ICAT: A Novel Algorithm to Robustly Identify Cell States Following Perturbations in Single Cell Transcriptomes

Dakota Y. Hawkins[1,2], Daniel T. Zuch[2,3], James Huth[2], Nahomie Rodriguez-Sastre[2], Kelley R. McCutcheon[2], Abigail Glick[2,3], Alexandra T. Lion[2,3], Christopher F. Thomas[2], Abigail E. Descoteaux[2,3,5], W. Evan Johnson[1,6], and Cynthia A. Bradham[1,2,3,5*]

[1]Bioinformatics Program, [2]Biology Department, [3]Program in Molecular and Cellular Biology and Biochemistry, and [5]Biological Design Center, Boston University, Boston, MA, USA

[6]Division of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA

*author for correspondence; e-mail: cbradham@bu.edu

Author e-mails: DYH, dyh0110@bu.edu; DTZ, dtzuch@bu.edu; JH, jhuth@bu.edu; NRS, chantal1@bu.edu; KRM, kmccutch@bu.edu; AG, aglick18@bu.edu; CFT: cfthomas@bu.edu; AED: abigaild@bu.edu; WEJ, wej@bu.edu

KEY WORDS: single cell, scRNA-seq, clustering, cell states, perturbation, cell type, semi-supervised, self-supervised

Abstract

Motivation

The detection of distinct cellular identities is central to the analysis of single-cell RNA sequencing experiments. However, in perturbation experiments, current methods typically fail to correctly match cell states between conditions or erroneously remove population substructure. Here we present the novel, unsupervised algorithm ICAT that employs self-supervised feature weighting and semi-supervised clustering to accurately resolve cell states across heterogeneous conditions.

Results

Using simulated and real datasets, we show ICAT is superior in identifying and resolving cell states compared to current integration workflows. While requiring no a priori knowledge of extant cell states or discriminatory marker genes, ICAT is robust to low signal strength, high perturbation severity, and disparate cell type proportions. We empirically validate ICAT in a developmental model and find that only ICAT identifies a perturbation-unique cellular response. Taken together, our results demonstrate that ICAT offers a significant improvement in defining cellular responses to perturbation in single-cell RNA sequencing data.

Availability and implementation

https://github.com/BradhamLab/icat

Introduction

From deconstructing tumor cell compositions (Schelker *et al.*, 2017), to reconstructing developmental pathways (Wagner *et al.*, 2018), single-cell RNA sequencing (scRNA-seq) has revolutionized scientists' ability to explore complex tissues and systems. Clustering of individual cells is central to the analysis of scRNA-seq data, in which cell states are identified by grouping cells with similar gene expression profiles (Luecken and Theis, 2019). Whether to ascertain the effect of drug treatment on cancer viability, or to infer the mechanistic role of a given gene in tissue specification, perturbation experiments have long been used to dissect and understand complex biological systems. Oftentimes this analysis is formalized by identifying changes in the abundance of cell identities between treatments (Luecken and Theis, 2019; Kang *et al.*, 2017; Haber *et al.*, 2017). Inherently, such analyses are based on the assumption that cell states are readily identified and matched between treatments. However, matching identities across perturbing treatments presents particular issues with scRNA-seq datasets due to dominating treatment effects (Fig. S1), where it is typical that scRNA-seq datasets segregate by treatments rather than cellular identities (Fig. S1) (Luecken and Theis, 2019; Kang *et al.*, 2017; Lähnemann *et al.*, 2020; Kagohara *et al.*, 2020). This challenge presents significant difficulties in correctly evaluating cellular responses to perturbation in scRNA-seq experiments.

Common techniques for identifying cell states in scRNA-seq perturbation experiments rely on either reference datasets to map cells onto known cell identities (Regev *et al.*, 2017; Schaum *et al.*, 2018), or an integration algorithm to minimize differences between control and perturbed cells by transforming cells into a shared space prior to clustering cells (Luecken and Theis, 2019; Perillo *et al.*, 2020). The first scenario is problematic for non-model systems, where such atlases may not exist, and is inappropriate when the goal of a given experiment is discovery of novel cell identities or subtypes; for these reasons, the latter approach is often preferred. Toward that end, scRNA-seq algorithms have been developed that

broadly integrate heterogeneous datasets (Luecken and Theis, 2019; Stuart and Satija, 2019; Hie *et al.*, 2019). Briefly, integration algorithms identify a shared space between separate datasets, then subsequently project cells into this embedding. While promising for combining datasets across modalities, integration methods can smooth over and eliminate biological signals, leading to incorrect identification and possibly wholesale erasure of captured cell states (Luecken and Theis, 2019; Luecken *et al.*, 2020; Büttner *et al.*, 2019; Tyler *et al.*, 2021).

We therefore developed a two-step algorithm to robustly and sensitively **I**dentify **C**ell states **A**cross **T**reatments (ICAT) in scRNA-seq data. By employing self-supervised feature weighting, followed by semi-supervised clustering, ICAT accurately identifies cell states across scRNA-seq perturbation experiments with high accuracy. ICAT does not require prior knowledge of marker genes or extant cell states, is robust to perturbation severity, and identifies cell states with higher accuracy than leading integration workflows within both simulated and real scRNA-seq perturbation experiments, while operating at lower signal resolutions.

Materials and methods

Algorithm overview

To overcome the inability of current clustering algorithms to accurately detect cell identities in perturbation experiments (Fig. S1), we developed a two-step algorithm (Fig. 1A). Using the Louvain method (Blondel *et al.*, 2008) by default, the first step separately clusters control cells to produce initial cluster labels for each cell. Cluster labels are then used to weight genes by their ability to predict the previously generated labels using Neighborhood Component Feature Selection (NCFS) (Yang *et al.*, 2012). NCFS weights genes by using gradient ascent to maximize a regularized leave-one-out classification accuracy metric. By the nature of regularization, most gene weights converge to zero, leaving only the most predictive genes with weights greater than 1. Applying the learned gene weights to the original expression matrix with both control and treated cells transforms the matrix into a "cluster-

