


# Subtle cell states resolved in single-cell data

Caleb Lareau

 Check for updates

## SEACells identifies rare cell states in large datasets, enabling atlas-scale studies.

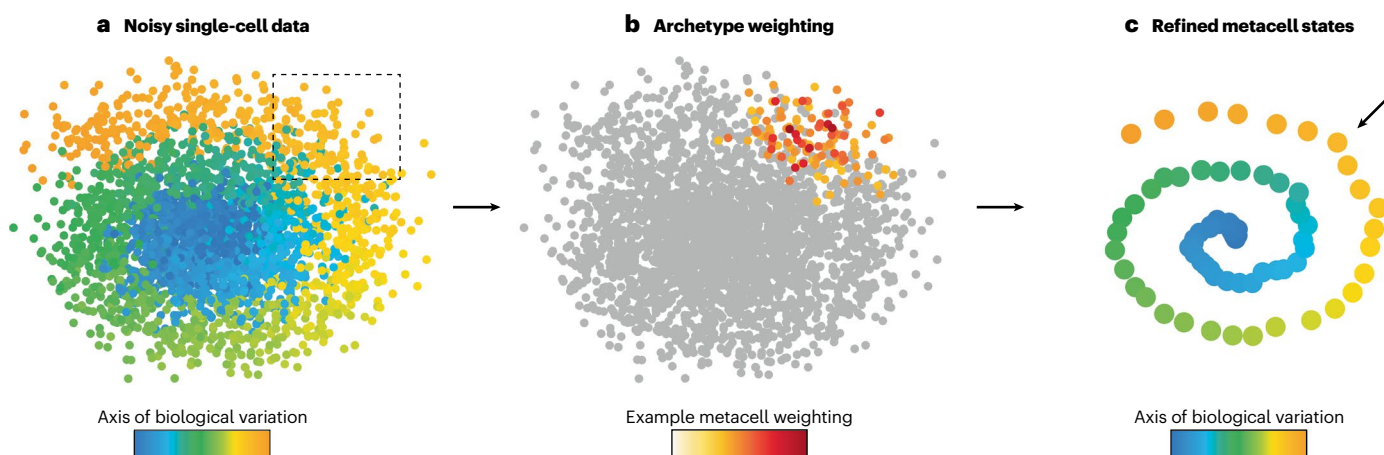
Massive-scale profiling of cell populations by single-cell sequencing is providing unparalleled insight into biological systems. However, the sparsity and noise of single-cell data hinder the detection and interpretation of rare and transitory cell states. In *Nature Biotechnology*, Persad et al.<sup>1</sup> describe SEACells, a data aggregation algorithm that finds the ‘sweet spot’ between sparse single-cell data and widely used clustering approaches that obscure biological variation. An improved method for defining ‘metacells’, SEACells outperforms existing approaches that infer relevant cell states from high-dimensional single-cell datasets. The power of SEACells is particularly apparent in the analysis of scATAC-seq (single-cell assay for transposase-accessible chromatin by sequencing) data, which are much more sparse than transcriptomic measurements, unlocking biological insights in high-dimensional data.

A tenet of single-cell genomics assays is that measurements are incomplete and noisy; nevertheless, the profiles represent a useful approximation of the transcriptome or epigenome of a cell. As technical limitations limit the utility of profiles from any one cell, most computational methods leverage thousands of parallel profiles to discern meaningful biological signals in the data. Consequently, a typical analysis workflow for single-cell genomics data infers clusters of dozens to hundreds of similar cells and annotates these coarse groups through prior knowledge of marker genes. Although this workflow has generally been effective in smaller-scale studies, the derivation of relatively few cell states guided by clustering techniques collapses heterogeneity in complex datasets, including meta-analyses across

multiple data cohorts. Thus, many single-cell analyses obscure novel and subtle biological variation underlying complex phenotypes.

An intermediate approach between analyses at the individual cell and cluster levels is to represent large, sparse single-cell sequencing data as ‘metacells’ – small groups of cells that together represent a specific biological state. Though the true signal can be difficult to discern with sparse measurements, technical variation can be minimized after identifying ‘pockets’ of cells in high-dimensional space that represent approximate replicates sampled from an underlying cell state. After identifying these pockets, cellular archetypes – linear combinations of the observed data – can better reflect biological variation after aggregation. As a conceptual example, Fig. 1a illustrates noisy data at the single-cell level in a reduced-dimensionality space. Following an archetype analysis, per-cell weights from the observed data can be estimated per archetype (Fig. 1b). The result of the computation is a new representation of the high-dimensional data in a form where biological axes of variation, such as differentiation trajectories, can be pronounced (Fig. 1c). This concept is particularly relevant for epigenomic assays like droplet-based scATAC-seq<sup>2,3</sup>, in which only 1–10% of peaks are detected per cell (compared to ~10–45% of expressed genes detected per cell)<sup>4</sup>, reflecting a greater degree of data sparsity.

