Title: Sorting apples from oranges in single-cell expression comparisons

Authors: Fiona K. Hamey and Berthold Göttgens

Corresponding author: Berthold Göttgens, E-mail: bg200@cam.ac.uk

Affiliations:

University of Cambridge, Department of Haematology, Cambridge Institute for Medical Research & Wellcome and MRC Cambridge Stem Cell Institute, Hills Road, Cambridge CB2 0XY, UK.

Two methods for comparing single-cell expression datasets 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 analysing these 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 cells between samples from different conditions, technologies and even species.

Single-cell RNA sequencing (scRNA-seq) has made it possible to extract 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 visualisation. These methods reveal similarities or differences between

cells, but 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 profile, and has helped to decipher the heterogeneity present within populations, 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 in such a process by computationally ordering cells along trajectories on a so-called 'pseudotime' axis that aims to reconstructs the process[3].

One of the most exciting applications of single-cell profiling is to compare 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 against reference samples, thus providing an important rationale for ongoing efforts to generate gold-standard datasets 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. Whilst the algorithm attempts to match cells to a reference set, cells remain unassigned if they do not show similar gene expression patterns to the reference data. This is an essential consideration, as there will be an incomplete overlap in 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 the algorithm runs quickly on large datasets.

Since 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 dataset with the scmap-cell version of their algorithm.

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 is not only able to compare whole transcriptomes, but can also utilize specific genes or gene modules to assess differences between conditions. Alpert *et al*.[2] even analyze scRNA-seq data from pre-implanation embryos to identify gene modules with different patterns of temporal behaviour across human and mouse development, demonstrating the ability of their algorithm to constrast data from very difference sources.

Since scmap and cellAlign differ in their aim 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 datasets. Such an approach is explored in papers from the Satija[5] and Marioni[6] labs and may be necessary for comparisons such as finding genes differentially expressed between 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, yet was unexplored in the scmap and cellAlign papers, is the assessment of perturbations on the transcriptional landscape, particularly in the context of disease. Analysing perturbed cell populations from patients or mouse models against their wild-type counterparts should give insight into which populations or stages of differentiation are most affected and in what way their gene expression changes.

A major challenge when comparing data generated from 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. Both publications briefly touch on this; the creators of scmap note that their method struggles to find the nearest neighbours of cells with zero expression in many genes (often due to dropout, or failed capture during library generation), and the cellAlign authors discuss the need for scaling gene expression due to technical differences in the data. How reliably

comparisons between such technically different datasets can be made will certainly be explored and debated within the scRNA-seq field in future. Initiatives generating vast numbers of datasets requiring integration such as the Human Cell Atlas[4] are certain to help drive further innovation in this area.

References

1. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **XXX,** XXX (2018).
2. Alpert, A., Moore, L. S., Dubovik, T. & Shen-Orr, S. S. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* **XXX**, XXX (2018).
3. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32,** 381–386 (2014).
4. Regev, A. *et al.* The Human Cell Atlas. *bioRxiv* Preprint at http://biorxiv.org/content/early/2017/05/08/121202 (2017).
5. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. (2018). Available at http://dx.doi.org/10.1038/nbt.4096.
6. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. (2018). Available at http://dx.doi.org/10.1038/nbt.4091.
7. Ibarra-Soria, X. *et al.* Defining murine organogenesis at single-cell resolution

reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.* **20,** 127–134 (2018).

Figure caption:

Computational methods match up data from multiple experiments. (A) The concept behind scmap. Individual or grouped items from a new dataset can be matched to exisiting groups from a reference dataset. (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 species.

Competing interests:

The authors declare no competing financial interests.

Figure