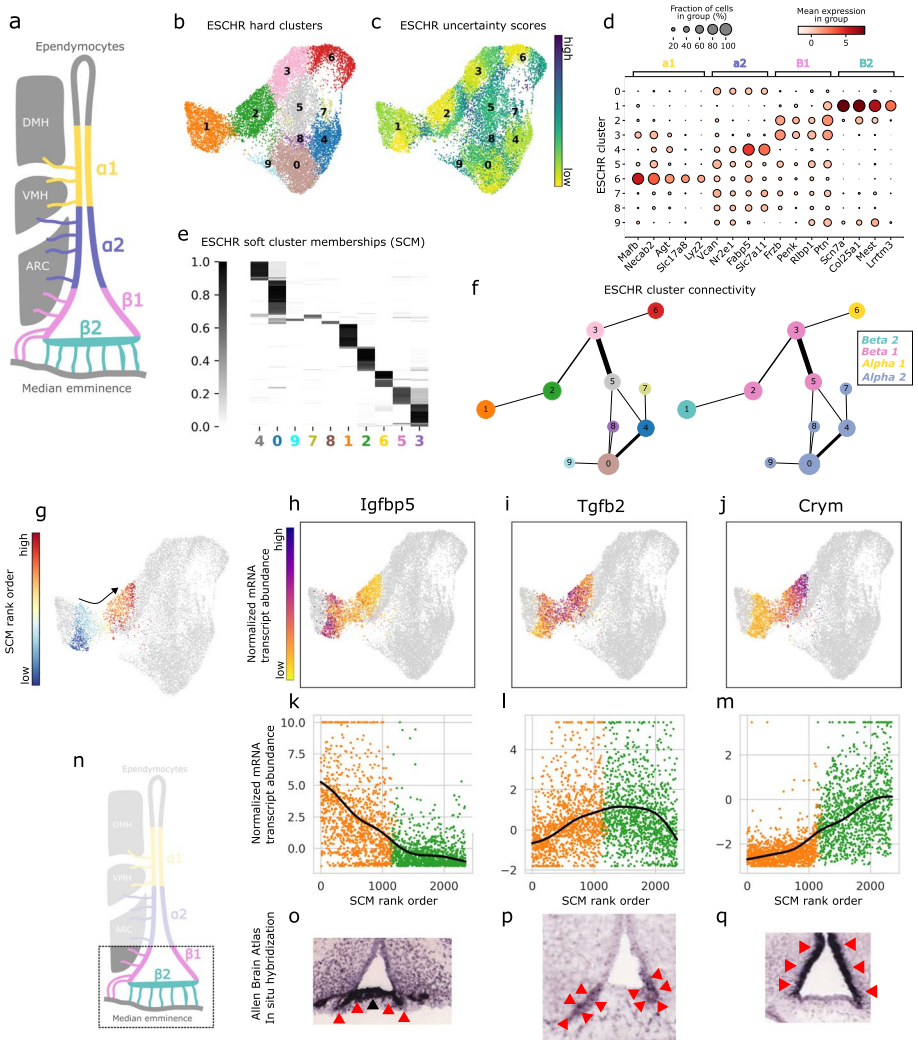

Fig. 5 ESCHR identifies continuity between and within canonical cell subtypes in static adult tissue. **a** Schematic illustration of canonical tanycyte subtypes in their anatomical context surrounding the third ventricle. **b** UMAP visualization with points colored by ESCHR hard cluster labels. **c** UMAP visualization with points colored by ESCHR uncertainty score. **d** Heatmap dotplot showing expression of marker genes for the canonical tanycyte subtypes across the ESCHR hard clusters. **e** Heatmap visualization of ESCHR soft cluster memberships. **f** Nodes represent ESCHR hard clusters and are located on the centroid of the UMAP coordinates for all data points assigned to that hard cluster. Node size is scaled to the number of data points in a given cluster. Edges exist between nodes that were determined to have significant connectivity by ESCHR cluster connectivity analysis, and edge thickness is scaled to the connectivity score. Node colors map to their ESCHR hard cluster colors from panel **b** (left) and to the color from panel **a** of the canonical subtype to which they primarily belong (right). **g** UMAP visualization with the subset of points that were included in the ordering analysis colored by ESCHR soft cluster membership (SCM) rank ordering score, and all others colored gray. **h**–**j** UMAP visualizations where points included in the ordering analysis are colored by their expression level and all others are colored gray. **k**–**m** Scatterplots showing normalized mRNA abundance on the *y*-axis and SCM rank order on the *x*-axis. Expression is bounded between the 2nd and 98th percentiles. Lines show Gaussian-smoothed B-splines fit to the data. **n** Schematic illustration of the anatomical region being shown in **o**–**q**. **o**–**q** In situ hybridization (ISH) of coronal brain sections, using probes specific for Igfbp5, Tgfb2, and Crym (Allen Mouse Brain Atlas). Red arrowheads indicate the areas of expression in the region of interest

this analysis of tanycyte subtypes demonstrates the utility of ESCHR for (1) identifying robust and biologically meaningful hard cluster assignments, (2) providing insight into the overlap and continuity between cell type clusters, and (3) providing a springboard for further analysis of expression level transitions via soft cluster membership ordering.

Discussion and conclusions

Clustering is a fundamental tool for single-cell analysis, used to identify groupings of cell types or cell states that serve as the basis for direct comparisons between biological samples or between specific cell types within a biological sample, as well as numerous further downstream applications. However, it has proven challenging to generate appropriate and consistent cell groupings when using previously available clustering methods on single-cell datasets, due to (1) continuity and overlap between cell types, (2) randomness and stochasticity built into the clustering algorithms, and (3) non-generalizable hyperparameter settings that were optimized for a specific dataset or data type. To overcome these limitations we developed ESCHR, a user-friendly method for ensemble clustering that captures both discrete and continuous structures within a dataset and transparently communicates the level of uncertainty in cluster assignment. Using a large collection of datasets representing a variety of measurement techniques, tissues of origin, species of origin, and dataset sizes, we benchmarked ESCHR's performance against several other clustering algorithms, demonstrating that ESCHR consistently provides the highest robustness and accuracy for clustering across all categories of this diverse dataset collection.

