Title:

Estimating and correcting index hopping misassignments in single-cell RNA-seq data

Short title:

Index hopping correction in single-cell RNA-seq data

Authors and affiliations:

Lingling Miao[1], Loren Collado[1], Savannah Barkdull[1], Yoshine Saito[1], Jay-Hyun Jo[1], Jungmin Han[1], Stefania Dell'Orso[2], Michael C. Kelly[3], Sean Conlan[4], Heidi H. Kong[1], and Isaac Brownell[1]

1 Dermatology Branch, NIAMS, NIH, Bethesda, Maryland, USA

2 Genomic Technology Section, NIAMS, NIH, Bethesda, Maryland, USA

3 Single Cell Analysis Facility, Cancer Research Technology Program, Frederick National Laboratory, Bethesda, Maryland, USA

4 Microbial Genomics Section, Translational and Functional Genomics Branch, NHGRI, NIH, Bethesda, Maryland, USA

Corresponding author:

Isaac Brownell, Isaac.brownell@nih.gov

ABSTRACT

Index hopping causes read assignment errors in data from multiplexed sequencing libraries. This issue has become more prevalent with the widespread use of high-capacity sequencers and highly multiplexed single-cell RNA sequencing (scRNA-seq) libraries. We conducted deep, plate-based scRNA-seq on a mixed population of mouse skin cells. Analysis of transcriptomes from 1152 cells identified four distinct cell types. To estimate the error rate in sample assignment due to index hopping, we employed differential expression analysis to identify signature genes that were highly and specifically expressed in each cell type. We quantified the proportion of misassigned reads by examining the detection rates of signature genes in other cell types. Remarkably, regardless of gene expression levels, we estimated that 0.65% of reads per gene were assigned to incorrect cell across our data. To computationally compensate for index hopping, we developed a simple correction method wherein, for each gene, 0.65% of the library's average expression level was subtracted from the expression of each cell. This correction had notable effects on transcriptome analyses, including increased cell-cell clustering distance and alterations in intermediate state assignments of cell differentiation. These findings underscore the potential impact of index hopping on experimental results. In conclusion, we devised a straightforward method to estimate and correct for the index hopping rate by quantifying misassigned genes in distinct cell types within an scRNA-seq library. This approach can be applied to any barcoded, multiplexed scRNA-seq library containing cells with distinct expression profiles, allowing for correction of the expression matrix before conducting biological analysis.

AUTHOR SUMMARY

Index hopping causes errors in multiplexed high-throughput sequencing data whereby inaccurate barcoding leads to the misassignment of sequence reads to an incorrect source. This can result in gene expression assignments to the wrong cell in single-cell RNA sequencing (scRNA-seq). We performed scRNA-seq on sorted mouse skin cells and identified four distinct cell types based on gene expression profiles. Using genes unique to each cell type, we discovered that approximately 0.65% of total reads per gene were incorrectly assigned to another cell due to index hopping, regardless of gene expression levels. To correct for the misassigned reads, we proportionally adjusted each cell's gene expression data. Applying this correction improved analyses allowing for enhanced cell clustering accuracy and refining the identification of intermediate cell states during cell differentiation. Our study emphasizes the significance of index hopping in scRNA-seq experiments and presents a practical correction approach to remove misassigned reads that can be applied to any scRNA-seq study involving cells with distinct gene expression profiles or where non-expressed genes are introduced as a quality control measure during library preparation.

INTRODUCTION

The advent of next-generation sequencing (NGS) has revolutionized genomic research, with contemporary sequencers allowing cost-effective sequencing of large-scale multiplexed libraries [1]. Multiplexing is a common strategy for sequencing multiple samples that involves adding a unique short DNA sequence (index) to each fragment in a sample and pooling samples together for sequencing. However, a notable limitation of

this approach is index misassignment, where sequence reads are erroneously paired with an index from another sample. Index misassignment can occur at various stages during NGS including reagent contamination, sample handling, PCR amplification, patterned flow cell "spread-of-signal", and bioinformatic errors [2, 3]. Among these mechanisms, a major contributor to index hopping (also known as index swapping or index switching), is incomplete library cleanup which allows free-floating index primers to randomly amplify DNA fragments from multiple samples within a pooled library during sequencing [4]. This phenomenon leads to data contamination through index misassignment. The impact of index hopping has grown significantly as high capacity modern sequencing technologies increasingly rely on multiplexing samples. This issue is particularly evident in single-cell RNA sequencing (scRNA-seq) studies that involve pooling large numbers of cDNA libraries for sequencing [4-7].

Index hopping poses a challenge in scRNA-seq, impacting both droplet and plate-based methods [4-7]. Recent investigations analyzing multiplexed droplet-based scRNA-seq datasets estimated sample index hopping probabilities ranging from 0.3% to 0.9% [4]. Plate-based scRNA-seq studies utilizing the "i5+i7" dual-index multiplexing strategy reported more variable and higher index hopping rates, ranging from 0.25% to 10% on Hiseq platforms [5-7]. The index hopping rate varies across different Illumina platforms, with advanced Hiseq systems, such as Hiseq 3000, which employ patterned flow cells and exclusion amplification chemistry, exhibiting higher index hopping levels compared to earlier platforms like Hiseq 2500 [3-5, 8, 9]. Illumina recommends reducing free index primers or the use of unique dual indexes as experimental approaches to mitigate index hopping when using dual-index multiplexing strategies [10]. However, unique dual

