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

J.A. Griffiths[1], A.T.L. Lun[1], A.C. Richard[1], K. Bach[2], J.C. Marioni[1,3,4]

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

Introduction

• Multiplexing is a widely-used procedure that pools DNA libraries together for efficient and technically robust sequencing.

• 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.

• This is particularly problematic for single-cell RNA-seq (scRNA-seq) where many libraries are multiplexed together.

Here, we have:

• Quantified the extent of HiSeq 4000 barcode swapping using Smart-Seq2 data.

• Identified how barcode swapping can compromise droplet-based scRNA-seq

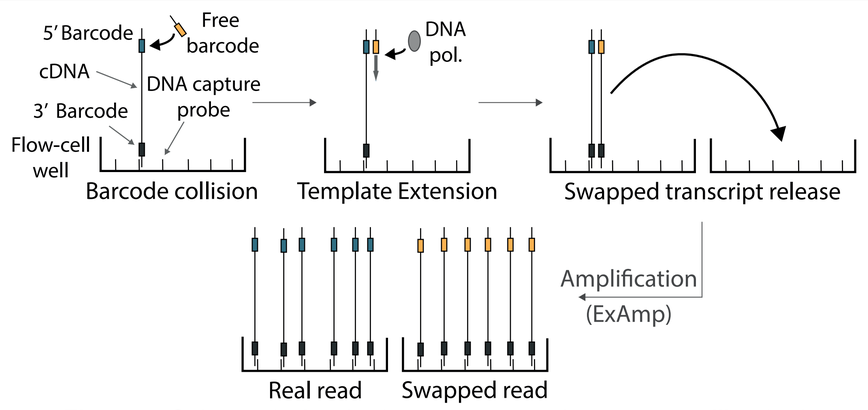

Figure 1 Schematic for the barcode swapping process. Illumina's new simultaneous seeding and cDNA extension is essential for the occurrence of barcode swapping.

Droplet-based assays

Droplet cDNAs have:
• An Illumina sample barcode
• A randomly drawn cell barcode (e.g. from a pool of ~750,000 for 10X).

Only the Illumina barcode is expected to swap.

There are two possible situations:
• If a cell barcode is present in multiple samples, **cell transcriptomes are homogenised** (two-way swapping).
• If a cell barcode is present in one sample only, **artefactual libraries are created** (one-way swapping).

Solution: exclude cells that share cell barcodes between samples

This step excludes cells affected by transcriptome homogenisation, and any artefactual cell libraries that have been created.

Swapping creates artefacts in compromised droplet scRNA-seq samples

In one 10X experiment, we noticed that two samples showed:
• Very small library sizes (**Figure 2**)
• An extreme degree of barcode sharing (**Figure 3**)

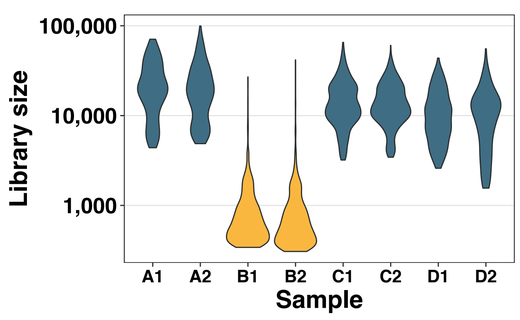

Figure 2 Called cells from samples B1 and B2 are much smaller than cells from other samples.

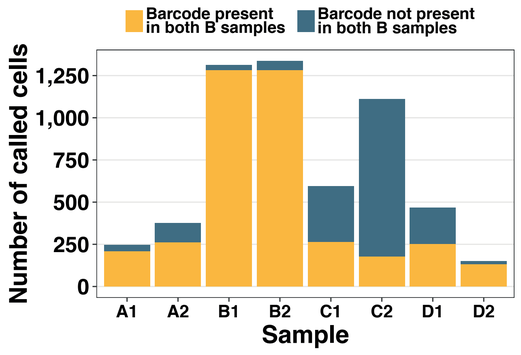

Figure 3 Samples B1 and B2 possess almost no unique barcodes.

• We hypothesised that library preparation had failed, and that **CellRanger identified swapped-in libraries as cells**.

• Cells whose barcodes were shared between the good and bad samples were larger than uniquely barcoded cells (**Figure 4**; $p<10^{-200}$).

• Larger cells swapped larger absolute amounts of transcript, which was then called as a cell by CellRanger.