One of the key design features of ESCHR is our approach using hyperparameter randomization during the ensemble generation step. While this was a deliberate design choice to generate diversity among the base clusterings to enhance the robustness and generalizability of clustering, an additional benefit is that it removes the need to manually test and select an optimized set of hyperparameters for each dataset. This design also affords several avenues for potential future improvements to the ESCHR algorithm, such as expanding the number of hyperparameters randomized in order to generate an even more diverse clustering ensemble. For example, we currently use k-nearest neighbor (kNN) graphs for the base Leiden clustering steps, but mutual nearest neighbor (mNN) or shared nearest neighbor (sNN) have shown good performance in other frameworks [3, 41, 70], and may improve ESCHR performance if incorporated as an additional hyperparameter to vary. ESCHR may also benefit from expanding the set of distance metrics utilized. We currently restrict our analysis to euclidean and cosine distances due to their efficient implementations within our chosen fast approximate nearest neighbor (ANN) package [71]. However, recent research has demonstrated the efficacy of a broader range of distance metrics for capturing diverse data structural properties [72]. While not all of these metrics may be applicable in an ANN context, several may hold the potential for enhancing the quality of our clustering outcomes. Additionally, the current version of ESCHR uses only Leiden community detection for clustering in the ensemble stage, but additional base clustering methods could be explored and potentially incorporated in future versions. Finally, our empirical identification of optimal ranges for ESCHR's numeric hyperparameters was somewhat limited by the time and memory required for running these experiments with many, sometimes large, datasets

and very wide search spaces. It is therefore possible that there may be more optimal default ranges or more sophisticated regimes for hyperparameter randomization and selection that could improve ESCHR's performance.

Another key design feature of ESCHR is our soft clustering approach for generating the final consensus results. Single-cell data is inherently complex and heterogeneous, and clustering methods often make assumptions about the structure of the data that may not hold in practice. For example, hard clustering methods assume discrete groups of single cells, which rarely exist in biological data [15]. Many clustering algorithms make further assumptions about the shapes and other properties of these discrete groups. In the opposite direction, toward continuity rather than discreteness, numerous methods have been developed for trajectory inference in single-cell datasets [4], but these methods also make assumptions about dataset structure, for example, many force a branched tree structure. ESCHR's soft cluster outputs enable unified mapping of both discrete and continuous grouping structures, without the need for assumptions about the shape and properties of the dataset. To illustrate this concept, we used ESCHR to identify tanycyte subtypes and reveal the transitional continuity between them (Fig. 5a–q, Additional File 1: Fig. S10), which is notable because assumptions about lineage relationships or dynamic developmental processes in this static adult tissue would be inappropriate and could lead to inaccuracies and distortion. Instead, ESCHR can identify and characterize discrete and continuous patterns simultaneously, even in the same dataset, without relying on assumptions about data shape and properties.

One of ESCHR's most useful outputs is the per-cell uncertainty score, which enables users to estimate clustering uncertainty and interpret hard clustering results more effectively. The Impossibility Theorem for clustering states that it is impossible for any clustering method to satisfy the three proposed axioms of good clustering, and therefore all clustering algorithms must make trade-offs among the desirable features, and no clustering result can be perfect [17]. Because of this, it is critical to evaluate the guaranteed uncertainty in a clustering result before using it for direct comparisons, downstream analyses, or hypothesis generation. ESCHR uncertainty scores, which are derived from the degree of cluster overlap for each datapoint as indicated by their soft cluster assignments, provide a useful proxy for this uncertainty and difficulty in cluster assignment. These scores can be visualized alongside hard cluster assignments to facilitate a more discerning interpretation of clustering results. We have validated the utility of these uncertainty scores by demonstrating that (1) they identify areas of ground truth continuity due to cells transitioning between cell states in simulated scRNA-seq data (Fig. 2b, d), (2) they are significantly higher for inaccurately assigned data points (Fig. 2g), and (3) they are significantly negatively correlated with the level of agreement between clustering algorithms (Fig. 2h). Altogether, these findings demonstrate that ESCHR uncertainty scores provide meaningful insights into clustering uncertainty.

To make the advantages of ESCHR clustering easily accessible to the research community, we have made ESCHR available as a Python module on GitHub (https://github.com/zunderlab/eschr), packaged as an extensible software framework that is compatible with the scverse suite of single-cell analysis tools [33]. We have provided tutorials for how to incorporate it into existing single-cell analysis workflows as well as for how to use it as a standalone analysis framework. In conclusion, our results demonstrate

that ESCHR is a useful method for single-cell analysis, offering robust and reproducible clustering results with the added benefits of per-cell uncertainty scores and soft clustering outputs for improved interpretability. By emphasizing ease of adoption, clustering robustness and accuracy, generalizability across a wide variety of datasets, and improved interpretability through soft clustering outputs and the quantification of uncertainty, we aim to support the responsible and informed use of clustering results in the single-cell research community.

Methods

ESCHR Framework

ESCHR takes as input a matrix, M, with n instances (e.g., cells) as rows and d features (e.g., genes/proteins) as columns. It does not perform internal normalization or correction, so input data are expected to have already been preprocessed appropriately. ESCHR can be thought of in three primary steps: base clustering to generate the ensemble, consensus determination, and output/visualization.