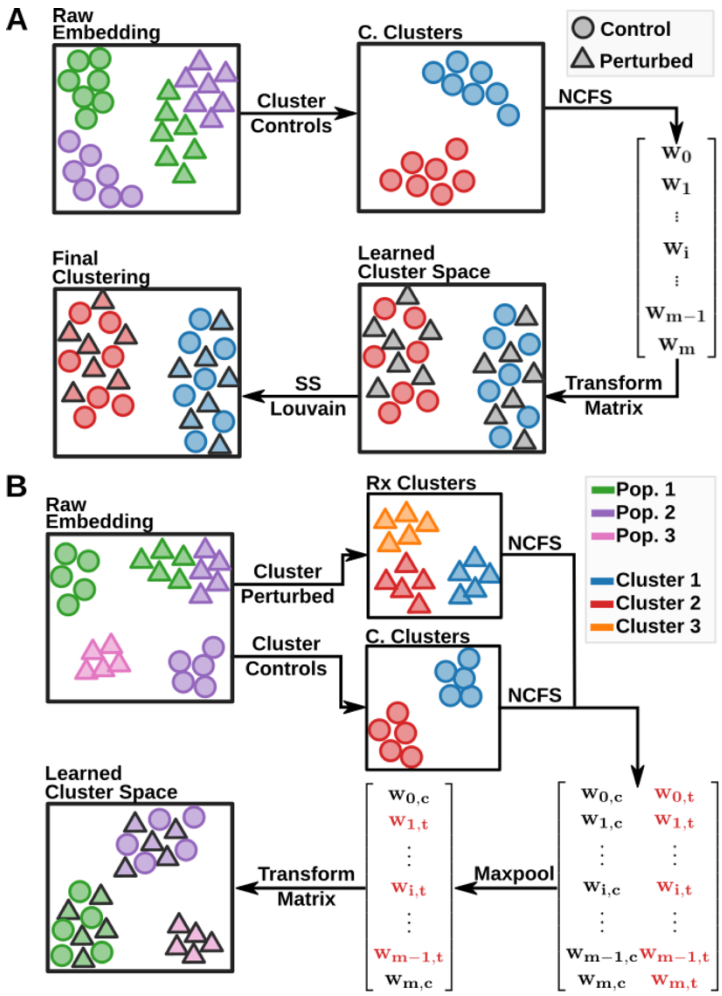


Figure 1. Overview of the ICAT algorithm. A. The schematic illustrates the ICAT$_C$ implementation of ICAT. To identify cell states across treatments, ICAT first performs self-supervised feature weighting to find genes that discriminate cell identities among control cells alone, followed by semi-supervised clustering using the newly transformed expression matrix. To learn feature weights, ICAT clusters control cells using Louvain community detection, then uses the newly generated cluster labels as input into NCFS to weight genes by their predictiveness. After applying the learned gene weights to the original gene expression matrix, ICAT clusters both treated and control cells together using a semi-supervised implementation of Louvain community detection. During this process, ICAT holds the previously identified cluster labels for the control cells immutable. **B.** The schematic illustrates the ICAT$_{C+T}$ implementation, which expands feature weighting to treated cells to identify asymmetrical populations between treatments. Cells are split along treatments and independently clustered using the Louvain method, then cluster labels are used to learn gene weights using NCFS in each treatment set independently. To retain asymmetrically informative genes, weights for each treatment are concatenated row-wise and subsequently reduced to the maximum weight using a row-wise maxpool. The reduced weight vector is then used to transform the original count matrix.

defined" space, where the distances between cells are dominated by highly weighted marker genes. After this transformation, ICAT retains previously identified cluster labels for control cells, then performs semi-supervised Louvain community detection (Blondel *et al.*, 2008; Traag, 2021), whereby control labels are held immutable (Fig. 1A).

This strategy achieves two goals: 1. ICAT makes comparing treatment effects on cell states straightforward. By implementing semi-supervised clustering with immutable control labels, ICAT ensures identities discovered in control cells are retained when treated cells are introduced. 2. ICAT produces easily interpretable embeddings. Because NCFS is a sparse method, it overtly separates informative and non-informative genes. Thus, the expression patterns of a few highly weighted genes directly account for cell location in

the NCFS-transformed space. This is a noted benefit over typical integration methods, where transformations are obscured behind neighbor-based approaches (Stuart and Satija, 2019; Hie *et al.*, 2019). These two features together produce interpretable results that ease compositional analysis between treatments.

While similar in approach, ICAT provides several advantages compared to identifying shared cell states via marker gene mapping: first, ICAT does not require previously known marker genes, making ICAT appropriate for situations with previously uncharacterized cell states. Second, by clustering control and treated cells together, ICAT takes advantage of the structure in both datasets compared to clustering both independently. This alleviates issues with imperfect matches between marker genes. Third, by foregoing common marker gene presence-absence comparisons to map cell types, ICAT accurately identifies activated states, which is otherwise difficult. Finally, by generating cell state labels through a clustering step rather than a classification approach, ICAT is able to group cells with treatment-specific phenotypes and identify out-of-sample states.

One caveat of our approach is that if asymmetrical populations exist in treated cells, but not controls, their identifying genes will likely be collapsed by the control-only feature-weighting step. To avoid this artifact, ICAT can be optionally enabled to perform feature-weighting in all treatment types to identify asymmetrical populations across conditions (Fig. 1B, Fig. 2A). To reduce gene weights to a single gene-weight vector, only the maximum gene weight produced across the conditions is retained. During semi-supervised clustering, labels for control cells are retained during label initialization, while treated cells are re-initialized as singleton clusters.

Data simulation

To assess ICAT, we initially tested performance in simulated single-cell RNAseq data (Fig. 2A-E). We simulated data using a zero-inflated Negative Binomial distribution to model gene counts per Büttner and colleagues (Büttner *et al.*, 2019). To simulate perturbations, we sampled a scalar multiplier from a $\Gamma(2,2)$ distribution, then applied this multiplier to randomly selected genes as perturbation targets. To generate distinct cell types, we randomly selected a separate set of genes to act as markers for each

population. Selected marker genes had their average expression shifted by a scalar multiplier following a $\Gamma(3,3)$ distribution to create population substructure. To test ICAT's ability to detect both asymmetrical populations and activated states, we simulated five populations of cells: three that are present in both control and treated cells, with one of these exhibiting elevated marker gene expression upon treatment and referred to as "stimulated" (population C1 and P(C1)+), and two treatment-only populations (populations P4 and P5) (Fig. 2A-B).