96 indexing may not be practical with large samples due to the limitation in barcode
97 numbers, especially in high-volume scRNA-seq systems. In addition, unique indexing is
98 not a complete solution as index hopping has been observed in droplet-based scRNA-
99 seq and multiplexed long-read sequencing which use unique sample barcodes [4, 5,
100 11]. A few recent studies have explored using computational algorithm to estimate and
101 correct the contamination effect of index hopping in plated-based and droplet-based
102 scRNA-seq data [4-6]. However, their approach requires knowing the ground truth of
103 non-multiplexed sequencing.
104
105 Given the limitations of existing strategies, there is a need for an objective method to
106 estimate and correct for index hopping in scRNA-seq data. In this study, we performed
107 an in-depth analysis of plate-based scRNA-seq data obtained from mouse skin cells.
108 We developed a simplified method that leverages distinct expression profiles among
109 cells to estimate the index hopping rate and implemented a correction approach to
110 adjust for the contamination caused by index hopping. We further demonstrate the
111 impact of correcting for index hopping before conducting biological analysis of scRNA-
112 seq data. Our findings contribute to advancing the understanding of index hopping and
113 provide a practical solution for its estimation and correction in scRNA-seq data analysis.
114

115 ## RESULTS

116 **Index hopping rates can be estimated by the frequency of non-self genes in**
117 **scRNA-seq data**

118 We performed plate-based scRNA-seq on FACS sorted dissociated mouse skin cells.
119 The skin is a complex tissue that encompasses multiple cell types including, but not

120 limited to, Merkel cells, dermal papilla fibroblasts, hematopoietic cells, and epithelial
121 keratinocytes (Fig. S1). The transcription factor SRY-Box Transcription Factor 2 (SOX2)
122 is expressed in Merkel cells and dermal papilla cells [12], whereas Growth Factor
123 Independence 1 (GFI1) is expressed in Merkel cells and hematopoietic cells [13, 14].
124 We sorted single cells expressing GFP from the skin of *Sox2*$^{GFP/+}$ or *Gfi1*$^{GFP/+}$ reporter
125 mice at various developmental stages, along with GFP-negative epidermal cells
126 expected to be keratinocytes (Fig. 1A). Single cells were sorted into lysis buffer in 96
127 well plates, and the Smart-seq2 method was used to generate cDNA from these sorted
128 individual cells. Cleaved cDNA was tagged with dual index adaptors prior to being
129 pooled into three 384-sample sequencing libraries (Fig. 2A, S2A-C). On average, each
130 cell yielded ~3 million uniquely mapped reads, allowing for the detection of
131 approximately 6000 genes per cell. By employing Seurat clustering, we grouped the
132 individual cells into four clusters based on their transcriptome similarities (Fig. 1B).
133 Furthermore, by identifying highly differentially expressed genes (marker genes) within
134 these four clusters, we were able to distinguish them as four distinct cell types: Merkel
135 cells (MCs), dermal papilla cells (DCs), hematopoietic cells (HCs), and keratinocytes
136 (KCs) (Fig. 1C). As the pooled libraries included all four cell types, index hopping had
137 the potential to misassign reads from one cell type to another based on their sharing of
138 i7 or i5 indexes (Fig. 2A, S2A-B).
139