Consistent with other published manuscripts in this domain, we will use the following notation. Let $X = \{x_1, x_2, ..., x_n\}$ denote a set of objects to be clustered, where each x_i is a tuple of some d-dimensional feature space for all$i = 1...n$. Let $X_s = \{x_1, x_2, ..., x_r\}$ denote a random subset of X where all of $x_1, ..., x_r$ are between 1 and$n$. $\mathbb{P} = \{P_1, P_2, ..., P_m\}$ is a set of partitions, where each $P_i = \left\{C_1^i, C_2^i, ..., C_{q_i}^i\right\}$ is a partition of an independent instantiation of X_s and contains q_i clusters. C_j^i is the j th cluster of the i th partition, for all $i = 1...m$. $t = \sum_{i=1}^{m} q_i$ is the total number of clusters from all ensemble members. Where $\mathbb{P}_X$ is the set of all possible partitions with the set of objects X and $\mathbb{P} \subset \mathbb{P}_X$, the goal of clustering ensemble methods is to find a consensus partition $P^* \epsilon \mathbb{P}_X$ which best represents the properties of each partition in$\mathbb{P}$. Additionally, the more general terminology of "instance" and "feature" will generally be used rather than domain-specific terms such as cells and genes/proteins.

Hyperparameter-randomized ensemble clustering

The ESCHR ensemble is generated with Leiden community detection as the base clustering algorithm [55]. Leiden is applied using Reichardt and Bornholdt's Potts model with a configuration null model [73]. Diversity is generated among ensemble members through a combination of data subsampling and Leiden hyperparameter randomization. The subsampling percentage varies for each ensemble member and is selected from a Gaussian distribution with the mean μ scaled to dataset size within the range 30 to 90. After subsampling a random subset X_s from X, principal components analysis (PCA) is applied to generate the most informative features for this data subspace. A default value of 30 or one less than the number of features if the number of features is less than 30 is used for the number of PCs. In the subsequent clustering step, three numerical hyperparameters are randomized for each ensemble member: (1) k, the number of neighbors for building a k-nearest neighbors (kNN) graph; (2) the choice of distance metric for building the kNN graph; and (3) r, a resolution parameter for the modularity optimization function used in Leiden community detection. The numerical hyperparameters k and r are randomly selected from within empirically established ranges (Additional File 1: Fig. S2). The distance metric is selected between either euclidean or cosine, because these

choices are efficiently implemented for fast calculation of approximate nearest neighbors (ANN) in our chosen implementation, nmslib [71]. Since each ensemble member is independent, we implemented parallelization via multiprocessing for this stage of the algorithm. Ensemble size is set at a default of 150 based on experiments demonstrating that this was sufficient to reach convergence to a stable solution (Additional File 1: Fig. S2).

Bipartite graph clustering and consensus determination

Bipartite graph clustering was used to obtain consensus clusters from the ESCHR ensemble. This approach was selected because methods that compute consensus using unipartite projection graphs of either instance or cluster pairwise relations suffer from information loss [58]. For these calculations, the biadjacency matrix is defined as: $B = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$ where A is an $n \times t$ connectivity matrix whose rows correspond to instances {1... n} and columns correspond to the ensemble clusters {1... t}. $A_{i,j}$ is an indicator that takes value 1 if instance i belongs to the j th cluster and 0 otherwise. Using this, we then create a bipartite graph$G = (V, W)$. The weights matrix $W = B$, and$V = V_1 \cup V_2$, where V_1 contains n vertices each representing an instance of the data set$X$; V_2 contains t vertices each representing a cluster of the ensemble (see Fig. 1a "Ensemble bipartite graph"). Given our bipartite graph G, we can define a community structure on G as a partition $P_1 = \{C_1, C_2, ..., C_{k_1}\}$ containing pairwise disjoint subsets ofV_1 and $P_2 = \{D_1, D_2, ..., D_{k_2}\}$ containing pairwise disjoint subsets ofV_2, such that all V_1 nodes in a specific C_i are more connected to a particular subset of V_2 than the rest of the nodes in V_1 are, and likewise (but opposite) for a given D_j ofV_2. Optimal P_1 and P_2 are computed with the Leiden algorithm for bipartite community detection with the Constant Potts Model quality function [57, 74]. This approach was designed to overcome the resolution limit of previous bipartite community detection approaches [75, 76]. There is one hyperparameter for this approach, the resolutionγ, which indirectly influences the number of clusters for P_1 and P_2 by modulating the density of connections within and between communities [74]. To avoid the need for external hyperparameter tuning, we implemented an internal hyperparameter selection strategy at this stage. First, ESCHR generates a set of potential consensus labelings across an internally-specified range of γ values. Since ARI can be used as a similarity measure between two different clustering results, ESCHR then calculates the pairwise ARI between each of the final consensus labelings generated using each different γ value. Finally, ESCHR selects the result that has the highest sum of similarity to all other results from the set of potential consensus labelings (the medoid) to return as the final consensus result. In experiments to validate this approach, we found that the number of and the memberships in the final consensus hard clusters is robust to the setting of this resolution parameter, indicating that more extensive optimization is not required (Additional File 1: Fig. S4d–e). To obtain the final consensus result, we collapse the base ensemble clusters contained in V_2 into the P_2 meta-cluster to which they were assigned. This results in each vertex of V_1 having a weighted edge to each meta-cluster equal to the sum of its edges with constituent base clusters ofV_2. The resulting weighted bipartite graph G^* therefore represents the final consensus clusteringP^*, with n vertices representing the instances, q^* vertices

representing the final consensus clusters, and weighted edges representing the membership of instance i in each of the q^* clusters ofP^*.