Using simulated data, we first assessed ICAT's robustness against increasing perturbation severity, testing five conditions with varying proportions of perturbed genes, ranging from 1% to 25% (Fig. 2C, Table S1). Next, we assessed performance with changing resolutions between cell types by simulating 10 different experimental conditions with a varying number of average marker genes per cell type. Mean marker gene numbers ranged from 10-105 per population (0.67-7.0% of total genes) (Fig. 2D, Table S2). In both cases, each experiment was simulated 15 times, for a total of 90 independent datasets in the first case (Fig. 2C) and 150 in the second case (Fig. 2D). Finally, to test performance as cell type proportions deviate from equality, we simulated three different experimental conditions with varying distributions of each cell type. Experiments ranged from all cell types having 100 cells per treatment, to the most disparate where the minimum and maximum population size per treatment were 50 and 150, respectively (Fig. 2E, Table S3). To isolate the effect of differential abundance, proportion simulations did not include an activated cell type, and instead were composed of three shared populations and two treatment-unique populations (Fig. 2E). Each experimental proportion was simulated 15 times.

When assessing ICAT's performance in simulated data, we considered three criteria reflecting "good" clustering between control and treated cells: 1. Global label conservation: the extent of mapping between cluster labels and known cell identities, which is ideally one-to-one to reflect accurate detection of cell identities, was assessed with the Adjusted Rand Index (Hubert and Arabie, 1985) (ARI). 2. Treatment mixing: the local homogeneity of shared cell types without undo isolation by treatment status was assessed with the Local Inverse Simpson's Index (Korsunsky *et al.*, 2019) (LISI); 3. Unique label separation: the accurate detection of populations-of-interest was measured using population-specific F1

measures (Luecken *et al.*, 2020). Such metrics were used to measure the ability to accurately detect asymmetric populations that exist only in single conditions (F1-Unique), the ability to discriminate stimulated cell states from non-stimulated analogues (F1-Stim), and the identification of rare cell types (F1-Rare). ARI, LISI, and all F1 metrics were standardized between 0 and 1, where 1 indicated stronger performance.

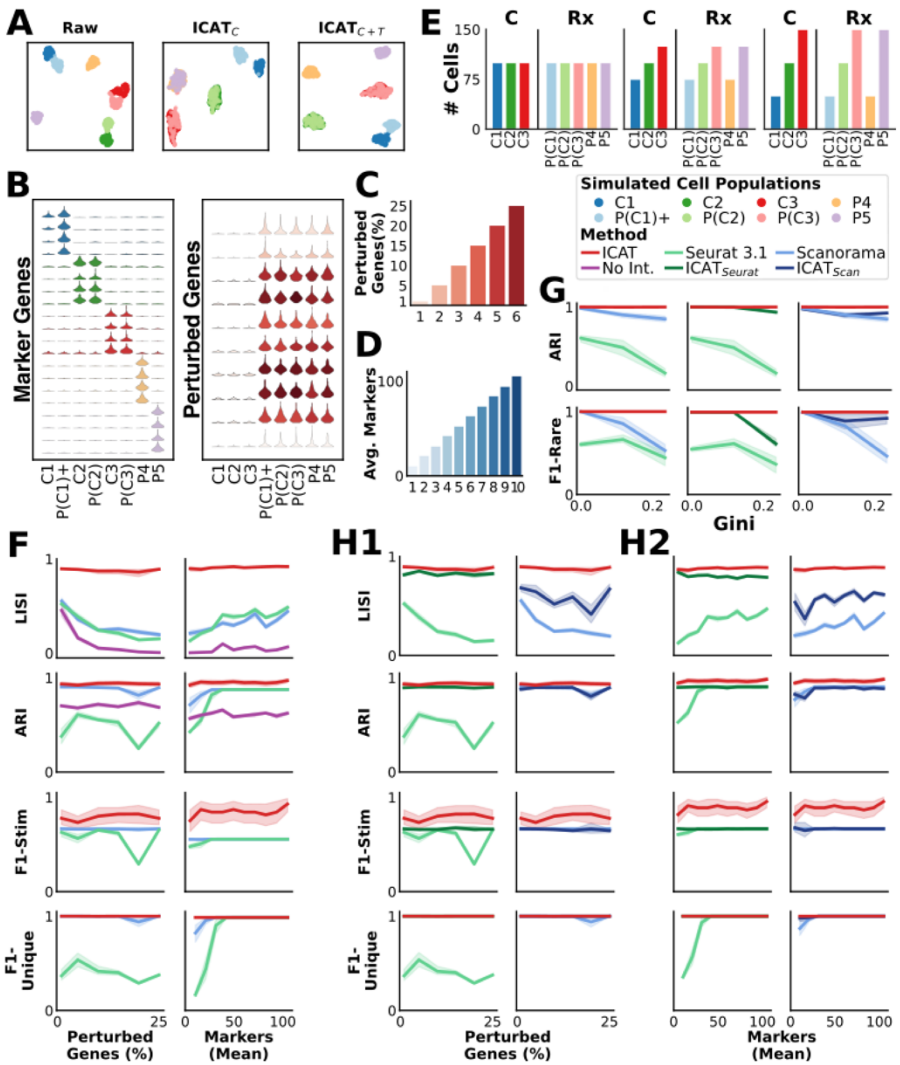
Please see Supplemental Methods for additional details.

Results

ICAT is robust to perturbation severity and better resolves cell states at lower resolutions than leading integration workflows.

We compared ICAT against two state-of-the-art integration methods: Seurat (Stuart and Satija, 2019; Satija *et al.*, 2015) and Scanorama (Hie *et al.*, 2019). Seurat was chosen for its ubiquity in single-cell analysis, while Scanorama was chosen due to its strong performance in benchmarks comparisons (Luecken *et al.*, 2020). In both cases, datasets were integrated between treatments, then the results were clustered via Louvain community detection. For appropriate comparisons between methods, we standardized nearest neighbor and resolution parameters for each method during clustering. To establish baseline performance, we also tested the "No Integration" scenario in which cells were clustered using Louvain community detection without accounting for treatment status.

When testing perturbation severity or variable marker gene numbers (Fig. 2C-D), ICAT outperformed both Scanorama and Seurat across all metrics while exhibiting highly stable performance across both increasing perturbation intensities and at varying cell-state resolutions (Fig. 2F). ICAT provides near-perfect treatment mixing, with LISI scores nearing 1 for all signal levels. In contrast, Scanorama and Seurat only approach 0.5 LISI scores at the highest signal levels, indicating that ICAT exhibits a substantially improved ability to correctly and robustly mix cell identities across treatments (Fig. 2F).