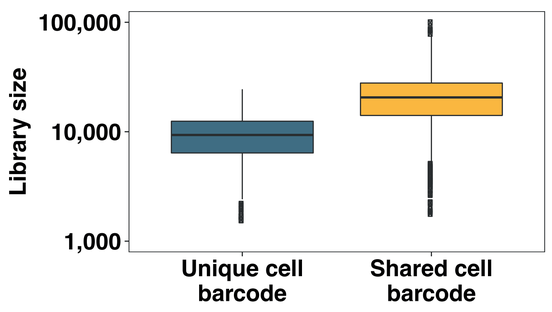

Figure 4 Good-quality cells whose barcodes were observed in the poor-quality samples were larger than those seen only in good-quality samples. This suggests that larger cells were swapping more transcript, consistent with our barcode swapping hypothesis.

We have drawn three conclusions:

- **Cell barcode sharing between samples is a useful QC metric.**
- **Identifying cells by library size alone is not robust.**
- **Care should be taken when multiplexing cells of very different sizes: barcode swapped artefacts may swamp small cells.**

Plate-based assays

For plate-based scRNA-seq protocols (e.g. Smart-seq2):
• Transcripts are labelled with a sample barcode at each end of the molecule.
• One barcode indexes the row-position of the cell on its microwell plate, and the other the column index.
• Swapping of individual barcodes therefore moves reads between cells on the same row or column of a plate.

The rate of swapping on HiSeq 4000 is 2.2%

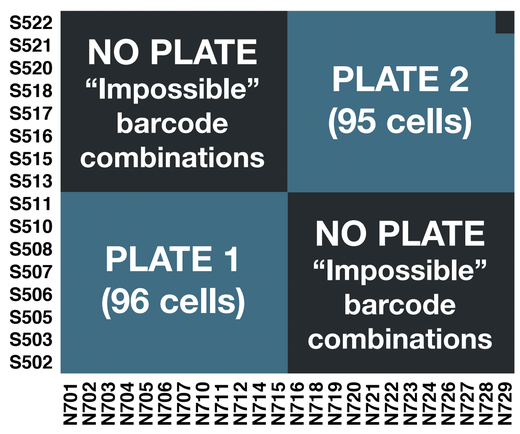

Figure 5 We multiplexed two 96-well plates on HiSeq 4000. Blue shaded barcode combinations were loaded with cells; grey shaded combinations should not give sequencing reads.

• We assumed that most swapped transcripts only changed one of their barcodes.

• This implies that **the libraries of each impossible combination derived from cells that share exactly one barcode with it** (**Figure 6**).

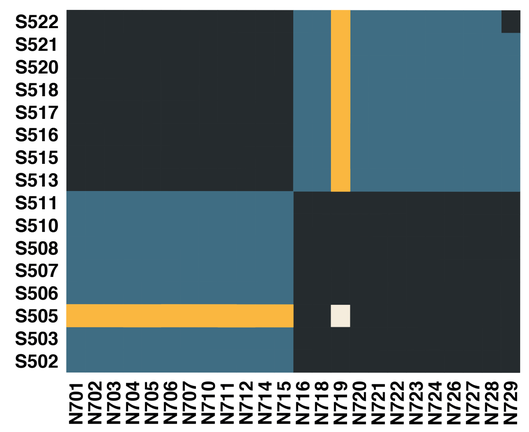

Figure 6 The barcode combination S505-N719 (white) recieves swapped reads from libraries that share exactly one barcode - these are highlighted in yellow.

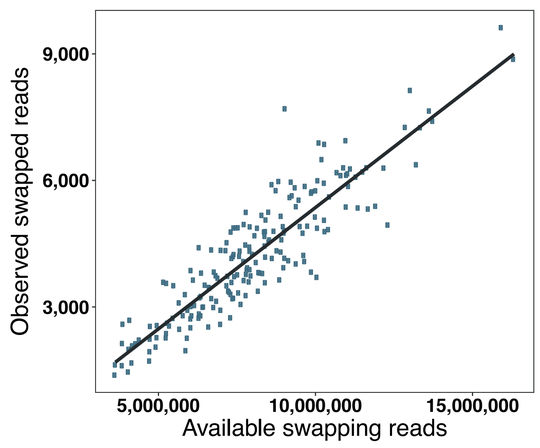

Figure 7 The number of "impossible" barcode reads is proportional to the number of reads of real cells that share a single barcode (see **Figure 6**). We estimated from this gradient a swapping rate of 2.19±0.08%.

• In a multiplexed experiment (**Figure 5**), we observed that **1% of reads** derived from "impossible" barcode combinations.

• Did these reads arise from barcode swapping?

Our assumption appeared to be correct (**Figure 7**). Using the gradient of this relationship, we estimated swapping rates of:

- **2.19±0.08% on HiSeq 4000**
- **0.22±0.01% on HiSeq 2500**

Acknowledgements

Our complete analysis is available at:
`github.com/MarioniLab/BarcodeSwapping2017`

Thanks to:
Göttgens lab, for data access assistance.
Sinha et al., for publicising barcode swapping.