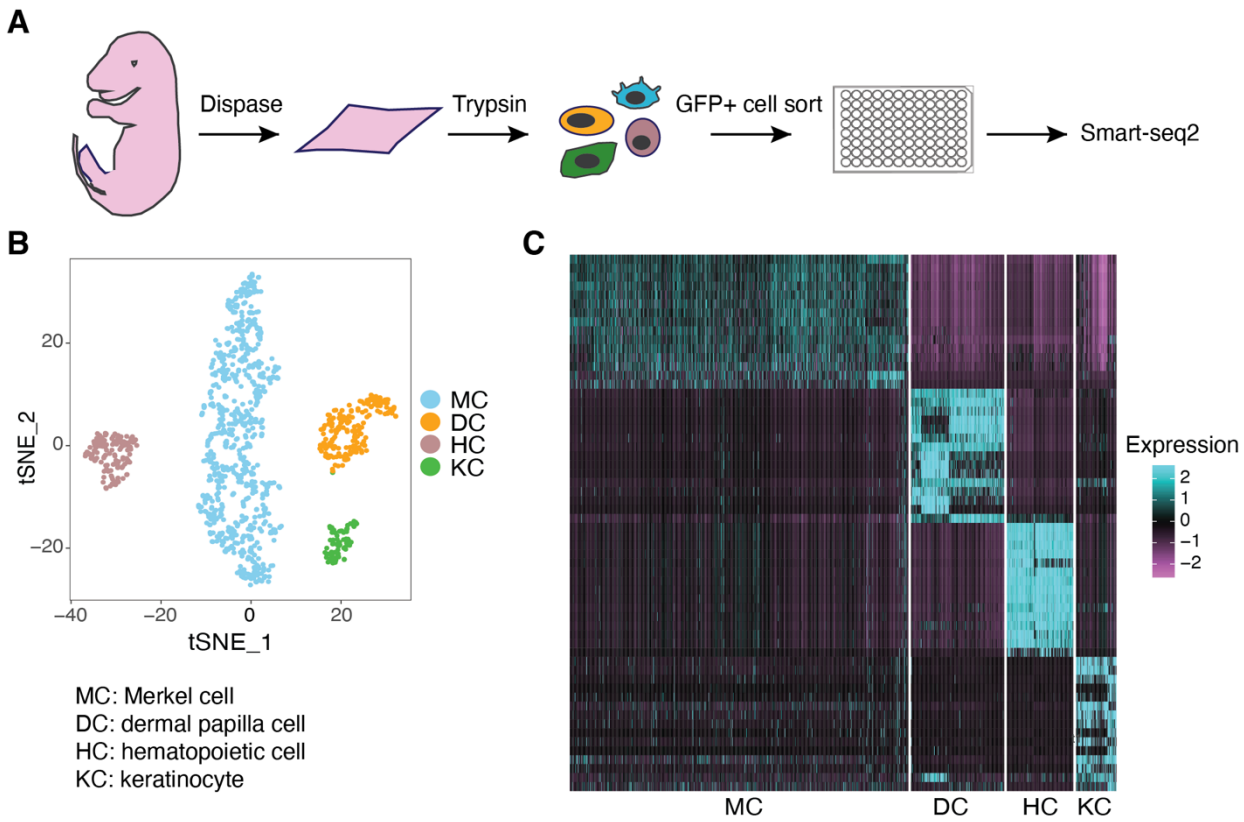


Figure 1. scRNA-seq identifies different skin cell types.

(A) Overview of scRNA-seq experimental procedure.

(B) tSNE plot showing clustering of the gene expression profiles of single skin cells.

(C) Heatmap showing highly differentially expressed genes (marker genes) for each cell cluster identified by Seurat clustering.

Index hopping occurs when cDNA from one cell is indexed erroneously with another cell's barcodes. After library pooling, unincorporated free index primers associated with a row or column on the barcoding plate can potentially be added to cDNA from another row/column creating the risk for index hopping and misassignments of reads. Detection of such an event is difficult when the two wells contain similar transcriptomes. However, if the two cells are of different lineages and have distinct gene expression, index hopping contamination can be measured by detecting non-self genes (transcript not expressed in that cell type) in a specific cell or from an empty well in plated-based

155 sequencing libraries. To identify non-self genes, we used the differentially expressed
156 marker genes for each of the four cell types in our library (Fig. 1C). We were able to
157 detect marker genes for each of the four cell types labeled with indexes for empty wells.
158 We then calculated the average library expression levels for all genes by dividing the
159 total transcripts per million (TPM) for each gene in the library by the number of cells in
160 the library. Similarly, we calculated the average detection levels for each gene in empty
161 wells by dividing the total TPM from the empty wells by the number of empty wells.
162 Comparing the average gene detection levels in the library and in empty wells, we
163 observed a positive correlation (Fig. 2B, S2D), indicating that misassignments were
164 likely a result of a stochastic process like index hopping rather than technical errors in
165 sample preparation.
166

