Detection and removal of barcode swapping in single-cell RNA-seq data.

Jonathan A. Griffiths[1], Arianne C. Richard[1,2], Karsten Bach[3], Aaron T.L. Lun[1,†], John C Marioni[1,4,5,*]

1 Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, CB2 0RE, United Kingdom.
2 Cambridge Institute for Medical Research, University of Cambridge, Cambridge, CB2 0XY, United Kingdom.
3 Department of Pharmacology, University of Cambridge, Cambridge, CB2 1PD, United Kingdom.
4 EMBL-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom.
5 Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, United Kingdom.

† aaron.lun@cruk.cam.ac.uk
***** marioni@ebi.ac.uk

Recent reports show that the DNA barcodes used to label multiplexed libraries can "swap" on patterned flow-cell Illumina sequencing machines, including the HiSeq 4000, HiSeq X, and NovaSeq (Sinha *et al.* , 2017). This results in mislabelling whereby reads assigned to one sample derive from molecules in another, thus compromising the interpretation of many -omic assays. Barcode swapping is particularly problematic for single-cell RNA-seq (scRNA-seq), where many libraries are multiplexed together. For example, barcode swapping can lead to cells that appear to falsely express particular genes, potentially yielding spurious correlation patterns that can confound clustering and other downstream analyses.

A number of widely-used scRNA-seq library preparation methods isolate and process individual cells in wells of a microwell plate, before performing library preparation in parallel (Picelli *et al.* , 2014). A unique combination of sample barcodes characterises the library associated with each cell, typically by adding a different barcode at each end of a cDNA molecule. One barcode indexes

the row position for each cell on the microwell plate and the other barcode indexes the column position. Swapping of either or both barcodes therefore moves transcripts between cell libraries.

We used two datasets to quantify the fraction of barcode-swapped reads in these experiments. We define the swapped fraction to be the fraction of total cDNA reads across all sequencing libraries multiplexed on a single flow cell lane that were mislabelled. In a first experiment (see Supplementary File section 3), we used cell library sizes to determine a swapped fraction of 2.19±0.08%. Applying the same model to the same libraries sequenced on a Hiseq 2500, we identified a swapped fraction of 0.22±0.01%, an order of magnitude lower than the HiSeq 4000 rate. In a second experiment (Nestorowa *et al.* (2016), see Supplementary File section 4), we used expression levels of individual genes to determine a mean swapped fraction of 2.275±0.359% across plates, consistent with the first experiment. Notably, these rates are substantially lower than those previously reported (Sinha *et al.* , 2017). We did not observe correlation between the swapped fraction and the concentration of free barcode in each library (Figure 1A). This suggests that, for typical experimental concentrations of free barcode (2-6nM), other factors play a greater role in determining the rate of barcode swapping.

New single-cell RNA-seq protocols use microfluidic systems to massively multiplex library preparation by capturing individual cells in droplets (Macosko *et al.* , 2015; Zheng *et al.* , 2017). These protocols label cells by incorporating a randomly chosen cell barcode in addition to a sample barcode. Each sample typically contains thousands of cells, each with its own cell barcode. Only the sample barcode is expected to swap, thus moving transcripts between samples leaving the cell identifier unchanged. In addition, droplet methods use unique

molecular identifiers (UMIs) to collapse PCR duplicate reads. This means that barcode swapping is particularly concerning for droplet-based experiments: a sample containing hundreds of reads of a particular molecule receives the same single UMI count as another sample containing a single swapped read. In this manner, the effect of barcode swapping on droplet expression profiles is effectively inflated in UMI data compared to read count data.

We consider two major effects of barcode swapping in droplet-based experiments. Firstly, it is possible that the same cell barcode is used in two or more samples. Between these samples, swapping of transcripts with this cell barcode will homogenise cell transcriptomes, similar to the effect observed in plate-based assays. Alternatively, consider a situation where a "donor" sample contains a cell barcode that is not present in another "recipient" sample. For transcripts labelled with this unique barcode, swapping will produce a new artefactual cell library in each recipient sample. This new cell library will contain a similar expression profile to the original cell in the donor sample.

