Sorting apples from oranges in single-cell expression comparisons

Fiona K Hamey & Berthold Göttgens

Two methods for comparing single-cell expression data sets help address the challenge of integrating data across conditions and experiments.

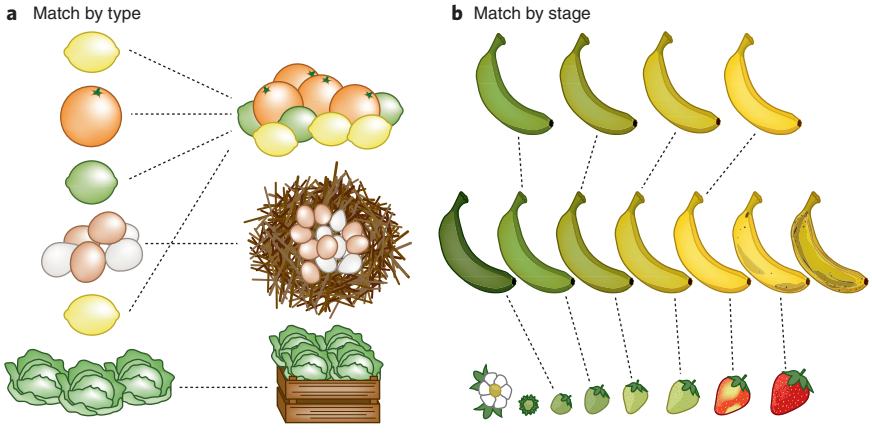
New single-cell molecular profiling techniques are rapidly transforming biomedical research across a diverse range of tissues and organisms. One of the main challenges in analyzing such data arises from so-called batch effects that result from technical differences between samples and hamper robust comparisons between experiments. Publications from the Hemberg[1] and Shen-Orr[2] laboratories now present two methodologies for comparing cell samples from experiments involving different conditions, technologies and even species.

Single-cell RNA sequencing (scRNA-seq) has made it possible to obtain biological insights through the bioinformatic analysis of large numbers of individual cells. Many studies rely on dimensionality-reduction techniques to project data onto two or three dimensions for visualization. These methods reveal similarities and differences between cells, but they do not easily lead to quantifiable comparisons. In parallel, unsupervised clustering is often used to group single cells by the similarity of their gene expression profiles, and has helped scientists decipher population heterogeneity, for example, by identifying previously unknown cell types. A single sample commonly contains heterogeneous cell populations that may be at different stages of a directional process such as differentiation or response to a perturbation. scRNA-seq profiles have been used to investigate gene expression changes during such a process by computationally ordering cells along trajectories on a so-called pseudotime axis that aims to reconstruct the process[3].

One of the most exciting applications of single-cell profiling is the comparison of gene expression between states to investigate how cells change across conditions. In particular, this has implications for understanding disease and identifying potential therapeutic targets. An emerging practice is for researchers to compare their data with data for reference samples, which provides an important rationale for ongoing efforts to generate gold-standard data sets such as the Human Cell Atlas initiative[4]. It is often desirable to combine scRNA-seq data from multiple experiments, yet differences due to sample origin, preparation and sequencing, rather than cell state, can make this challenging.

Kiselev *et al.*[1] present an approach for mapping cells from a new experiment onto an annotated reference (**Fig. 1a**). Their algorithm, scmap-cluster, calculates distances in gene expression space to match cells to their most similar cluster in the reference data. scmap first identifies a subset of features on which to perform calculations. Interestingly, the authors find that selecting genes with a higher than expected frequency of zero expression produces more accurate mappings than selecting highly variable or random genes, an observation that may be useful for other types of scRNA-seq data analysis. Although the algorithm attempts to match cells to a reference set, cells remain unassigned if they do not show gene expression patterns similar to those in the reference data. This is an essential consideration, as there will be incomplete overlap among the cell types present for many comparisons. The authors have made a praiseworthy effort to render their method user-friendly by providing both an R package and a web version, and ensuring that the algorithm runs quickly on large data sets.