Scanorama offered comparable performance to ICAT in global label conservation (ARI) and the ability to detect asymmetrical populations (F1-Unique) (Fig. 2F), while ICAT was uniformly better at accurately distinguishing the stimulated population from its non-stimulated analogue (Fig. 2F, F1-Stim). Seurat underperformed both Scanorama and

Figure 2. ICAT outperforms current methods for cell state identification and is robust to experimental conditions. A. UMAP projections of raw (left), ICAT$_C$ processed (middle), and ICAT$_{C+T}$ processed count matrices for the simulated data. Projections show ICAT$_C$ and ICAT$_{C+T}$ correctly mix shared populations (red and green), while only ICAT$_{C+T}$ isolates asymmetrical populations (purple, yellow). ICAT performance for simulated data was further evaluated using the ICAT$_{C+T}$ implementation only. **B.** Select marker and perturbed gene expression patterns are displayed as violin plots for three simulated control cell types (1-3) under normal (C) and perturbed (P) conditions. P(C1)+ is a stimulated and perturbed version of cell type 1; perturbation-specific cell types P4 and P5 express distinct marker genes. **C.** The percentage of perturbed genes used to assess robustness to perturbation severity is shown; values range from 1-25%. **D.** The average number of marker genes per cell identity used to assess robustness to signal strength is shown; values range from 10 to 105 (0.7 - 7% of total genes). **E.** The set of cell identity proportions used to test the ability to identify rare cell states is shown. The number of cells per treatment-label pair ranges from 50 to 175. **F.** Method performance is compared as the fraction of perturbed genes increases (left), and as the average number of marker genes per population increases (right). Results are depicted as averages with 95% confidence intervals shown by shading. **G.** Method performance is compared as the proportion of cell types is varied (left). The Gini coefficient reflects the degree of population size disproportion among cell states. **H.** The performance of integration methods alone or in combination with ICAT (as ICAT$_{Seurat}$ or ICAT$_{Scan}$) is compared across the same metrics for the perturbation (H1) and signal strength experiments (H2).

ICAT in label conservation and asymmetrical population detection (ARI, F1-Unique), and offered comparable treatment-mixing performance to Scanorama (LISI), indicating that Seurat does not accurately separate known populations relative to Scanorama or ICAT (Fig. 2F).

For each of the perturbation severity tests and with low marker gene numbers, Seurat was surprisingly outperformed by the no-integration scenario for ARI (Fig. 2F) which can be explained by Seurat's general under-clustering of cells (Fig. S2A). Scanorama exhibited a similar drop in ARI performance at low marker numbers, but consistently outperformed no integration (Fig. 2F). These results show ICAT more accurately and robustly resolves cell identities across experimental conditions than either Seurat or Scanorama integration workflows.

ICAT most accurately detects rare populations.

ICAT was unaffected by differences in cell identity proportions as assessed by global and rare population label conservation via ARI and the F1-Rare metric, respectively, while Seurat and Scanorama each exhibited decreased performance as the disparity in proportions increased (Fig. 2G, left). The performance of ICAT was robust and stronger than both Scanorama and Seurat in each measure.

ICAT improves and stabilizes the performance of integration workflows.

Outside of requiring a data matrix, ICAT is agnostic to input data. Accordingly, we also evaluated clustering performance for two extended methods: $ICAT_{Seurat}$ and $ICAT_{Scan}$, whereby cells were first integrated using Seurat or Scanorama, respectively, then clustered with ICAT. In this way, ICAT can be utilized either as an alternative to integration methods, or as an additional step after integration.

Compared to the typical Seurat workflow, $ICAT_{Seurat}$ led to significant improvements in all three simulated experiments, leading to greater robustness to perturbation (Fig. 2H1, left), increased signal sensitivity (Fig. 2H2, left), and better isolation of rare cells (Fig. 2G, middle). When compared to Scanorama, $ICAT_{Scan}$ better isolate rare cells (Fig. 2G, right), and improved performance in some metrics in perturbation and signal experiments. Gains were more modest compared to $ICAT_{Seurat}$, primarily since Scanorama alone already exhibits strong performance (Fig. 2H1-2, right). In particular, $ICAT_{Scan}$ substantially improved treatment-mixing (LISI) across conditions compared to Scanorama alone (Fig.

2H1, right). Interestingly, at both signal extremes, stand-alone ICAT scored higher in the F1-Stim metric than either combined or original integration methods (Fig. 2H2), indicating that ICAT better preserves discriminatory signals compared to either integration algorithms.

In summary, the results from simulated data demonstrate that ICAT offers robust and accurate cell state identification that exceeds both Seurat and Scanorama, two current state-of-the-art integration algorithms, across a range of simulated conditions including severe perturbations, low marker gene numbers, and variations in relative cell abundance. Further analysis shows that identifying cell states with ICAT following integration produces higher quality clusters compared to either current integration workflow alone.

ICAT produces higher quality clusters in real datasets than integration methods.

To assess ICAT's ability to correctly cluster cells in real perturbation experiments, we made use of three publicly available datasets (Table S4) (Tian *et al.*, 2019; Kagohara *et al.*, 2020; Kang *et al.*, 2017). While no true benchmark datasets currently exist for defining cell identities in perturbation experiments, we selected datasets that either contain genotyped cell lines, or are from well-studied cell types with confidently labeled cells. The first dataset, by Tian et al., was created as a standard to evaluate methods developed for scRNA-seq data analysis (Tian *et al.*, 2019). To evaluate ICAT, we used their CellMix dataset that was produced by creating various mixtures of nine cells from combinations of H2228, HCC827, and H1975 adenocarcinoma cell lines, sequencing the cell mixtures, then downsampling the resulting counts to produce pseudo-cells that are in-line with true single-cell data. While the Tian dataset was developed to test integration across batch effects, we reasoned that variance introduced by batch effects is similar to perturbation-induced variation. To better mimic distinct phenotypes, we selected only four cell "types", consisting of pure H2228, HCC827, and H1975 cells, along with a single mixture with equal proportions of each cell line, denoted "Mixed" (Fig. S3, Table S4). The second dataset from Kagohara et al. (Kagohara *et al.*, 2020) offers a simple perturbation experiment in which three genotyped cell lines of head and neck squamous carcinoma cells were treated with either PBS (control) or the chemotherapeutic cetuximab (Table S4). Finally, we assessed performance in a