To identify whether barcode swapping was creating artefactual cells, we tested whether samples from droplet-based experiments shared more cell barcodes than expected by chance. For both of the HiSeq 4000-sequenced experiments tested, at least one sample comparison exhibited excessive sharing. This was not observed for any comparison in an experiment sequenced on HiSeq 2500, or when one of the aforementioned HiSeq 4000 experiments was re-sequenced on a HiSeq 2500 (Supplementary Figures 28-31).

To mitigate the effect of swapping, cell barcodes that are shared across samples may be removed prior to downstream analysis. This excludes both homogenised and swap-derived artefactual libraries. However, such an approach may discard a large amount of data when many cells are captured. To resolve this, we have implemented an algorithm that removes swapped molecules be-

fore cell identification. Specifically, we considered molecules across multiplexed samples that contain the same combination of UMI, cell barcode, and aligned gene. Due to combinatorial complexity, these molecules are extremely unlikely to arise by chance, and are almost certain to be generated by barcode swapping. For each combination, we calculated the fraction of all observed reads from each sample. Where one sample contained the vast majority of all reads for a molecule ($\geq 80\%$), we considered this as the sample-of-origin, and removed the molecule count from all other samples. Where this was not the case, we removed the molecule from all samples (Figure 1C). In one experiment, a technical failure in two samples resulted in a considerable number of swap-derived artefacts (Figure 1D, Supplemental File section 5.3). Application of our approach successfully removed swap-derived artefactual cells from this experiment (Figure 1E, Supplementary File section 6.2.3). In contrast, the method removed only a negligible number of molecules between experiments that were not multiplexed together ($<0.0005\%$, Supplementary File section 6.2.4). Our method is implemented in the DropletUtils R package (Lun, 2018) for 10X Genomics experiments, and can be easily applied in a conventional analysis pipeline.

We have estimated the fraction of barcode-swapped reads on the HiSeq 4000 at around 2.25%, which is lower than previously reported (Sinha *et al.* , 2017). Different amounts of free DNA barcode did not affect our swapping fraction estimates, suggesting that other factors are primarily responsible for swapping fraction variability between plates. We recommend that plate-based scRNA-seq experiments that reuse cell barcodes should continue to be sequenced on non-patterned flow-cell machines (e.g. HiSeq 2500), to minimise barcode swapping. We have implemented a method for removing swapped transcripts from 10X Genomics data without removing en-

tire cell libraries. This permits the cost-effective use of the highest-throughput sequencing machines (e.g. HiSeq 4000) for large-scale droplet scRNA-seq experiments while excluding the confounding effects of barcode swapping.