Hard and soft clustering outputs

Let $\Theta \epsilon \mathbb{R} n \times q^*$ be a nonnegative matrix where each row,$\Theta_i := (\Theta_{i1}, ..., \Theta_{ik_2})$, contains nonnegative numbers that sum to less than or equal to one, representing the membership of instance i in each of the q^* clusters ofP^*. Θ_{ij} is calculated by dividing the weight of the edge between instance i and consensus cluster D_j by the sum of all edge weights for instance$i$. We refer to this matrix as the soft membership matrix and to each row as the association vector v for each instance. To determine hard clustering assignments, each instance is assigned to the meta-cluster with the highest entry in its association vector$v$, with ties broken randomly. A "core" cell of a given cluster j will have $\Theta_{ij} = 1$ and zeros elsewhere, while a "transitional" instance may have up to q^* non-zero membership values. To describe the degree to which a given instance is "core" versus" transitional," we define an "uncertainty score,"Ω, for each instance as the highest membership value in its association vector ($\Omega = \max(v)$). We can additionally calculate the mean of all instance memberships in a given cluster to yield a measure of each cluster's discreteness, which we call the "cluster stability score"$s = \frac{1}{n}\Sigma_{i \epsilon n}\theta_{i,j}$.

Cluster connectivity mapping

To map the connectivity structure of clusters, we first calculate the sum-of-squares-and-cross-products matrix (SSCP) of the soft membership matrix Θ, which is calculated as $S = \Theta \prime \Theta$ and then consider $S_{i,j}$ to be an uncorrected measure for connectivity between consensus clusters i and j. To correct for connectivity that may result from random chance, we first estimate a null distribution of connectivity scores accounting for the following attributes of Θ: (1) the association vector v for a given instance are proportions and can sum to no more than 1 (with cells summing to less than one being potentially outliers) and (2) the distribution of values is not uniformly distributed and will be differently skewed for different datasets depending on overall levels of continuity or discreteness. In practice, we achieve this by independently shuffling the association vector, v, for each instance to generate a random sample. We then calculate the SSCP for 500 iterations of this randomization procedure. Using this empirical null distribution, we then calculate a p-value for each observed edge and prune edges that do not meet a default alpha value cutoff of 0.05. Thus the final corrected connectivity is defined as the ratio of the cross-product of instance memberships between a given two clusters normalized to the cross-product of instance memberships expected under constrained randomization.

Exploring soft cluster continuity for the MNIST dataset

To visualize the transition between clusters 1a and 1b (Fig. 4f) that was identified by connectivity mapping of ESCHR clustering results, we devised a simple approach for creating a one-dimensional ordering of the instances in a transitional zone based on their membership in the connected clusters of interest. Specifically, the ordering score,δ_i, of cell i having m_j membership in the cluster at position j along the cluster path of interest was calculated as: $\delta_i = \Sigma_{j=1}^{N} m_j \cdot j$, where N is the number of clusters in the path. To obtain the relevant cells of interest for ordering, we used the

following criteria: cells were included if they had (1) > 90% membership in either one of the 2 clusters of interest or (2) > 5% membership in both clusters and a combined membership of > 80% in the 2 clusters. We then visualized the progression of cluster memberships using a stacked bar plot of rank-ordered data points. The ordered data points were then partitioned into "core" datapoints and "transitional" datapoints for each cluster based on bimodality observed in the ordering scores for the cells assigned to each hard cluster.

While a linear ordering approach could in principle be used to create an ordering across a path of more than two connected clusters, it would likely only be effective in cases where connectivity mapping identifies a linear path of successively connected clusters. In cases such as the example in Fig. 4 where connectivity mapping identified a ring of 3 connected clusters (9a, 9b, 4), this approach will generally not work as well since a group of more than two clusters with nonlinear connectivity may exhibit more complex continuity structures than could be captured with a simple linear ordering. We therefore devised another method for distilling the core continuity structure for cases of greater than two clusters and nonlinear connectivity paths. We first performed principal components analysis (PCA) on the columns of the soft membership matrix Θ that correspond to the hard clusters selected for analysis, thereby capturing the primary axes of variation contained within these soft memberships. We then projected the data onto the first two PCs and used this to gain insight into the continuity structure by (1) visualizing the data points belonging to the relevant hard clusters projected into the space of these first two PCs and (2) identifying and exploring the features most highly correlated with these PCs.

Exploring soft cluster continuity for the Tany-Seq dataset

To visualize transitional zones between connected clusters in the Tany-Seq dataset [46], we applied the same one-dimensional linear ordering approach that was used to examine clusters 1a and 1b of the MNIST dataset in Fig. 4f. To identify marker features associated with the one-dimensional soft cluster transition paths, we calculated the Pearson correlation between each feature and the vector of cluster memberships for each cluster in the path. Features were then selected based on their correlation with each of the clusters individually and based on the sum of their correlations across the clusters. The three genes in Fig. 5h–j were selected from the top ten features identified through each of these methods based on their expression patterns and the availability of in situ hybridization images of sufficient quality in the Allen Mouse Brain Atlas [69]. To handle outliers for the expression heatmap UMAP plots and the expression scatterplots in Fig. 5h–j, values were thresholded to fall between the 2nd percentile and 98th percentile. The curves overlaid on the expression scatter plots in these panels were generated by first fitting B-splines with degree 3 (cubic) to the points included in the scatterplot. To generate a smoothed curve, a Gaussian kernel with a sigma of 10 was applied to the results of the spline function evaluated at 100 evenly spaced points within the range of the number of points included in the scatter plot. This is approximately equivalent to the behavior for large data sizes of the "geom_smooth" function from the R package ggplot [77].

Clustering evaluation metrics

Extrinsic evaluation metrics measure the distance of a clustering result to some external set of labels. When these labels are ground truth class labels, we can consider these to be measures of accuracy. However, they can also be used in other contexts such as with an "external" set of clustering labels. There are numerous metrics that can be used to measure this distance between a given set of predicted cluster labels and a ground truth or other set of external labels. Each of these metrics introduces some type of bias in evaluating the accuracy and robustness of clustering results, as discussed further below. To diversify these biases, we selected 2 metrics from different categories: Adjusted Rand Index (ARI) from the category of methods that employ peer-to-peer correlation and Adjusted Mutual Information (AMI) from the information theoretic measures [63]. We use both ARI and AMI to evaluate accuracy and robustness in our systematic benchmarking in Fig. 3 and in Supplementary Figures S2 and S7 (Fig. 3; Additional File 1: Figs. S2 and S7).