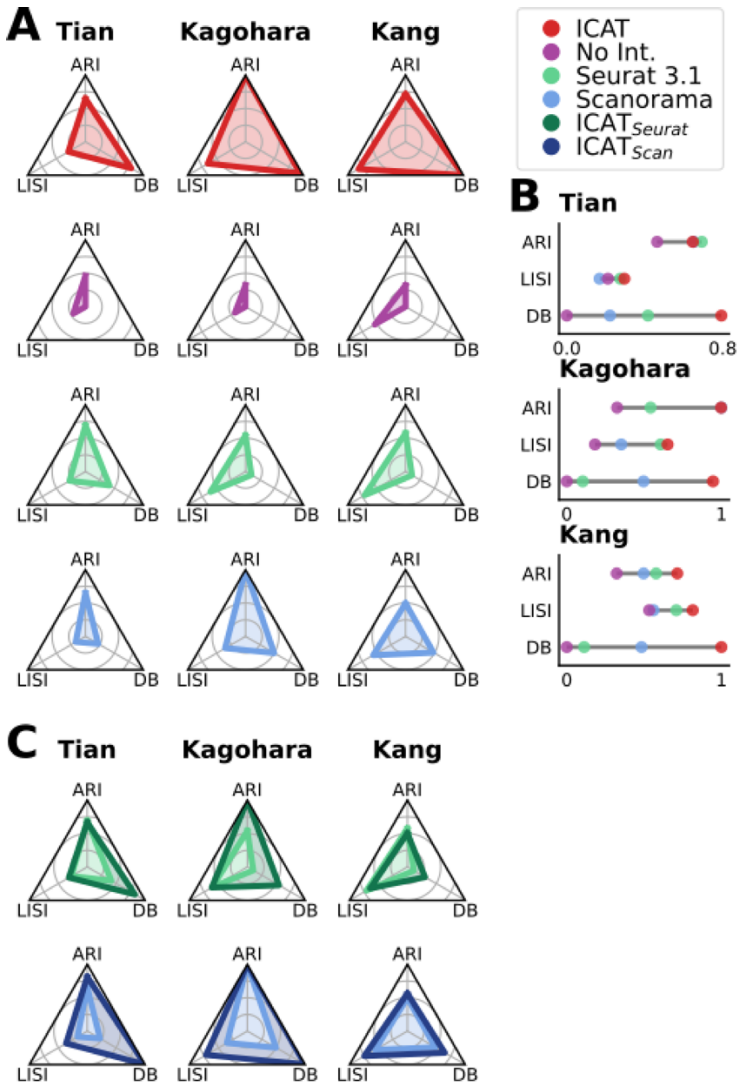


Figure 3: ICAT outperforms current integration methods in identifying cell states across treatments in real datasets. A. Spider plots compare the performance for each algorithm within each dataset for ARI, LISI, and DB quality metrics. **B.** Lollipop plots highlight the differences in the metrics for each method across the same datasets. **C.** Spider plots comparing Seurat to ICAT$_{Seurat}$ and Scanorama to ICAT$_{Scan}$, respectively. All metrics are scaled from 0 to 1, where 1 is best (closest to the apex of each corner). Abbreviations: DB Davies-Bouldin metric, and otherwise as in Fig. 2. ICAT performance for the Tian dataset was evaluated using the ICAT$_{C+T}$ implementation, while Kagohara and Kang datasets were evaluated using the ICAT$_{C}$ implementation.

peripheral blood mononuclear cell dataset collected by Kang et al. (Kang *et al.*, 2017) with eight cell types in control and IFN-β-stimulated conditions (Table S4). Prior to analysis, each dataset was preprocessed by normalizing cell counts and selecting highly variable genes using *Scanpy*. To ensure each tested algorithm analyzed the same input data, we provided the same *Scanpy* preprocessed data to each method.

Because NCFS feature-weighting is a computationally intensive task, and since the Kagohara and Kang datasets each contain > 20,000 cells, it was necessary to subselect cells for feature weighting. To this end, ICAT selects a representative sample of cells via submodular optimization using the *apricot* (Schreiber *et al.*, 2020) Python package. Submodular optimization via *apricot* was able to effectively represent the data space using only a small number of cells (Fig. S4).

To measure ICAT's performance in these datasets, we again evaluated clustering quality using ARI and LISI metrics to measure global label conservation and treatment status mixing, respectively.

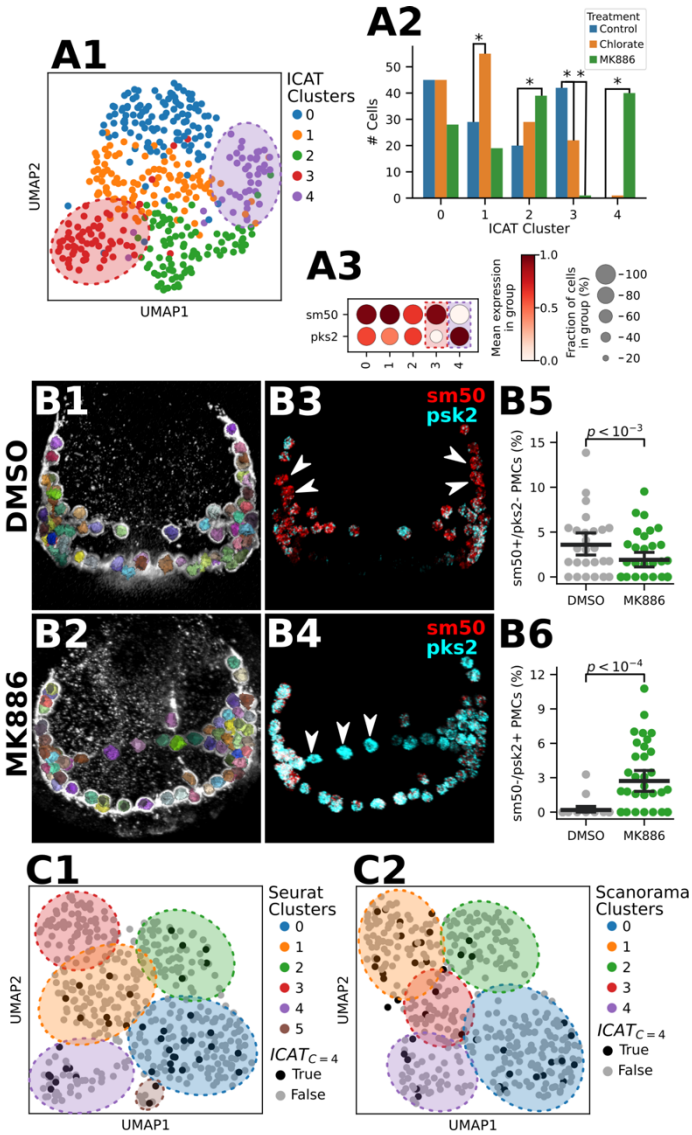
Because cell identity is resolution dependent and likely imperfect in real datasets (Luecken and Theis, 2019; Lähnemann *et al.*, 2020), we also calculated the label-free Davies-Bouldin (DB) metric to assess the density of clusters produced by each method (Davies and Bouldin, 1979). The DB metric measures similarity between a cluster and the next most similar cluster. To facilitate comparison, each score was scaled between 0 and 1 such that 1 represents better clustering while 0 represents worse clustering.