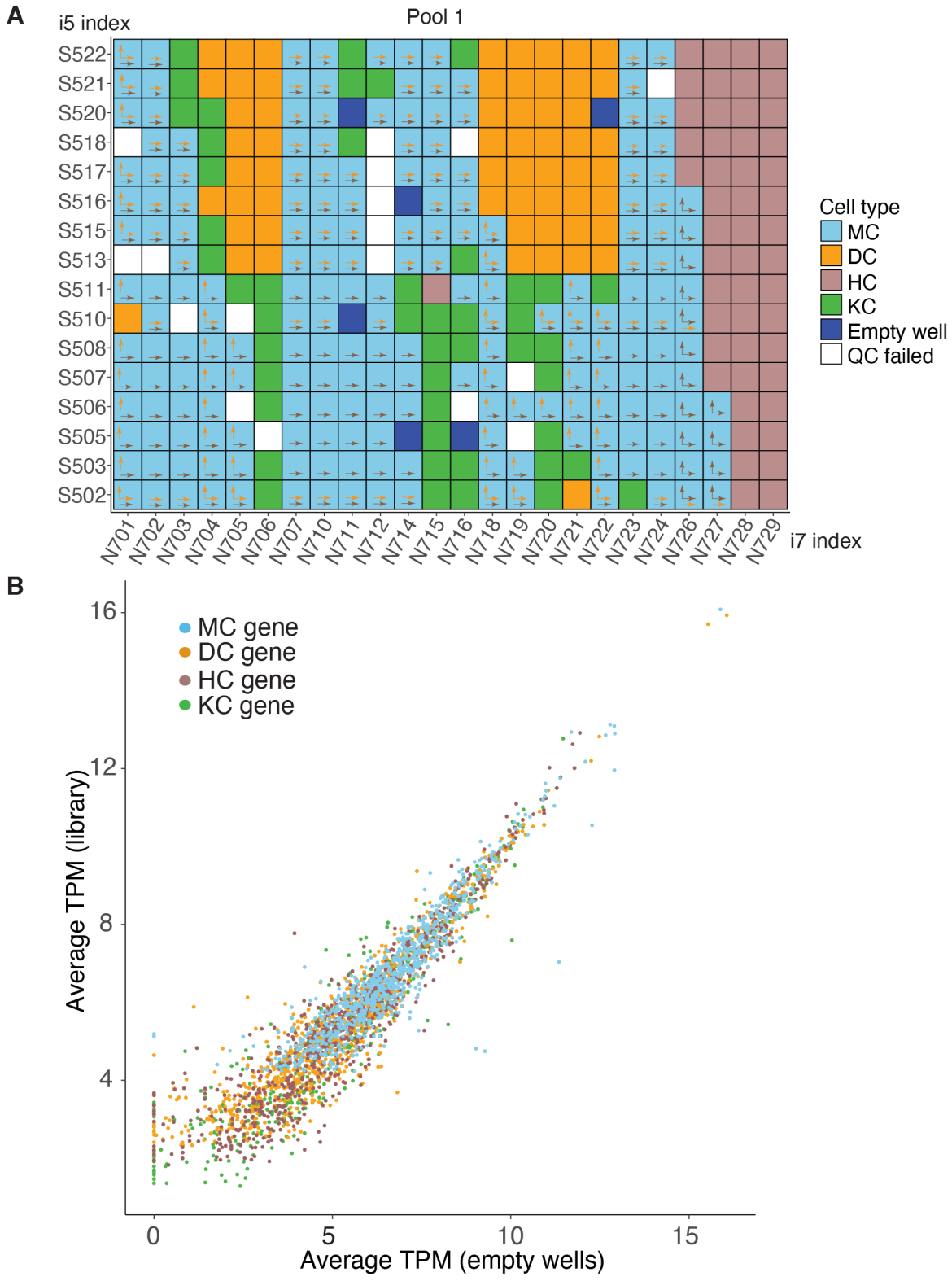


Figure 2. Index hopping can be detected when occurring between differing cell types.

(A) Overview of the indexing used for 384 single-cell cDNA libraries in the first pooled sequencing library. Cell types together with empty wells and quality control (QC) failed wells are color-coded. Arrows show for each individual Merkel cells (MCs), the potential index hopping with dermal cells (DCs, orange arrow) or hematopoietic cells (HCs, brown arrow). Vertical arrows show i7 index hopping, and horizontal arrows show i5 index hopping.

(B) Scatter plot showing the correlation of average TPM of cell type marker genes in empty wells and the entire library, with values transformed using a log2 scale (log2(average TPM +1). Color labels marker genes for the 4 cell types.

To assess index hopping in our library, we quantified non-self gene detection in the four cell types. Considering the location of single cell samples within the multi-well plates and the application of dual index primer sets in our libraries, there was the potential for dual index hopping (misassignment can happen by either i5 or i7 hopping), single index hopping (misassignment can happen only by i5 or i7 hopping), and no potential index hopping between DCs and MCs, while dual index hopping and i5 index hopping were possible between HC and MC (Fig. 2A, S2A-B). To estimate the rate of read misassignments, we calculated the following:

Firstly, we computed the average transcripts per million (TPM) of a gene in MCs by dividing the total TPM for that gene in MCs by the number of MCs (gene A as an example, $\overline{TPM}_{geneA,MC}$). The data were shown on a log2 scale (log2($\overline{TPM}_{geneA,MC} + 1$).

$$\overline{TPM}_{geneA,MC} = \frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} TPM_i$$

$\overline{TPM}_{geneA,MC}$: average TPM for gene A in MC

n_{MC} : number of single MCs

Then, to determine the ratio of reads assigned to individual MCs (detection ratio) for each gene (gene A as an example, $\varphi_{geneA,cellB:lib}$), we divided the TPM of each gene in single MC by its average TPM in the library.

$$\overline{TPM}_{geneA,lib} = \frac{1}{n_{lib}} \sum_{i=1}^{n_{lib}} TPM_i$$

$\overline{TPM}_{geneA,lib}$: average TPM for gene A in sequencing library

n_{lib} : number of single cells in library

$$\varphi_{geneA,cellB:lib} = \frac{TPM_{geneA,cellB}}{\overline{TPM}_{geneA,lib}}$$

$\varphi_{geneA,cellB:lib}$: for gene A, the detection ratio between single cell B and average expression in library