The ARI is the corrected-for-chance version of the Rand index, which measures the agreement between two sets of partition labels U and V [78]. The ARI is defined as:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}$$

where a is the number of pairs of two objects in the same group in both U and V; b is the number of pairs of two objects in different groups in both U and V; c is the number of pairs of two objects in the same group in U but in different groups in V; and d is the number of pairs of two objects in different groups in U but in the same group in V. Random clusterings have an expected score of zero and identical partitions have a score of 1. ARI is biased towards solutions containing (1) balanced clusters (i.e., similar size clusters within each partition) and (2) similar cluster numbers and sizes between the two partitions [13]. ARI was calculated using the implementation in sklearn (v 1.0.1).

AMI is the corrected-for-chance version of Mutual Information, which quantifies the amount of information that can be obtained about one random variable (in this application, a list of cluster labels) by observing the other random variable (another list of cluster labels) [79]. Let $C = \{C_1, C_2, ..., C_{tc}\}$ and $G = \{G_1, G_2, ..., G_{tg}\}$ be the predicted and ground truth labels on a dataset with n cells. AMI is then defined as:

$$\mathrm{AMI}(C,G) = \frac{I(C,G) - E\{I(C,G)\}}{\max\{H(C),H(G)\} - E\{I(C,G)\}}$$

Here, $I(C,G)$ represents the mutual information between C and G and is defined as:

$$I(C,G) = \sum_{p=1}^{tc}\sum_{q=1}^{tg}\left|C_p \cap G_q\right| log\frac{n\left|C_p \cap G_q\right|}{\left|C_p\right| \times \left|G_q\right|}$$

$H(C)$ And $H(G)$ are the entropies: $H(C) = -\sum_{p=1}^{tc}\left|C_p\right| log\frac{\left|C_p\right|}{n}$ and $H(G) = -\sum_{p=1}^{tg}\left|G_p\right|\log\frac{\left|G_p\right|}{n}$. $E\{I(C,G)\}$ is the expected mutual information between two random clusters. Random clusterings have an expected score of zero and identical partitions have a score of 1. AMI is biased towards solutions containing pure clusters, with a "pure cluster" being defined as a cluster in one set of labels that contains instances

from only one cluster of the other set of labels to which it is being compared [13]. AMI was calculated using the implementation in sklearn (v 1.0.1).

Systematic benchmarking

For benchmarking ESCHR, we selected the following clustering algorithms for comparison: (1&2) K-means and agglomerative hierarchical clustering (from scikit-learn version 1.0.1) [80], (3) SC3 (version 1.10.1 from Bioconductor) [24], (4) SC3s (version 0.1.1 through Scanpy) [25], (5) Seurat (version 4.1.1 from CRAN) [40], (6) SINCERA (version 1.0 from https://github.com/xu-lab/SINCERA) [39], (7) Scanpy (version 1.8.2 from Anaconda) [35], (8) Phenograph (version 1.5.7) [41], (9) scCAN (version 1.0 from https://github.com/bangtran365/scCAN) [38], and (10) scAIDE (version 1.0 https://github.com/tinglabs/scAIDE) [37].

Clustering algorithms were excluded from our benchmarking comparison if they did not meet the following selection criteria: (1) software freely available; (2) code publicly available; (3) can run on multiple data modalities (e.g. not scRNA-seq-specific); (4) no unresolved errors during install or implementation; (5) does not require additional user input during the algorithm (other than prior information); and (6) able to complete analysis of datasets with $\geq$ 100,000 data points and 2000 features.

For the included methods, we followed the instructions and tutorials provided by the authors of each software package. For the K-means, SC3s, and Agglomerative methods, which require pre-specification of cluster number, we calculated distortion scores over a range of cluster numbers for K-means clustering and used the elbow method to select the optimal cluster number for use across all three methods. Default values were used for all other hyperparameters for each tool, as is common practice for most realistic use cases [6, 81]. ESCHR was also run with all default settings, which is the intended usage. For all benchmarking analyses, the memory was set to 100 GB of RAM on the University of Virginia (UVA) Rivanna High Performance Computing (HPC) cluster.

No random seeds were intentionally fixed, but from inspecting the respective codebases, we believe it is likely that there remained internally fixed random seeds for some functions within some of the tested methods. Many common methods have internally fixed random seeds and/or default hyperparameters with fixed random seeds. This practice may mask a lack of robustness of these methods, and should only legitimately serve to replicate exact analyses when that is desired by the end user.

To assess the accuracy of methods in clustering our synthetic and image datasets, which have ground truth labels, we used the two extrinsic evaluation metrics defined above (ARI, AMI). For these purposes, each of the five independent runs of a given method was scored against the ground truth labels. Since it is nearly universal in papers describing new single-cell analysis methods, we also applied this analysis to evaluate the accuracy of each of the methods for our collection of "real" datasets using available published labels. However, we stress that we do not think this is a reliable or effective measure for evaluating clustering methods, as we detail further in Supplementary Note 1 (Additional File 1: Supplementary Note 1).

We also used the extrinsic metrics ARI and AMI to evaluate the stability and reproducibility of hard clustering results. In line with standard practice for benchmarking stability/robustness [82], we performed repeated runs with 5 random subsamples (90%) of

each dataset for every method. This simulates slight differences in data collection and/or preprocessing, and if the clustering is capturing a true underlying structure rather than overfitting to noise it should be detected regardless of the exact set of cells that are sampled for the analysis. We then calculated pairwise scores for each metric between each of the 5 independent runs of a respective method and then took the mean across replicate pairs to obtain the final score per dataset-method.