In the Tian dataset, ICAT outperformed both existing integration methods in LISI and DB metrics and was comparable to both in ARI, in which ICAT was slightly worse than Seurat and Scanorama (Fig. 3A-B, Table S5). ICAT outperformed both integration methods in the Kagohara and Kang datasets (Fig. 3A-B). Aside from ARI in the Kagohara dataset, in which Scanorama and ICAT scored equally well (Fig. 3B), ICAT produced the highest scores for ARI, LISI, and DB metrics (Fig. 3B). Overall, ICAT generated higher quality clusters in real datasets compared to existing integration algorithms (Fig. 3A-B, Table S5).

Compared to Scanorama, $ICAT_{Scan}$ led to greater ARI, LISI and DB scores across the three analyzed datasets (Fig. 3C, Table S5), aside from ARI in the Kagohara dataset where Scanorama and $ICAT_{Scan}$ scored equally well (Fig. 3C). For Seurat, $ICAT_{Seurat}$ resulted in improvements that exhibit dataset and metric dependence: cell states were more readily identified in the Kagohara dataset, shown by the dramatic increase in ARI scores, and clusters were tighter and better separated as shown by large increases in DB scores (Fig. 3C). In contrast, ARI scores were maintained or marginally reduced with $ICAT_{Seurat}$ compared to standalone Seurat for the Tian and Kang datasets. The results demonstrate that ICAT offers a marked improvement in cell state identification in multi-condition scRNA-seq experiments compared to the current integration workflows Seurat and Scanorama.

ICAT alone accurately predicts subpopulation response to perturbation *in vivo.*

To empirically test ICAT predictions on cell state compositions following perturbation, we generated scRNA-seq data using the SMART-Seq2 protocol (Picelli *et al.*, 2014). Specifically, we isolated and sequenced the skeletogenic lineage in sea urchins known as primary mesenchyme cells (PMCs) at 18 hours post-fertilization, and compared cells from embryos subjected to chemical treatments that perturb skeletal patterning, specifically, MK-886 (MK) and chlorate (Pidgeon *et al.*, 2007; Simionato



et al., 2021; Piacentino *et al.*, 2016). PMCs are of interest because this population of cells receives patterning cues during development that cause their diversification into subsets; since MK and chlorate each perturb distinct patterning cues, we reasoned that these differences would be reflected in PMC subpopulation composition (Piacentino *et al.*, 2016; Lyons *et al.*, 2014; Sun and Ettensohn, 2014; Piacentino *et al.*, 2015). After quality control and cell filtering, we retained expression data for 435 cells (147 control, 134 MK886, and 154 chlorate-treated cells).

ICAT clustered the combined PMC expression profiles into five distinct cell states (Fig. 4A1). The impacts of MK

Figure 4: ICAT most accurately defines subpopulation response to perturbation. SMART-seq2 was performed on cells isolated from controls or from embryos treated with perturbants chlorate or MK-886. **A.** ICAT predicts five clusters from the combined data (A1), with treatment-dependent subpopulation compositions (pairwise Fisher's Exact test, fdr < 0.05) (A2). ICAT-calculated gene weights plotted by rank show *sm50* and *pks2* as highly informative for cell-state membership, with 37 total genes being considered informative (weight > 1) (A3). MK-induced differences are defined by reciprocal expression patterns of ICAT-identified genes, *sm50* and *pks2,* in subpopulations 3 (sm50+/pks2-) and 4 (sm50-/pks2+) (A4). **B.** ICAT predictions for control (B1,3,5) and MK-866 (B2,4,6) are validated by HCR FISH analysis. **B1-2.** Automated detection and segmentation of PMCs *in vivo* from 3-D image data is shown as various colors; each color represents an individual cell. **B3-4.** The expression of *sm50* (red) and *pks2* (cyan) transcripts are shown in the same embryos; arrowheads indicate PMCs that express only *sm50* or only *pks2*. **B5-6.** MK-866 treated embryos exhibit statistically significantly fewer *sm50+/pks2-* PMCs (B5) and more *sm50-/pks2+* PMCs (B6) than controls, consistent with the predictions from ICAT (A). Dot plot error bars denote 95% confidence intervals of the mean with each dot representing an individual embryo (n_{DMSO} = 26, n_{MK886} = 37; B5 (Binomial GLM, μ_{DMSO}= 3.59%, μ_{MK886}= 1.91%; β_{MK886}= -0.71, $p < 10^{-3}$; B6 Binomial GLM, μ_{DMSO}= 0.19%, μ_{MK886}= 2.71%; β_{MK886}= 2.60, $p < 10^{-4}$). **C.** Integration methods fail to identify the MK886-enriched *sm50+/pks2-* subpopulation (ICAT cluster 4). UMAP projections show clusters generated by Seurat (C1) or Scanorama (C2) for the same data, showing with the cells comprising the MK866-specific asymmetric *sm50+/pks2-* population (ICAT cluster 4, black dots) distributed across the multiple clusters generated by each method. The clusters produced by each integration method are indicated by colored regions.

exposure were more dramatic than those of (Fig. 4A2); we therefore focused on the effects of MK. Two clusters showed stark compositional differences between control and MK-treated cells, in that cluster 3 lacks MK cells compared to controls, while clusters 2 and 4 show significant enrichment for MK cells (G-test fdr < 0.05, post hoc Fisher's Exact test fdr < 0.05; Fig. 4A2, Table S6-7). Since MK-treated cells dominated cluster 4 such that it appears to be an asymmetric population, we more closely examined clusters 3 and 4. Investigating highly predictive genes identified by ICAT found that clusters 3 and 4 were distinguished by inverse expression of genes *pks2* and *sm50* (Fig. 4A3-4), such that cluster 3 displayed a *sm50+/pks2-* phenotype, while cluster 4 exhibited a reciprocal *sm50-/pks2+* pattern.