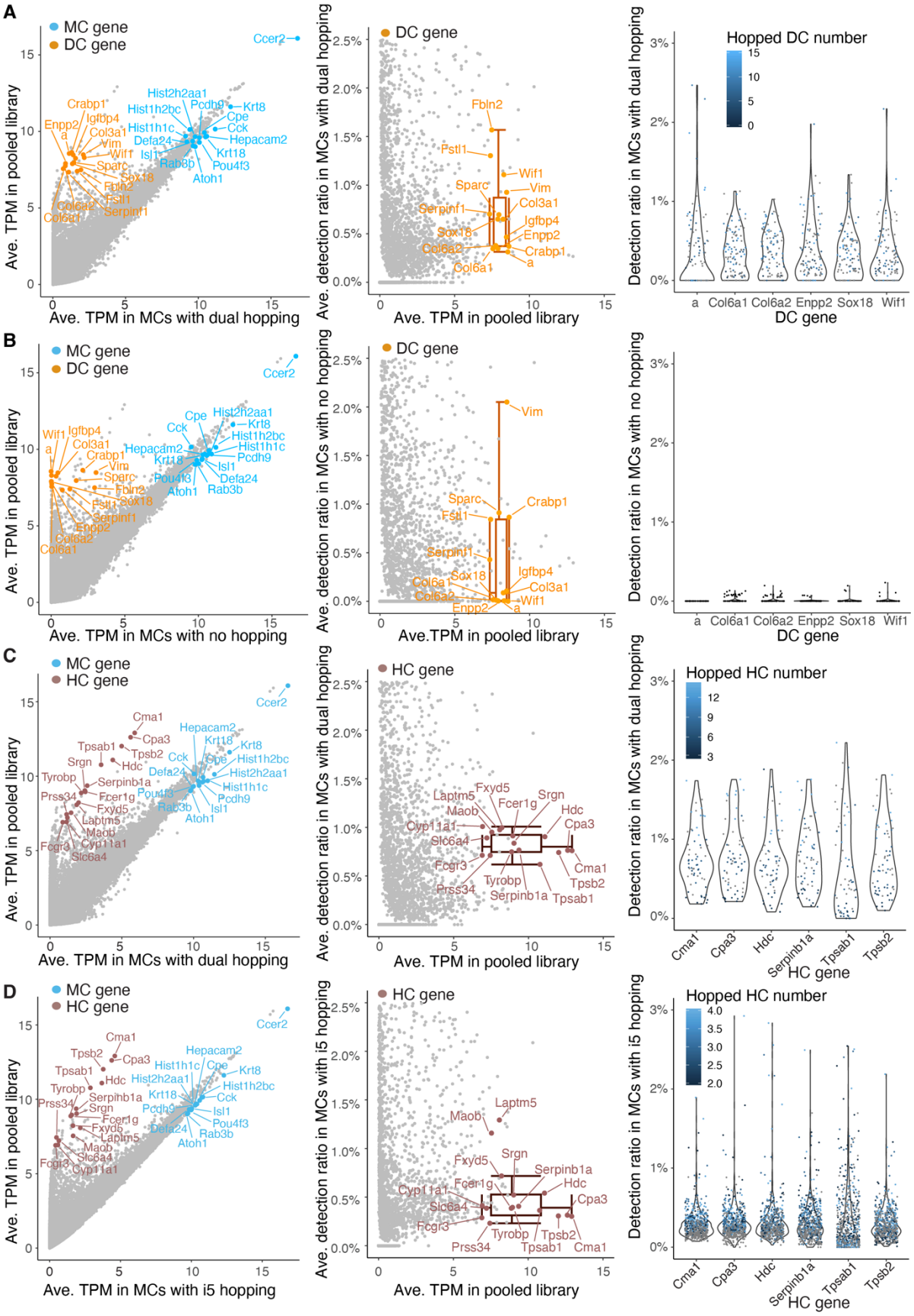
And we were able to calculate the average detection ratio of each gene between MCs and the entire library (gene A as an example, $\overline{\varphi}_{geneA,MC:lib}$).

$$\overline{\varphi}_{geneA,MC:lib} = \frac{1}{n_{MC}} \sum_{i=1}^{n_{MC}} \varphi_{geneA,cell_i:lib}$$

$\overline{\varphi}_{geneA,MC:lib}$: for gene A, the average detection ratio between MCs and average expression in library

n_{MC} : number of single MCs

If index hopping played a significant role in misassignment, the detection ratio of non-self genes in MCs would correspond to the index hopping rate. Notably, the detection ratio of top DC genes in MCs was higher in MCs with dual index hopping from DCs compared to those with single index hopping from DCs. Conversely, when no index hopping occurred between MC and DC, the detection ratio of DC genes in MCs was nearly zero (Fig. 3A-B, S3A-B). Similarly, the detection ratio of top HC genes in MCs with dual index hopping from HCs was higher than those with i5 index hopping from HCs (Fig. 3C, D). Furthermore, the detection ratio of DC and HC genes in MCs exhibited a positive correlation with the number of index-hopped DCs (Fig. S3C) and HCs (Fig. S3D), respectively.


A
MC gene
DC gene
Ave. TPM in pooled library
Ave. TPM in MCs with dual hopping
Ave. detection ratio in MCs with dual hopping
Ave. TPM in pooled library
Hopped DC number
Detection ratio in MCs with dual hopping
DC gene
B
MC gene
DC gene
Ave. TPM in pooled library
Ave. TPM in MCs with no hopping
Ave. detection ratio in MCs with no hopping
Ave.TPM in pooled library
Detection ratio in MCs with no hopping
DC gene
C
MC gene
HC gene
Ave. TPM in pooled library
Ave. TPM in MCs with dual hopping
Ave. detection ratio in MCs with dual hopping
Ave. TPM in pooled library
Hopped HC number
Detection ratio in MCs with dual hopping
HC gene
D
MC gene
HC gene
Ave. TPM in pooled library
Ave. TPM in MCs with i5 hopping
Ave. detection ratio in MCs with i5 hopping
Ave. TPM in pooled library
Hopped HC number
Detection ratio in MCs with i5 hopping
HC gene

218

Figure 3. Non-self marker genes are detected at different levels in MCs with dual, single and no index hopping to other cell types.

