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

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

1 Cancer Research UK Cambridge Institute, University of Cambridge, CB2 0RE, UK
2 Department of Pharmacology, University of Cambridge, CB2 1PD
3 EMBL-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, CB10 1SD, Cambridge, UK
4 Wellcome Trust Sanger Institute, Wellcome Genome Campus, CB10 1SD, Cambridge, UK

***** marioni@ebi.ac.uk

Multiplexing is a widely-used procedure that allows multiple DNA libraries to be pooled together for efficient sequencing. However, recent reports suggest that the DNA barcodes that label different libraries can "swap" on patterned flow-cell Illumina sequencing machines, including the HiSeq 4000, HiSeq X, and NovaSeq, thereby mislabelling molecules [1, 2]. This may compromise many types of -omic assays, but it is particularly problematic for single-cell RNA-seq (scRNA-seq), where many libraries are multiplexed together.

A number of widely used plate-based scRNA-seq library preparation methods isolate and process individual cells in wells of a microwell plate, before performing library preparation in parallel [3]. A unique combination of sample barcodes labels the library of each cell, typically with one barcode at each end of a cDNA molecule. One barcode provides a row index for each cell on the microwell plate and the other barcode provides a column index. Barcode swapping therefore moves transcripts between cells.

We generated a dataset (see Supplementary Files, "Richard data") where two plates of single-cell libraries were multiplexed for sequencing on the HiSeq 4000 using two mutually exclusive barcode sets. We expect to only observe reads labelled with combinations of barcodes from the same plate. Reads labelled with one barcode from one plate and one barcode from the other should not exist, as these barcodes were never mixed in library preparation. However, such “impossible” reads were observed at 1.1% of the frequency of the expected barcode combinations.

We assumed that swapping was a rare event, such that very few transcripts swapped both barcodes. Therefore, most of the movement of transcripts due to barcode swapping will occur between libraries that share exactly one barcode, as these transcripts require only a single swap to move from one library to another. Indeed, the number of reads for each impossible barcode combination is proportional to the sum of library sizes for all expected libraries that share exactly one barcode (**Figure 1A**). From the gradient, we estimated the rate of swapping to be 2.19±0.08%. Applying the same model to the same libraries sequenced on a HiSeq 2500 yielded a swapping rate estimate of 0.22±0.01%.

To confirm this result, we used a published dataset (see Supplementary Files, “Nestorowa data”) where the same libraries were sequenced on both the HiSeq 2500 (less affected by barcode swapping) and the HiSeq 4000 [4]. For each cell, we modelled the HiSeq 4000 counts as a linear sum of the HiSeq 2500 counts from the same cell, from cells that share exactly one barcode, and cells that share no barcodes. This yielded estimates typically between 1-3% for the swapping rate between cells sharing one barcode, consistent with results from the Richard data (**Figure 1B**). We also estimated the swapping rate between cells that do not share any barcodes as 1% or less for 75% of plates. Note that these rates measure an increase in swapping compared to the HiSeq 2500, not an absolute rate.

To identify whether barcode swapping was affecting all genes, we again utilised the Nestorowa data. For each gene, we modelled the HiSeq 4000 counts