To experimentally validate ICAT's predicted enrichment of the *sm50-/pks2+* cell state and depletion of *sm50+/pks2* cells, in MK886-treated PMCs, we performed single molecule fluorescence *in situ* hybridization (FISH) to quantify expression levels for both genes in DMSO (control) and MK-treated embryos (Choi *et al.*, 2018) (Fig. 4B, Fig. S5). To accurately map FISH signals to PMCs, embryos were stained for PMCs using the PMC-specific 6a9 antibody. Using Ilastik (Berg *et al.*, 2019), we trained a random forest classifier (out-of-bag error = 0.14) to predict and segment individual PMCs in each confocal image stack (Fig. 4B1-2, Fig. S6). Following methods for relative FISH quantification from the original paper (Choi *et al.*, 2018), we calculated cell-level expression values for each PMC by first restricting fluorescence signal to PMC-labeled voxels, then calculated the average relative intensity for each gene in every PMC (Fig. 4B3-4, Fig. S5). Across embryos, we labeled individual PMCs as *sm50+* or *sm50-* if their standardized *sm50* expression values were above the 75th or below the 25th percentiles of all PMCs, respectively (Fig. S7). The procedure was repeated for *pks2* to determine *pks2+* and *pks2-* PMCs. We found that MK-treated embryos displayed both a significant loss of the *sm50+/pks2-* phenotype, as well as a significant enrichment of the *sm50-/pks2+* cell state relative to controls (adjusted $p < 0.05$; Fig. 4B5-6, Table S9-10): verifying the ability of ICAT to correctly identify affected cell states in scRNA-seq datasets from perturbed samples.

Integrating the PMC scRNA seq data using either Seurat or Scanorama failed to identify the enriched *sm50-/pks2+* cell state in MK-treated embryos, and instead incorrectly merged those cells with

other cell states (Fig. 4C, Fig. S8). In fact, Seurat failed to identify any dependence of cell state composition on treatment status (G-test fdr > 0.05; Fig. S8A2, Table S11). Scanorama, on the other hand, correctly identified the depletion of *sm50+/pks2-* PMCs in MK embryos (Fig S8B2-3, Table S12-13; G-test fdr < 0.05, post hoc Fisher's Exact test fdr < 0.05), but importantly, failed to identify *sm50-/pks2+* enrichment in MK embryos (Fig. S8B2-3, Table S11-12; G-test fdr > 0.05). ICAT is the only method among these three to correctly identify the cell-state specific sm50 and pks2 responses to MK exposure, which accounts for most of the differences in cell groupings between the methods (Jaccard similarity, Fig. S8A3, B3). Of the compared methods, Seurat failed to detect the MK-specific cell identity; moreover, Seurat failed to produce even qualitatively correct results since it did not identify any treatment-specific effects on PMCs (Fig. S8A2), contradicting the known disruptions to PMCs in response to disruption of the ALOX signaling pathways (Piacentino *et al.*, 2016). These results demonstrate that ICAT significantly improves the computational capability to accurately detect and describe cellular responses to perturbation in scRNA-seq data.

Discussion

One of the key promises of scRNA-seq technology is to more accurately define complex biological systems by offering a more finely resolved perspective on transcriptional landscapes than was previously possible. However, due to technical and biological noise, taking advantage of such a powerful tool in diverse biological conditions has proven to be challenging. We therefore developed the ICAT algorithm to improve the accuracy of cell state identification from scRNA-seq data that includes perturbation treatments. By combining sparse feature weighting with semi-supervised Louvain community detection, ICAT not only correctly identifies shared and disparate cell states, but also produces interpretable results that aid in understanding analytical outcomes and selecting the most suitable downstream analyses.

When comparing the performance of ICAT to state-of-the-art integration algorithms in simulated datasets, we found that ICAT performs more robustly in response to perturbation severity, readily identifies cell states at lower signal levels, better isolates asymmetrical and activated cell states, and more readily identifies rare cell populations when compared to either Seurat or Scanorama. ICAT also outperforms Seurat and Scanorama in three benchmark datasets, further solidifying ICAT's ability to better identify cell states across biological conditions compared to either integration method alone. By testing ICAT on datasets with > 20,000 cells (Kang *et al.*, 2017; Kagohara *et al.*, 2020), we demonstrated ICAT's ability to analyze the large datasets, necessary for any practical scRNA-seq tool. Depending on the experimental system, researchers may still elect to use integration algorithms. Excitingly, the choice between ICAT and integration methods is not mutually exclusive; we show that following Scanorama or Seurat with ICAT produces higher quality clusters compared to either of these integration workflows alone.

By experimentally validating two uniquely predicted cell states in a developmental model using single molecule FISH, we empirically demonstrated that the capability of ICAT to accurately identify perturbed cell states surpasses current state-of-the-art integration workflows. ICAT was the only method among the three tested that correctly identified these distinct states and their perturbation responses, whereas both Seurat and Scanorama integration workflows failed, highlighting ICAT as a substantial advancement in identifying and describing cell state behaviors across biological conditions and perturbations.

With the growing ubiquity of scRNA-seq experiments and the development of ever more sophisticated experimental designs, we anticipate that ICAT will play an important role in highly resolving cell state-specific responses to treatments and differing biological conditions. In the future, the ICAT framework can easily be extended to operate on other single-cell modalities, such as scATAC-seq, and will offer a potential route to intelligently cluster multi-modal data from heterogeneous conditions.

Declarations

Availability of data and materials.

ICAT, *ncfs*, and *sslouvain* Python packages are freely available on pip and github (github.com/BradhamLab). Scripts to perform simulated data generation and benchmark comparisons are available as a SnakeMake workflow at (github.com/BradhamLab/ICAT_manuscript). Archived data files for both simulated and real data are also available at the same repository. The pipeline to align and quantify SMART-seq2 reads is freely available at (github.com/BradhamLab/scPipe), and the raw sequencing data is available at GEO accession GSE164240. Likewise, all scripts to segment PMCs and quantify HCR FISH signal are available at (github.com/BradhamLab/hcr_quantify) and the image data is available upon request.

Acknowledgements

We thank Prof. Masanao Yajima (Dept. of Mathematics and Statistics, Boston University) for helpful discussions. We thank Prof. Charles Ettensohn (Carnegie Mellon University, PA, USA) for the PMC antibody. We thank Jose Ordovas-Montanes and Sanjay Prakadan of the Shalek lab (MIT) for help with scRNA-seq library preparation. We thank Dr. Todd Blute for his advice and help with FACS and microscopy imaging.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by NSF IOS 1656752 (CAB).