(A) and **(B)** MCs which had dual (A) and no (B) index hopping potential from DCs: Left, scatter plot showing the average TPM for individual genes in single MCs correlated with their average TPM in the library. TPMs are displayed on a log2 scale (log2 (average TPM + 1)). Middle, scatter plot showing the average detection ratios of DC genes in single MCs. Colored labels indicate the top marker genes for different cell types. Box plots show the median and interquartile range ± SD of average detection ratio for color-labeled genes. Ave., average. Right, violin plot showing the detection ratio of the top 6 DC marker genes in individual MCs

(C) and **(D)** MCs which had dual (C) and i5 (D) index hopping potential from HCs: Left, scatter plot showing the average TPM in single MCs correlated with their average TPM in the library. TPMs are displayed on a log2 scale (log2 (average TPM + 1)). Middle, scatter plot showing the average detection ratio of HC genes in single MCs. Colored labels indicate the top marker genes for different cell types. Box plots show the median and interquartile range ± SD of average detection ratio for color-labeled genes. Ave., average. Right, violin plot showing the detection ratio of the top 6 HC marker genes in individual MCs.

Adjusting gene expression levels by the estimated index hopping rate compensates for misassigned reads

Having observed that index hopping was relatively consistent, we sought to correct for its effect. A group of highly-expressed HC and DC marker genes with no expected expression in MCs was detected in MCs at levels consistent with their average TPM in the library (Fig. 4A). We referred to this distinctive group of genes with their misassigned reads in MCs as the “side population”. The detection ratios of these genes in MCs ($\overline{\varphi}$) were consistently below 1%, regardless of their varying expression levels in the libraries (Fig. 4B). Notably, there were other highly-expressed HC and DC marker genes that appeared to have native expression in MC as indicated by their detection ratios being greater than 1%. These non-specific marker genes clustered away from the side population genes (Fig. 4A-B). To estimate average index hopping rates in MC, we employed a methodology inspired by the Rank Ordering of Super-Enhancers (ROSE) tool for super-enhancer identification from ChIP-seq data [15]. We first ranked and

248 plotted all genes increasingly by the reciprocal of their average gene detection ratio ($\overline{\varphi}$)
249 in MCs. This plot demonstrated a distinct inflection point where the reciprocal of the
250 detection ratio in MCs began increasing rapidly. To geometrically define this point, we
251 rescaled the data so that the x and y axes ranged from 0 to 1. Next, we determined
252 where a tangent to the rank curve had a slope of 1. The gene detection ratio at this
253 point was calculated to be 0.65%, which we used as the estimated index hopping rate in
254 our sequencing library. Genes ranked below this threshold were classified as
255 predominantly MC-expressed genes, whereas genes ranked above this threshold were
256 considered as genes whose detection in MCs was due, in part, to index hopping
257 misassignment (Fig. 4C). This classification was validated by identifying the rank order
258 position of the differentially expressed marker genes for each cell type (Fig. S4A-D).
259 Notably, genes in the side population of top HC and DC genes were ranked well above
260 the inflection point, and the non-specific marker genes ranked near the inflection point
261 (Fig. 4D). Consistent with MC and KC having shared epithelial progenitors and similar
262 transcriptomes, the detection ratio of KC marker genes in MCs was typically higher than
263 the index hopping rate reflecting the relative lack of specific KC marker genes (Fig. S5A,
264 B). Following this pattern, in cases where marker genes appeared to be specific to the
265 two cell types being compared, a distinct side population was observed when visualizing
266 average expression levels and detection ratios (Fig. S5C-H).
267