Persad et al. refine the metacell inference with the aim of capturing subtle cell states that have been missed by previous metacell methods<sup>5–7</sup>. The key advance in SEACells is the use of a cell-archetype analysis in which metacell composition is estimated from the kernel matrix (representing cell-by-cell similarity) rather than the data matrix (empirically defined as the cell-by-feature) as used by other methods<sup>5–7</sup>. Conceptually, as the number of cells  $n$  has increased exponentially in single-cell analyses, embedding cells in this  $n$ -by- $n$  space leverages



**Fig. 1 | Key components of the SEACells approach to analysis of single-cell datasets.** **a**, A schematic of noisy single-cell data plotted in two dimensions, wherein an axis of biological variation is obscured. The rectangle represents a cellular group along the axis of biological variation to be summarized via SEACells. **b**, Example of the weighting of the single cells from the highlighted

region used to infer an individual metacell via the SEACells archetype analysis. **c**, Output of SEACells algorithm that resolves the simulated biological variation by mitigating noise. The arrow indicates the computed metacell from the previous panels.

the massively parallel profiling of cells in current genomics assays. Consequently, SEACells can, in complex settings (such as differentiation or tumorigenesis), retain representative biological heterogeneity and intermediate states often lost in applying previous metacell approaches<sup>5–7</sup> and continue to yield relevant biological information in increasingly large, atlas-scale datasets.

The authors demonstrate the power of SEACells in several applications. In a characterization of the continuous differentiation of CD34<sup>+</sup> hematopoietic stem and progenitor cells, they recover biologically relevant metacells, including gradations of erythroid progenitors associated with known transcription factor activity. Crucially, the metacell ‘state space’ is shown to be largely independent of the density or number of cells sampled, providing a compelling example of how the kernel space improves metacell identification over previous methods. Beyond cell state identities, the authors infer transcription factor activities and regulatory modules underlying hematopoietic differentiation that had been inaccessible in traditional analyses schemes. In another application to an atlas of >600,000 cells derived from patients with COVID-19, they reveal expression signatures of CD4<sup>+</sup> T<sub>H</sub>17 genes in individuals with persistent severe disease, an insight that would be missed using conventional analytical strategies. Thus, the practical utility of SEACells when applied to large datasets from independent atlases enables harmonization to reveal distinct biological insights, whereas alternative approaches would either fail to capture the disease-associated variation in the data or could not scale to the cell numbers derived from multiple atlases.

Whereas SEACells fills a key analytical gap in single-cell workflows, other approaches have emerged for rapidly summarizing large datasets from varied sources. Many of these tools rely on reference-based projections to produce annotations from existing datasets that are readily interpretable for commonly studied cell types (for example, peripheral blood mononuclear cells in healthy individuals). One example is a recent dictionary-learning approach implemented in Seurat<sup>8</sup>. However, reference-based projection methods require prespecified axes of variation, which can eliminate unexpected sources of transcriptional or epigenomic changes, such as cell-state perturbations in disease. Thus, an opportunity for future algorithm development would be to balance the preservation of de novo variation in datasets (supported by SEACells) and annotations of known cell states from cell dictionaries, as implemented in reference-based methods. Further, while SEACells demonstrated sensitivity to recover biologically relevant transitory cell states in various settings, future metacell algorithms that discern the appropriate number of groups without a user-defined parameter may enable more robust inferences of metacells.

Beyond the applications presented by Persad et al., the importance and utility of metacell analyses have been evident in certain disease-relevant settings. For example, metacell analysis was instrumental in the discovery of a rare population of brain mural cells that may be targeted by CD19-directed immunotherapies in a form of ‘on-target, off-tumor’ toxicity<sup>9</sup>. Identification of this CD19<sup>+</sup> mural cell state required scalable reanalysis of multiple cell atlases from previous studies, which reduced technical variation present at the single-cell level to reveal a biological state. In other words, the inference of a clinically relevant cell population necessitated the use of a metacell approach that could scale and sustain complex analyses within large datasets, a mode of analysis that may become increasingly common as comprehensive profiling of disease and healthy tissues becomes more widespread.

In sum, the work of Persad et al. offers an essential advance for single-cell analyses by providing a facile workflow that minimizes technical variations (such as batch effects) while preserving biological variability. Thus, SEACells is well positioned to catalyze future biological insights from massive-scale atlas datasets, such as the ever-growing Human Cell Atlas<sup>10</sup>. Beyond the utility of the method, I found the choice of the SEACells name to evoke a stirring image of seashells spread along a coastline – not only is each shell distinct and intelligible, but the full constellation of shells portrays a picture that is more than the sum of its parts. In a similar sense, the advances enabled by SEACells help us appreciate the landscape of the many distinct cell states that contribute to biological function and are perturbed in disease.

Caleb Lareau  

Department of Pathology, Stanford University, Palo Alto, CA, USA.

✉ e-mail: [clareau@stanford.edu](mailto:clareau@stanford.edu)

Published online: 17 May 2023

## References

1. Persad, S. et al. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-023-01716-9> (2023).
2. Satpathy, A. T. et al. *Nat. Biotechnol.* **37**, 925–936 (2019).
3. Lareau, C. A. et al. *Nat. Biotechnol.* **37**, 916–924 (2019).
4. Chen, H. et al. *Genome Biol.* **20**, 241 (2019).
5. Bilous, M. et al. *BMC Bioinformatics* **23**, 336 (2022).
6. Baran, Y. et al. *Genome Biol.* **20**, 206 (2019).
7. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. *Genome Biol.* **23**, 100 (2022).
8. Hao, Y. et al. *Cell* **184**, 3573–3587.e29 (2021).
9. Parker, K. R. et al. *Cell* **183**, 126–142.e17 (2020).
10. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. *Nature* **550**, 451–453 (2017).

## Competing interests

The author declares no competing interests.