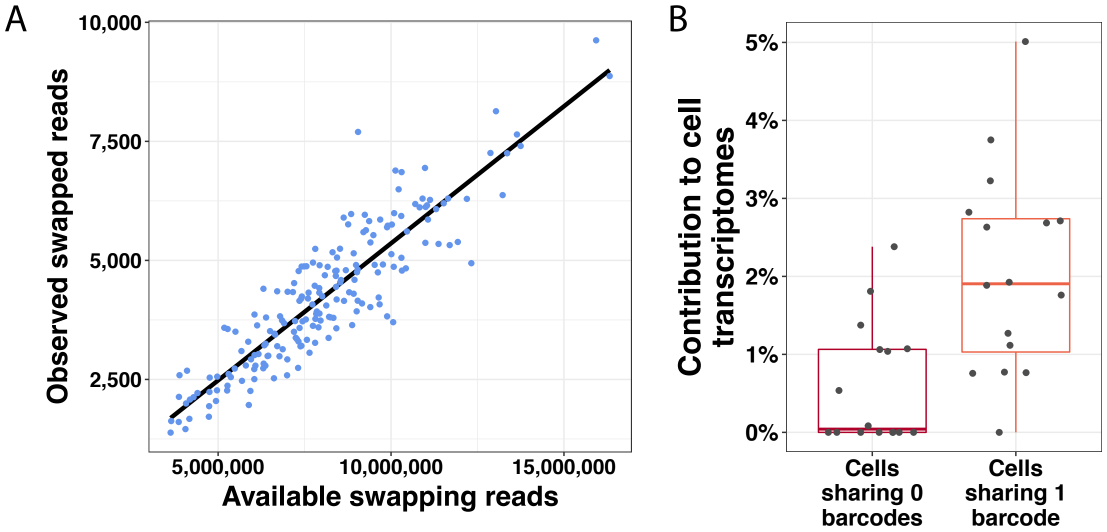

Figure 1: **Barcode swapping rate estimates.** (**A**) The number of observed reads in an impossible barcode combination is proportional to the sum of library sizes for all expected libraries that share exactly one barcode with that combination. We estimated a swapping rate of 2.19±0.08% from the gradient. (**B**) We fitted a linear model that uses HiSeq 2500 libraries to infer the proportion of HiSeq 4000 libraries that are derived from other samples due to barcode swapping. Cells that share a single barcode contribute greater fractions of their transcriptome to each other, consistent with the barcode swapping mechanism presented in [2]. We estimated the mean swapping rate as 2.275±0.359%.

as a linear function of the HiSeq 2500 counts. After including an additional term that accounted for swapping between cells that share one barcode, we observed a significant improvement to the model fit for 90% of genes, suggesting a transcriptome-wide barcode swapping effect (Supplementary Figure 19).

New single-cell RNA-seq protocols use microfluidic systems to automate stages of library preparation by capturing individual cells in droplets [5, 6]. 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 while retaining an identical cell identifier.

As all samples use the same cell barcode set, 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 plate based assays.

Alternatively, consider a situation where a “donor” sample contains a cell barcode that is not present in any other “recipient” samples. 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 a HiSeq 2500 experiment (Supplementary Figures 22 to 24). In all experiments, the fraction of shared barcodes was low (below 3%).

We suggest that the extent of cell barcode sharing across samples should be quantified as part of quality control of droplet-based scRNAseq experiments, as excess sharing is symptomatic of barcode swapping and the presence of artefactual cells. All cells with cell barcodes shared across samples should be removed prior to downstream analysis of droplet-based data. This procedure will exclude both homogenised libraries as well as any swap-derived artefactual cells.

We have confirmed the existence of barcode swapping in scRNA-seq data from a HiSeq 4000 machine. We have estimated the rate of

swapping between libraries at 1-3%, which is lower than previously reported [2]. Swapping also occurs on the HiSeq 2500 at approximately 1/10th of the rate as on the HiSeq 4000. The effects of swapping on droplet-based data may be easily removed, though there is no obvious solution for plate-based methods. Whether or not barcode swapping will compromise downstream biological conclusions remains to be explored; clearly, however, caution will be required when analysing swapping-affected single cell RNA-seq data.

References

[1] J. Hadfield, "Index mis-assignment between samples on HiSeq 4000 and X-Ten," Dec. 2016. Available at `http://enseqlopedia.com/2016/12/index-mis-assignment-between-samples-on-hiseq-4000-and-x-ten/`.

[2] R. Sinha, G. Stanley, G. S. Gulati, C. Ezran, K. J. Travaglini, E. Wei, C. K. F. Chan, A. N. Nabhan, T. Su, R. M. Morganti, S. D. Conley, H. Chaib, K. Red-Horse, M. T. Longaker, M. P. Snyder, M. A. Krasnow, and I. L. Weissman, "Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing," *bioRxiv*, Apr. 2017.

[3] S. Picelli, O. R. Faridani, A. K. Bjorklund, G. Winberg, S. Sagasser, and R. Sandberg, "Full-length RNA-seq from single cells using Smart-seq2," *Nature Protocols*, vol. 9, pp. 171–181, Jan. 2014.

[4] S. Nestorowa, F. K. Hamey, B. P. Sala, E. Diamanti, M. Shepherd, E. Laurenti, N. K. Wilson, D. G. Kent, and B. Göttgens, "A single cell resolution map of mouse haematopoietic stem and progenitor cell differentiation," *Blood*, pp. blood–2016–05–716480, Jan. 2016.

[5] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets," *Cell*, vol. 161, pp. 1202–1214, May 2015.

[6] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj,

A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, p. 14049, Jan. 2017.

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.

ACKNOWLEDGMENTS

We thank 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.