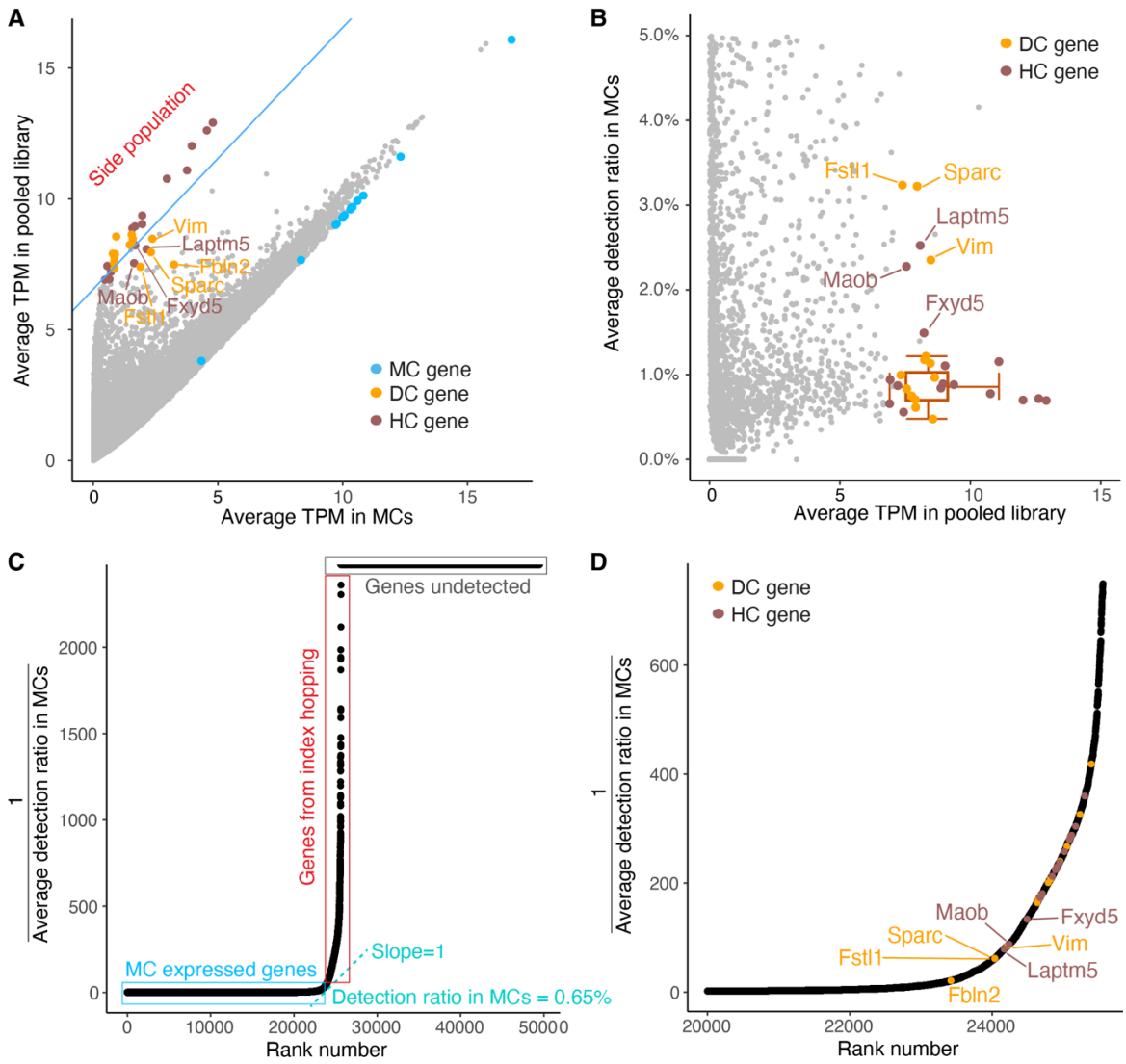


Figure 4. Index hopping rate can be estimated using non-self genes

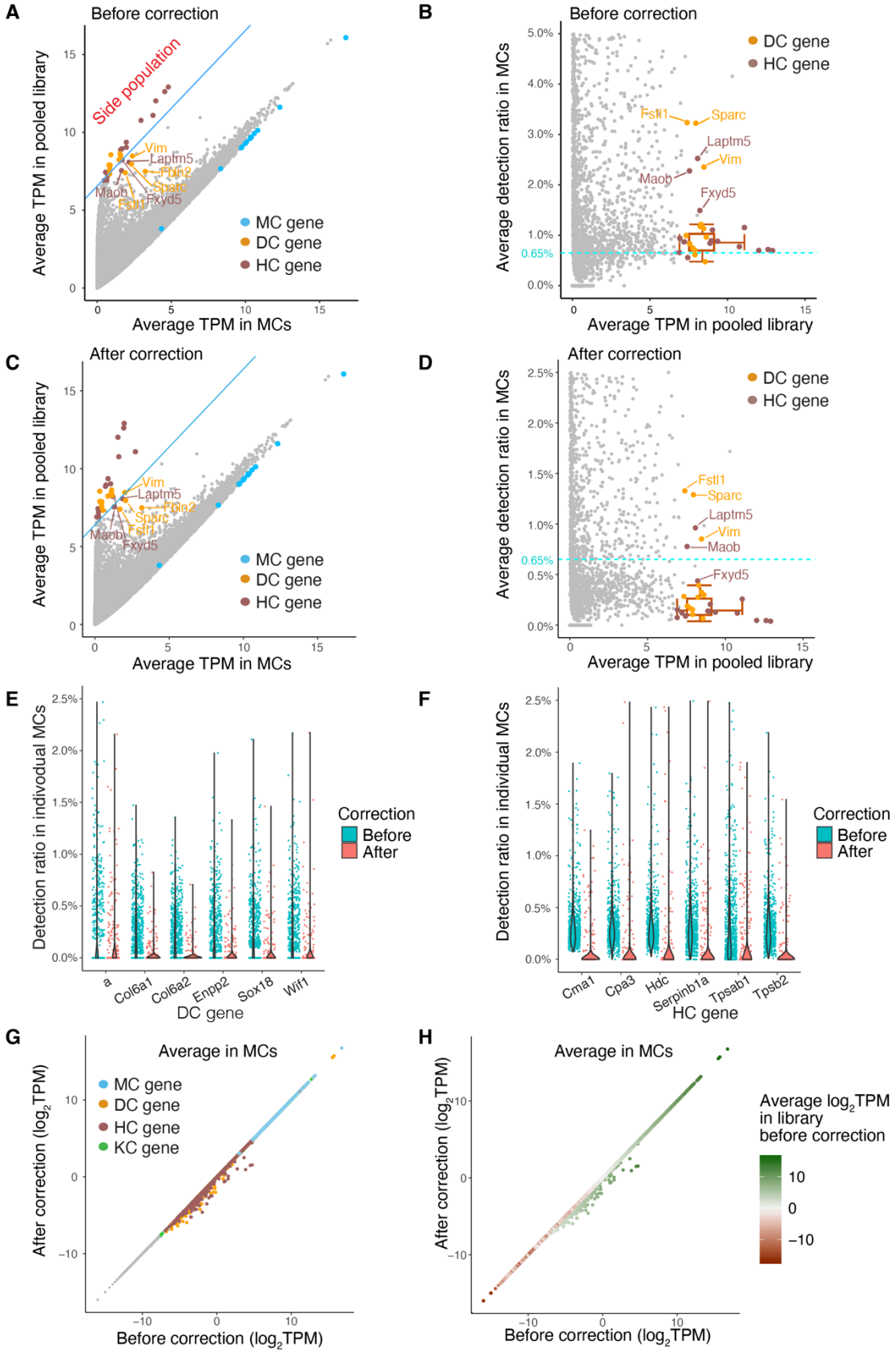
(A) Scatter plot showing the average TPM of top HC and DC genes in single MCs correlated to their average TPM in library. Blue line separates highly-expressed HC and DC "side" population genes where the level in MCs is proportional to their average TPM in the library. TPMs are displayed on a log2 scale (log2 (average TPM + 1)). Colors indicate the top MC, HC and DC marker genes.