To calculate both the method and replicate "agreement scores" for comparison with ESCHR uncertainty scores (Fig. 2j), we first constructed contingency matrices between all pairs of replicates and methods and mapped the cluster labels from the result with more clusters to the result with fewer clusters. Using the shared labels between a given pair of clustering results we could then calculate per-instance agreement (binary) within the pair of results. The final per-instance score was calculated as the mean agreement across all possible combinations.

Statistical analyses

Statistical comparisons were performed using the "scipy.stats" and "statannotations" Python packages [83, 84]. The two-sided Wilcoxon signed-rank test with Bonferroni correction was used to compare the performance of ESCHR versus each alternative method in the systematic benchmarking panels shown in Fig. 3. Comparisons were calculated using dataset means across replicates for all tested datasets. $N=21$ for comparisons using synthetic datasets and $N=45$ for comparisons using real datasets. The two-sided Mann–Whitney-Wilcoxon was used for comparing uncertainty scores between accurately and inaccurately assigned cells in Fig. 2g, with $N=126{,}545$ and $N=750{,}955$ for inaccurate and accurate groups respectively. The resulting p-value was below the threshold of calculation in a standard Python computing environment and was reported as zero, so we have reported this as $p<0.00001$ in the figure.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03386-5.

Additional file 1. All supplementary notes and figures, tables S5-S6 [91, 92].

Additional file 2. Tables S1-S4.

Additional file 3. Review history

Acknowledgements
The authors acknowledge Research Computing at The University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication. URL: https://rc.virginia.edu. The authors would also like to thank Kristen Naegle and Sean Chadwick for critical feedback on the manuscript.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 3.

Authors' contributions
S.M.G and E.R.Z. conceptualized the ESCHR clustering approach. S.M.G. developed the ESCHR method and Python package. S.M.G. and E.R.Z. planned all analysis and benchmarking experiments. S.M.G. performed all analysis and benchmarking experiments, and prepared figures. S.M.G. and E.R.Z. wrote the manuscript.

Authors' X handles
X handles: @Sarah_M_Goggin (Sarah M. Goggin), @zunderlab (Eli R. Zunder).

Funding
Research reported in this publication was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award number R01NS111220 to E.R.Z. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. S.M.G. was further supported by the Transdisciplinary Big Data Science Training Grant, award number T32LM012416.

Availability of data and materials
All "real" datasets are publicly available. The 45 datasets used in our study are described in Supplementary Table 2 (Additional File 2: Table S2), including information and links to download preprocessed datasets. The only processing step we performed was log2 transformation for scRNA-seq datasets and arcsinH transformation for mass cytometry datasets if a given dataset was not yet scaled. The MNIST dataset was downloaded from keras datasets [85] and was preprocessed as recommended by the accompanying documentation.
Synthetic Gaussian datasets were generated using sklearn "make_blobs" [80] using various combinations of object number, feature number, cluster number, cluster size, and cluster standard deviations. Anisotropic transformations were then applied to the resulting datasets by multiplying pairs of features (an $n \times 2$ subset of the full data matrix) by different 2×2 matrices filled with random values between −2 and 2. These datasets are available at: https://doi.org/https://doi.org/10.5281/zenodo.12746558 [86].
Simulated scRNA-seq datasets with 1000 "cells" and 1000 "genes" were generated using the DynToy package, which simulates different complex trajectory models based on real single-cell gene expression data [62]. These datasets are available at: https://doi.org/10.5281/zenodo.12786322 [87].
ISH images were downloaded from the ABA portal (http://mouse.brain-map.org) [88] and are freely available.
ESCHR is available under the MIT License and can be installed via PyPi (https://pypi.org/project/eschr/) [89] or GitHub (https://github.com/zunderlab/eschr) [32]. The source code of the ESCHR version used for making the figures in this manuscript is available on zenodo (https://doi.org/10.5281/zenodo.13380410) [90]. ESCHR is compatible with standard Python single-cell data structures and can be easily incorporated into scverse workflows or used as a standalone framework.
Further information and requests for resources should be directed to and will be fulfilled by Eli Zunder (ezunder@virginia.edu).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 1 February 2024 Accepted: 30 August 2024
Published online: 16 September 2024

References

1. Svensson V, da Veiga Beltrame E, Pachter L. A curated database reveals trends in single-cell transcriptomics. Database (Oxford). 2020;2020:baaa073.
2. Xie B, Jiang Q, Mora A, Li X. Automatic cell type identification methods for single-cell RNA sequencing. Comput Struct Biotechnol J. 2021;19:5874–87.
3. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20:273–82.
4. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37:547–54.
5. Zappia L, Theis FJ. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. Genome Biol. 2021;22:301.
6. Schneider I, Cepela J, Shetty M, Wang J, Nelson AC, Winterhoff B, et al. Use of "default" parameter settings when analyzing single cell RNA sequencing data using Seurat: a biologist's perspective. JTGG. 2020;5:37–49.
7. Lu Y, Phillips CA, Langston MA. A robustness metric for biological data clustering algorithms. BMC Bioinformatics. 2019;20(Suppl 15):503.
8. Krzak M, Raykov Y, Boukouvalas A, Cutillo L, Angelini C. Benchmark and parameter sensitivity analysis of single-cell RNA sequencing clustering methods. Front Genet. 2019;10:1253.
9. Tang M, Kaymaz Y, Logeman BL, Eichhorn S, Liang ZS, Dulac C, et al. Evaluating single-cell cluster stability using the Jaccard similarity index. Bioinformatics. 2021;37:2212–4.
10. Gibson G. Perspectives on rigor and reproducibility in single cell genomics. PLoS Genet. 2022;18:e1010210.
11. Patterson-Cross RB, Levine AJ, Menon V. Selecting single cell clustering parameter values using subsampling-based robustness metrics. BMC Bioinformatics. 2021;22:39.
12. Renedo-Mirambell M, Arratia A. Identifying bias in network clustering quality metrics. PeerJ Comput Sci. 2023;9:e1523.