Author Contributions

The study was conceived by CAB and DYH. The computational work was performed by DYH, with advice from WEJ. JH and DTZ collected the urchin scRNA-seq data. NRS, CFT, KRM, ATL, and AED performed the imaging experiments. DYH, KRM, and AG trained the PMC prediction model. The manuscript was written by DYH, CAB, DTZ and NRS, and edited by all coauthors.

References

Berg,S. *et al.* (2019) ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods 2019*, 1–7.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.

Bradham,C.A. and McClay,D.R. (2006) p38 MAPK is essential for secondary axis specification and patterning in sea urchin embryos. *Development*, **133**, 21–32.

Büttner,M. *et al.* (2019) A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods*, **16**, 43–49.

Chen,S. *et al.* (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Choi,H.M.T. *et al.* (2018) Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development*, **145**, dev165753.

Davies,D.L. and Bouldin,D.W. (1979) A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-1**, 224–227.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Gross,J.M. *et al.* (2003) LvTbx2/3: a T-box family transcription factor involved in formation of the oral/aboral axis of the sea urchin embryo. *Dev. Camb. Engl.*, **130**, 1989–1999.

Haber,A.L. *et al.* (2017) A single-cell survey of the small intestinal epithelium. *Nature*, **551**, 333–339.

Haghverdi,L. *et al.* (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, **36**, 421–427.

Hie,B. *et al.* (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.*, **37**, 685–691.

Hogan,J.D. *et al.* (2020) The developmental transcriptome for Lytechinus variegatus exhibits temporally punctuated gene expression changes. *Dev. Biol.*, **460**, 139–154.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Imbert,A. *et al.* (2021) FISH-quant v2: a scalable and modular analysis tool for smFISH image analysis.

Kagohara,L.T. *et al.* (2020) Integrated single-cell and bulk gene expression and ATAC-seq reveals heterogeneity and early changes in pathways associated with resistance to cetuximab in HNSCC-sensitive cell lines. *Br. J. Cancer*, **123**, 101–113.

Kang,H.M. *et al.* (2017) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, **36**, 89–94.

King,G. and Zeng,L. (2001) Logistic Regression in Rare Events Data. *Polit. Anal.*, **9**, 137–163.

Korsunsky,I. *et al.* (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods*, **16**, 1289–1296.

L. Lun,A.T. *et al.* (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.

Lähnemann,D. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**, 31.

Lam,S.K. *et al.* (2015) Numba: a LLVM-based Python JIT compiler. In, *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15. Association for Computing Machinery, New York, NY, USA, pp. 1–6.

Li,Y. *et al.* (2020) Genomic insights of body plan transitions from bilateral to pentameral symmetry in Echinoderms. *Commun. Biol.*, **3**, 1–10.

Liao,Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Luecken,M.D. *et al.* (2020) Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*, 2020.05.22.111161.

Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.
Lyons,D.C. *et al.* (2014) Specification to biomineralization: following a single cell type as it constructs a skeleton. *Integr. Comp. Biol.*, **54**, 723–33.
McClay,D.R. and Fink,R.D. (1982) Sea urchin hyalin: Appearance and function in development. *Dev. Biol.*, **92**, 285–293.
Nicholas Sofroniew *et al.* (2020) napari/napari: 0.3.7rc3 Zenodo.
Pedregosa,F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
Perillo,M. *et al.* (2020) Regulation of dynamic pigment cell states at single-cell resolution. *eLife*, **9**, e60388.
Piacentino,M.L. *et al.* (2015) Late Alk4/5/7 signaling is required for anterior skeletal patterning in sea urchin embryos. *Development*, **142**, 943–952.
Piacentino,M.L. *et al.* (2016) RNA-Seq identifies SPGs as a ventral skeletal patterning cue in sea urchins. *Dev. Camb. Engl.*, **143**, 703–14.
Picelli,S. *et al.* (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
Pidgeon,G.P. *et al.* (2007) Lipoxygenase metabolism: roles in tumor progression and survival. *Cancer Metastasis Rev.*, **26**, 503.
Regev,A. *et al.* (2017) The Human Cell Atlas. *eLife*, **6**.
Reichardt,J. and Bornholdt,S. (2006) Statistical mechanics of community detection. *Phys. Rev. E*, **74**, 016110.
Satija,R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**.
Schaum,N. *et al.* (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.
Schelker,M. *et al.* (2017) Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.*, **8**, 2032.
Schreiber,J. *et al.* (2020) apricot: Submodular selection for data summarization in Python. *J. Mach. Learn. Res.*, **21**, 1–6.
Seabold,S. and Perktold,J. (2010) statsmodels: Econometric and statistical modeling with python. In, *9th Python in Science Conference*.
Simionato,G.B. *et al.* (2021) Lack of 5-lipoxygenase in intramembranous and endochondral 129Sv mice skeleton and intramembranous healing. *Arch. Oral Biol.*, **131**, 105266.
Sokal,R. and Rohlf,F. (2013) Biometry : the principles and practice of statistics in biological research / Robert R. Sokal and F. James Rohlf. *SERBIULA Sist. Libr. 20*.
Stuart,T. and Satija,R. (2019) Integrative single-cell analysis. *Nat. Rev. Genet.*, 1.
Sun,Z. and Ettensohn,C.A. (2014) Signal-dependent regulation of the sea urchin skeletogenic gene regulatory network. *Gene Expr. Patterns*, **16**, 93–103.
Tian,L. *et al.* (2019) Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, **16**, 479–487.
Traag,V. (2021) vtraag/louvain-igraph.
Trombetta,J.J. *et al.* (2014) Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Curr. Protoc. Mol. Biol.*, **107**, 4.22.1-17.
Tyler,S.R. *et al.* (2021) PMD Uncovers Widespread Cell-State Erasure by scRNAseq Batch Correction Methods.
Virtanen,P. *et al.* (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods*, **17**, 261–272.
Wagner,D.E. *et al.* (2018) Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, eaar4362.
van der Walt,S. *et al.* (2014) scikit-image: image processing in Python. *PeerJ*, **2**, e453.

Wolf,F.A. *et al.* (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
Yang,W. *et al.* (2012) Neighborhood Component Feature Selection for High-Dimensional Data. *J. Comput.*, **7**.