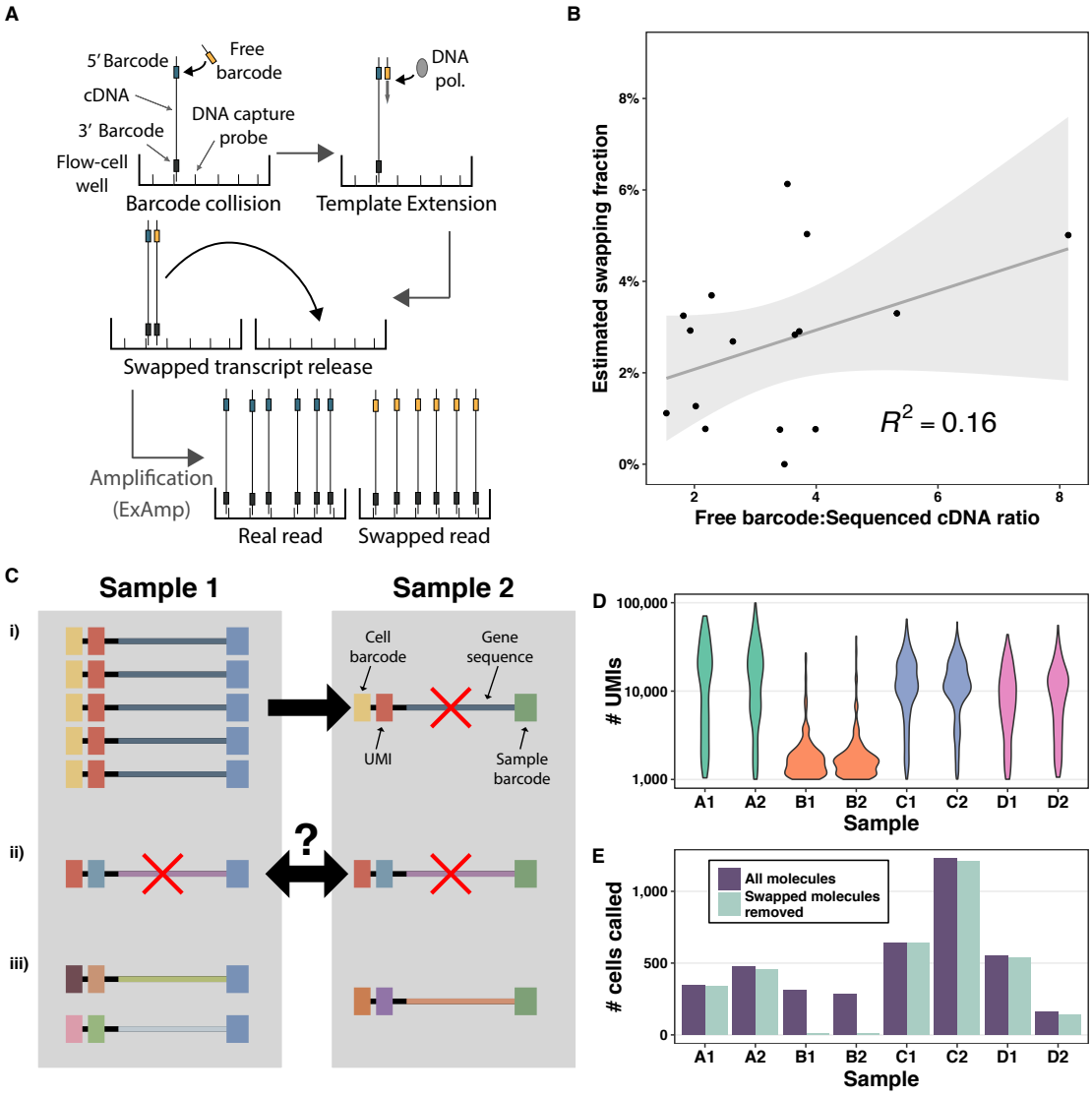

FIGURE CAPTION

(**A**) A schematic for the mechanism for barcode swapping. On new Illumina sequencing machines, flow cell seeding and DNA amplification take place simultaneously. As a result, free sample indexing barcodes remain in solution, and may be extended using library DNA molecules as a template. The transfer of mislabelled molecules between nanowells of the flow cell results in clustering and sequencing of artefactually labelled DNA molecules. (**B**) Estimated swapping fractions for different plates of the Nestorowa *et al.* (2016) dataset do not depend on the ratio of the concentration of free barcode to the concentration of cDNA of the correct length for sequencing. A linear regression fit is shown through the points, alongside its 95% confidence interval. The slope of the fit is not significant different from 0 (p = 0.129). Free barcode concentration was estimated using Agilent's Bioanalyser Expert software, considering DNA lengths of between 40 and 75 base pairs. Sequenced cDNA was considered between 400 and 800bp. (**C**) A summary of the algorithm to remove swapped molecules from droplet data. Molecules found in different samples with identical UMI, cell barcode, and aligned gene were considered to have swapped. Where the molecule existed mostly in one sample, we excluded the molecule from all other samples (i). Where the molecule is spread between samples, we excluded it from all samples (ii). Molecules in one sample only were retained (iii). (**D**) In one 10X Genomics experiment, two samples (B1, B2) contained called cells that appeared to be artefactual, as they had very few assigned molecules (as shown), and shared nearly all (>90%) of their cell barcodes with each other. We hypothesised that these were artefacts of barcode swapping. Cells were called using *emptyDrops* (Lun, 2018), using an adjusted $p < 0.01$ and a minimum of 1,000 UMIs. (**E**) The number of called cells for these samples is shown before and after application of our algorithm. After application, we exclude nearly all

cells from the compromised samples, confirming that we can successfully remove barcode-swapped artefacts from droplet scRNA-seq datasets.