13. Amigó E, Gonzalo J, Artiles J, Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf Retr Boston. 2009;12:461–86.
14. Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. [version 2; peer review: 3 approved]. F1000Res. 2018;7:1297.
15. Cembrowski MS, Menon V. Continuous Variation within Cell Types of the Nervous System. Trends Neurosci. 2018;41:337–48.
16. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature. 2017;541:331–8.
17. Kleinberg J. An Impossibility Theorem for Clustering. Advances in neural information processing systems. 2002;15.
18. Huh R, Yang Y, Jiang Y, Shen Y, Li Y. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. Nucleic Acids Res. 2020;48:86–95.
19. Burton RJ, Cuff SM, Morgan MP, Artemiou A, Eberl M. GeoWaVe: geometric median clustering with weighted voting for ensemble clustering of cytometry data. Bioinformatics. 2023;39:btac751.
20. Tsoucas D, Yuan G-C. GiniClust2: a cluster-aware, weighted ensemble clustering method for cell-type detection. Genome Biol. 2018;19:58.
21. Sagi O, Rokach L. Ensemble learning: A survey. WIREs Data Mining Knowl Discov. 2018;8:e1249.
22. Wan S, Kim J, Won KJ. SHARP: hyperfast and accurate processing of single-cell RNA-seq data via ensemble random projection. Genome Res. 2020;30:205–13.
23. Risso D, Purvis L, Fletcher RB, Das D, Ngai J, Dudoit S, et al. clusterExperiment and RSEC: s Bioconductor package and framework for clustering of single-cell and other large gene expression datasets. PLoS Comput Biol. 2018;14:e1006378.
24. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14:483–6.
25. Quah FX, Hemberg M. SC3s: efficient scaling of single cell consensus clustering to millions of cells. BMC Bioinformatics. 2022;23:536.
26. Zhu L, Lei J, Klei L, Devlin B, Roeder K. Semisoft clustering of single-cell data. Proc Natl Acad Sci USA. 2019;116:466–71.
27. Peters G, Crespo F, Lingras P, Weber R. Soft clustering – Fuzzy and rough approaches and their extensions and derivatives. Int J Approximate Reasoning. 2013;54:307–22.
28. Kanter I, Dalerba P, Kalisky T. A cluster robustness score for identifying cell subpopulations in single cell gene expression datasets from heterogeneous tissues and tumors. Bioinformatics. 2019;35:962–71.
29. Chen Z, Goldwasser J, Tuckman P, Liu J, Zhang J, Gerstein M. Forest Fire Clustering for single-cell sequencing combines iterative label propagation with parallelized Monte Carlo simulations. Nat Commun. 2022;13:3538.
30. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016;19:335–46.
31. Karim MR, Beyan O, Zappa A, Costa IG, Rebholz-Schuhmann D, Cochez M, et al. Deep learning-based clustering approaches for bioinformatics. Brief Bioinformatics. 2021;22:393–415.
32. Goggin SM, Zunder ER. ESCHR. Computer software. Github; 2024. https://github.com/zunderlab/eschr.
33. Virshup I, Bredikhin D, Heumos L, Palla G, Sturm G, Gayoso A, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat Biotechnol. 2023;41:604–6.
34. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177:1888-1902.e21.
35. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15.
36. Guo M, Xu Y. Single-Cell Transcriptome Analysis Using SINCERA Pipeline. Methods Mol Biol. 2018;1751:209–22.
37. Xie K, Huang Y, Zeng F, Liu Z, Chen T. scAIDE: clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types. NAR Genom Bioinform. 2020;2:lqaa082.
38. Tran B, Tran D, Nguyen H, Ro S, Nguyen T. scCAN: single-cell clustering using autoencoder and network fusion. Sci Rep. 2022;12:10267.
39. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. PLoS Comput Biol. 2015;11:e1004575.
40. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.
41. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015;162:184–97.
42. The MNIST Database of Handwritten Digits. https://yann.lecun.com/exdb/mnist/. Accessed 11 Feb 2023.
43. Brunner M, Lopez-Rodriguez D, Messina A, Thorens B, Santoni F, Langlet F. Pseudospatial transcriptional gradient analysis of hypothalamic ependymal cells: towards a new tanycyte classification. BioRxiv. 2023.
44. Campbell JN, Macosko EZ, Fenselau H, Pers TH, Lyubetskaya A, Tenen D, et al. A molecular census of arcuate hypothalamus and median eminence cell types. Nat Neurosci. 2017;20:484–96.
45. Langlet F. Tanycyte gene expression dynamics in the regulation of energy homeostasis. Front Endocrinol (Lausanne). 2019;10:286.
46. Sullivan AI, Potthoff MJ, Flippo KH. Tany-Seq: Integrated Analysis of the Mouse Tanycyte Transcriptome. Cells. 2022;11:1565.
47. Fong H, Kurrasch DM. Developmental and functional relationships between hypothalamic tanycytes and embryonic radial glia. Front Neurosci. 2022;16:1129414.
48. Dietterich TG. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. Berlin. Heidelberg: Springer Berlin Heidelberg; 2000. p. 1–15.
49. Ghosh J, Acharya A. Cluster ensembles. WIREs Data Mining Knowl Discov. 2011;1:305–15.
50. Strehl A, Ghosh J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. JMLR (J Mach Learn Res). 2002;3:583–617.
51. Ben-Hur A, Elisseeff A, Guyon I. A stability based method for discovering structure in clustered data. Pac Symp Biocomput. 2002:6–17.

52. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52:91–118.
53. Fred A. Finding consistent clusters in data partitions. In: Kittler J, Roli F, editors. Multiple Classifier Systems. Springer, Berlin Heidelberg: Berlin, Heidelberg; 2001. p. 309–18.
54. Naegle KM, Welsch RE, Yaffe MB, White FM, Lauffenburger DA. MCAM: multiple clustering analysis methodology for deriving hypotheses and insights from high-throughput proteomic datasets. PLoS Comput Biol. 2011;7:e1002119.
55. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9:5233.
56. Topchy A, Jain AK, Punch W. Combining multiple weak clusterings. In: Third IEEE International Conference on Data Mining. Melbourne, FL: IEEE Comput. Soc; 2003. p. 331–8.
57. Traag VA. leidenalg Documentation. Release 0102.dev9+gdc8ec1a.d20230927. Section 4.1.2 Bipartite:15–6.
58. Fern XZ, Brodley CE. Solving cluster ensemble problems by bipartite graph partitioning. In: Twenty-first international conference on Machine learning - ICML '04. New York: ACM Press; 2004. p. 36.
59. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10:5416.
60. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. JOSS. 2018;3:861.
61. van der Maaten L, Hinton G. Visualizing Data using t-SNE. JMLR. 2008;9:2579–605.
62. Cannoodt R, Saelens W, Deconinck L, Saeys Y. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. Nat Commun. 2021;12:3942.
63. Palacio-Niño J-O, Berzal F. Evaluation Metrics for Unsupervised Learning Algorithms. arXiv. 2019.
64. Rodríguez EM, Blázquez JL, Pastor FE, Peláez B, Peña P, Peruzzo B, et al. Hypothalamic tanycytes: a key component of brain-endocrine interaction. Int Rev Cytol. 2005;247:89–164.
65. Langlet F. Targeting Tanycytes: Balance between Efficiency and Specificity. Neuroendocrinology. 2020;110:574–81.
66. Yoo S, Kim J, Lyu P, Hoang TV, Ma A, Trinh V, et al. Control of neurogenic competence in mammalian hypothalamic tanycytes. Sci Adv. 2021;7:eabg3777.
67. Chen R, Wu X, Jiang L, Zhang Y. Single-cell RNA-Seq reveals hypothalamic cell diversity. Cell Rep. 2017;18:3227–41.
68. Deng G, Morselli LL, Wagner VA, Balapattabi K, Sapouckey SA, Knudtson KL, et al. Single-Nucleus RNA sequencing of the hypothalamic arcuate nucleus of C57BL/6J Mice After Prolonged Diet-Induced Obesity. Hypertension. 2020;76:589–97.
69. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445:168–76.
70. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics. 2015;31:1974–80.
71. Boytsov L, Naidan B. Engineering Efficient and Effective Non-metric Space Library. In: Brisaboa N, Pedreira O, Zezula P, editors. Similarity search and applications. Springer, Berlin Heidelberg: Berlin, Heidelberg; 2013. p. 280–93.
72. Watson ER, Mora A, Taherian Fard A, Mar JC. How does the structure of data impact cell-cell similarity? Evaluating how structural properties influence the performance of proximity metrics in single cell RNA-seq data. Brief Bioinformatics. 2022;23:bbac387.
73. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E. 2006;74:016110.
74. Traag VA, Van Dooren P, Nesterov Y. Narrow scope for resolution-limit-free community detection. Phys Rev E. 2011;84:016114.
75. Calderer G, Kuijjer ML. Community detection in large-scale bipartite biological networks. Front Genet. 2021;12:649440.
76. Xu Y, Chen L, Li B, liu W. Density-based modularity for evaluating community structure in bipartite networks. Inf Sci (Ny). 2015;317:278–94.
77. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. Cham: Springer; 2016.
78. Hubert L, Arabie P. Comparing partitions J of Classification. 1985;2:193–218.
79. Vinh N, Epps J, Bailey J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. J Mach Learn Res. 2010;11:2837–54.
80. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
81. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, da Costa L F, et al. Clustering algorithms: a comparative approach. PLoS One. 2019;14:e0210236.
82. Weber LM, Saelens W, Cannoodt R, Soneson C, Hapfelmeier A, Gardner PP, et al. Essential guidelines for computational method benchmarking. Genome Biol. 2019;20:125.
83. Jones E, Oliphant T, Peterson P, others. SciPy: Open source scientific tools for Python. Version 1.31. Computer software. 2024. https://github.com/scipy/scipy/tree/v1.13.0.
84. Charlier F, Weber M, Izak D, Harkin E, Magnus M, Lalli J, et al. Statannotations. Version 0.6. Computer software. 2023. https://github.com/trevismd/statannotations/tree/v0.6.0.
85. Chollet F, et al. Keras. Version 3.0. Computer software. 2023. https://keras.io/api/datasets/mnist.
86. Goggin S. Synthetic gaussian datasets. Zenodo. 2024. https://doi.org/10.5281/zenodo.12746558
87. Goggin S. DynToy simulated scRNA-seq datasets. Zenodo. 2024. https://doi.org/10.5281/zenodo.12786322
88. Allen Mouse Brain Atlas. https://mouse.brain-map.org/. Accessed 2 Mar 2023.
89. Goggin SM, Zunder ER. ESCHR. Computer software. PyPi. 2024. https://pypi.org/project/eschr/.
90. Goggin S. ESCHR v0.2.0. Zenodo. 2024. https://doi.org/10.5281/zenodo.13380410.
91. Yanai I, Lercher M. A hypothesis is a liability. Genome Biol. 2020;21:231.
92. Tyler SR, Bunyavanich S, Schadt EE. PMD Uncovers Widespread Cell-State Erasure by scRNAseq Batch Correction Methods. BioRxiv. 2021.

Publisher's Note