Because discrete clustering cannot readily capture continuous aspects of differentiation processes, Kiselev *et al.*[1] also outline a nearest-neighbor approach to accurately compare cells to an unclustered (e.g., pseudotime-ordered) reference data set with the scmap-cell version of their algorithm.



Katie Vicari/Springer Nature

Figure 1 | Computational methods match up data from multiple experiments. (**a**) The concept behind scmap. Individual or grouped items from a new data set can be matched to groups from an existing reference data set. (**b**) The concept behind cellAlign. Items arranged in ordered sequences can be matched to identify overlapping stages, even when the items originate from different sources, such as different species.

Fiona K. Hamey and Berthold Göttgens are at the Department of Haematology, Cambridge Institute for Medical Research, and Wellcome–MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK.
e-mail: bg200@cam.ac.uk

© 2018 Nature America, Inc., part of Springer Nature. All rights reserved.

For more in-depth comparison of pseudotime orderings, Alpert *et al.*[2] developed cellAlign. cellAlign uses dynamic time warping to align sections of two trajectories with shared expression patterns, thereby enabling the comparison of expression dynamics (**Fig. 1b**). Excitingly, cellAlign not only is able to compare whole transcriptomes, but also can utilize specific genes or gene modules to assess differences between conditions. Alpert *et al.*[2] even analyzed scRNA-seq data from preimplantation embryos to identify gene modules with different patterns of temporal behavior across human and mouse development, thereby demonstrating the ability of their algorithm to contrast data from very difference sources.

As scmap and cellAlign differ in their aims of either mapping or aligning data, the choice of approach will depend on the study in question. It is worth noting that neither method aims to 'batch correct' data to enable downstream analysis such as dimensionality reduction of the integrated data sets. Such an approach is explored in papers from the Satija[5] and Marioni[6] labs and may be necessary for certain comparisons, such as finding genes that are differentially expressed in different conditions. Moreover, it will be interesting to see how pseudotime comparisons may be adapted for comparative analyses of pseudospace orderings[7], where instead of being ordered by temporal progression, single cells are arranged by spatial coordinates inferred from the expression of positional landmark genes.

The application for which mapping or alignment may be the most revealing—unexplored in the scmap and cellAlign papers—is the assessment of perturbations to the transcriptional landscape, particularly in the context of disease. Analysis of perturbed cell populations from humans or from mouse models in comparison with their wild-type counterparts should give insight into which populations or stages of differentiation are most affected and in what way gene expression changes.

A major challenge when comparing data generated via different protocols is how to address the varying technical properties inherent to different methods, such as the huge variation in the number of genes detected per cell. Kiselev *et al.*[1] and Alpert *et al.*[2] both briefly touch on this: the creators of scmap note that their method struggles to find the nearest neighbors of cells with zero expression of many genes (often owing to dropout or failed capture during library generation), and the cellAlign authors discuss the need to scale gene expression because of technical differences in the data. How reliably comparisons between such technically different data sets can be made will certainly be explored and debated in the scRNA-seq field in the future. Initiatives to generate vast numbers of data sets requiring integration, such as the Human Cell Atlas[4], are certain to help drive further innovation in this area.

COMPETING INTERESTS
The authors declare no competing interests.

1. Kiselev, V.Yu., Yiu, A. & Hemberg, M. *Nat. Methods* **15**, 359–362 (2018).
2. Alpert, A., Moore, L.S., Dubovik, T. & Shen-Orr, S.S. *Nat. Methods* **15**, 267–270 (2018).
3. Trapnell, C. *et al. Nat. Biotechnol.* **32**, 381–386 (2014).
4. Regev, A. *et al. bioRxiv* Preprint at http://biorxiv.org/content/early/2017/05/08/121202 (2017).
5. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4096 (2018).
6. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4091 (2018).
7. Ibarra-Soria, X. *et al. Nat. Cell Biol.* **20**, 127–134 (2018).

© 2018 Nature America, Inc., part of Springer Nature. All rights reserved.