References

Lun, Aaron T.L. 2018. *DropletUtils: Utilities for Handling Single-Cell Droplet Data*. R package version 0.99.14.

Macosko, Evan Z., Basu, Anindita, Satija, Rahul, Nemesh, James, Shekhar, Karthik, Goldman, Melissa, Tirosh, Itay, Bialas, Allison R., Kamitaki, Nolan, Martersteck, Emily M., Trombetta, John J., Weitz, David A., Sanes, Joshua R., Shalek, Alex K., Regev, Aviv, & McCarroll, Steven A. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**(5), 1202–1214.

Nestorowa, Sonia, Hamey, Fiona K., Sala, Blanca Pijuan, Diamanti, Evangelia, Shepherd, Mairi, Laurenti, Elisa, Wilson, Nicola K., Kent, David G., & Göttgens, Berthold. 2016. A single cell resolution map of mouse haematopoietic stem and progenitor cell differentiation. *Blood*, Jan., blood–2016–05–716480.

Picelli, Simone, Faridani, Omid R., Bjorklund, Asa K., Winberg, Gosta, Sagasser, Sven, & Sandberg, Rickard. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, **9**(1), 171–181.

Sinha, Rahul, Stanley, Geoff, Gulati, Gunsagar Singh, Ezran, Camille, Travaglini, Kyle Joseph, Wei, Eric, Chan, Charles Kwok Fai, Nabhan, Ahmad N, Su, Tianying, Morganti, Rachel Marie, Conley, Stephanie Diana, Chaib, Hassan, Red-Horse, Kristy, Longaker, Michael T, Snyder, Michael P, Krasnow, Mark A, & Weissman, Irving L. 2017. Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv*, Apr.

Zheng, Grace X. Y., Terry, Jessica M., Belgrader, Phillip, Ryvkin, Paul, Bent, Zachary W., Wilson, Ryan, Ziraldo, Solongo B., Wheeler, Tobias D., McDermott, Geoff P., Zhu, Junjie, Gregory, Mark T., Shuga, Joe, Montesclaros, Luz, Underwood, Jason G., Masquelier, Donald A., Nishimura, Stefanie Y., Schnall-Levin, Michael, Wyatt, Paul W., Hindson, Christopher M., Bharadwaj, Rajiv, Wong, Alexander, Ness, Kevin D., Beppu, Lan W., Deeg, H. Joachim, McFarland, Christopher, Loeb, Keith R., Valente, William J., Ericson, Nolan G., Stevens, Emily A., Radich, Jerald P., Mikkelsen, Tarjei S., Hindson, Benjamin J., & Bielas, Jason H. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, **8**(Jan.), 14049.

DATA AVAILABILITY STATEMENT

A Github repository (`https://github.com/MarioniLab/BarcodeSwapping2017`) contains a detailed report that expands on the analyses described herein, describing models and showing results. The repository also contains a script to download the data, and the code used to generate the report.

FUNDING

J.A.G. was funded by Wellcome Trust award [109081]. A.C.R. was funded by MRC grant [MR/P014178/1]. K.B. was funded by a Cambridge Cancer Centre studentship. A.T.L.L. was funded by Cancer Research UK [A17197]. J.C.M. was funded by Cancer Research UK [A17197] and EMBL core funding. Sequencing of mouse T-cells was supported by Wellcome Trust grants [103930] and [100140].

ACKNOWLEDGMENTS

We thank Gillian Griffiths, Berthold Göttgens, and his group for help with data access.

AUTHOR CONTRIBUTIONS

J.A.G. performed data analyses; A.T.L.L. and K.B. contributed code; K.B. and A.C.R. generated data; J.A.G., A.T.L.L., and J.C.M. wrote the manuscript; all authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.