(B) Scatter plot showing the average gene detection ratios in single MCs. Box plots show median and interquartile range ± SD of average detection ratio in MCs for side population genes. Colors indicate the top HC and DC marker genes.

(C) Rank ordered scatter plot showing distribution of gene detection ratios in MCs across all genes from mm10 genome annotated by Ensembl. Blue box indicates the MC expressed genes, red box indicates genes presumptively detected due to index hopping, grey box indicates genes with no detection in MCs. The cyan tangent line with slope = 1 identifies the inflection point.

(D) Focused rank ordered scatter plot showing genes near the inflection point from plot in (C).

282 To address the background gene detection resulting from index hopping, we
283 implemented a simple correction method. We subtracted 0.65% of the average
284 transcripts per million (TPM) for each gene in our library from the single-cell gene TPM
285 matrix. Following the correction, the detection ratios of non-self genes in MCs (Fig. 5A-
286 F) and other cell types (Fig. S5A'-H') was notably reduced. Additionally, we investigated
287 the impact of the correction on individual gene TPM values in MCs. As expected, highly
288 expressed HC and DC marker genes showed notably reduced detection in MCs after
289 correction, whereas the reduction in MC marker genes in MCs was inconsequential
290 (Fig. 5G-H, S6A-D). Similarly, when applying the correction in HCs, only non-self genes
291 with high library TPM values were notably impacted, whereas HC marker genes
292 remained comparable to pre-correction levels in HCs (Fig. S6E-J).
293


A
Before correction
Side population
Average TPM in pooled library
Average TPM in MCs
Vim
Laptm5
Fbln2
Sparc
Maob
Fstl1
Fxyd5
MC gene
DC gene
HC gene
B
Before correction
Average detection ratio in MCs
Average TPM in pooled library
DC gene
HC gene
Fstl1
Sparc
Laptm5
Vim
Maob
Fxyd5
0.65%
C
After correction
Average TPM in pooled library
Average TPM in MCs
MC gene
DC gene
HC gene
D
After correction
Average detection ratio in MCs
Average TPM in pooled library
DC gene
HC gene
0.65%
E
Detection ratio in indivodual MCs
Correction
Before
After
a
Col6a1
Col6a2
Enpp2
Sox18
Wif1
DC gene
F
Detection ratio in individual MCs
Correction
Before
After
Cma1
Cpa3
Hdc
Serpinb1a
Tpsab1
Tpsb2
HC gene
G
Average in MCs
MC gene
DC gene
HC gene
KC gene
After correction (log2TPM)
Before correction (log2TPM)
H
Average in MCs
Average log2TPM in library before correction
After correction (log2TPM)
Before correction (log2TPM)

294

295 **Figure 5. Index hopping correction reduces the detection of non-self marker genes.**
296 **(A)** and **(B)** Same as Figure 4A and B – precorrection scatter plots for comparison.
297 **(C)** and **(D)** Scatter plots showing the average TPM (C) and detection ratios (D) of individual genes in single
298 MCs after index hopping correction. Colored labels indicate the top marker genes for different cell types.
299 The blue line in (C) is same as the one in (A) which separates the side population. The cyan dashed line in
300 (B and D) represents the estimated index hopping rate before correction (0.65%). TPMs are displayed on
301 a log2 scale (log2 (average TPM + 1)). Box plots depict the median and interquartile range ± SD of average
302 detection ratio for the side population genes after correction.
303 **(E)** and **(F)** Violin plots showing the detection ratios of the top 6 DC (E) and top 6 HC (F) marker genes in
304 MCs before and after correction.
305 **(G)** and **(H)** Scatter plots showing the log2 scaled average TPM of individual genes in single MCs before
306 and after correction. Colors indicate the marker genes in different cell types (G) and represent levels of
307 average TPM in library (H).
308

309 **Correcting for index hopping impacts single-cell analysis and interpretation**

310 To assess the impact of index hopping correction on the bioinformatic analysis of the
311 scRNA-seq data, we analyzed our data set with and without correction. To evaluate the
312 heterogeneity among single cells, we employed cell-to-cell distance, which quantifies
313 transcriptomic similarities between cells. A smaller cell distance indicates a higher level
314 of similarity between two cells [16]. We compared the cell distances among all cells in
315 the library before and after correction. As misassigned read contamination would be
316 expected to artificially make distinct cell types more similar, we anticipated that index
317 hopping correction would increase distances between cells. The density plot of cell
318 distances demonstrated that index hopping correction did indeed increase measured
319 heterogeneity within the DC and HC clusters (Fig. 6A, B). This suggests that index
320 hopping correction can mitigate artificial cell-cell similarities caused by randomly
321 misassigned reads across scRNA-seq libraries.
322

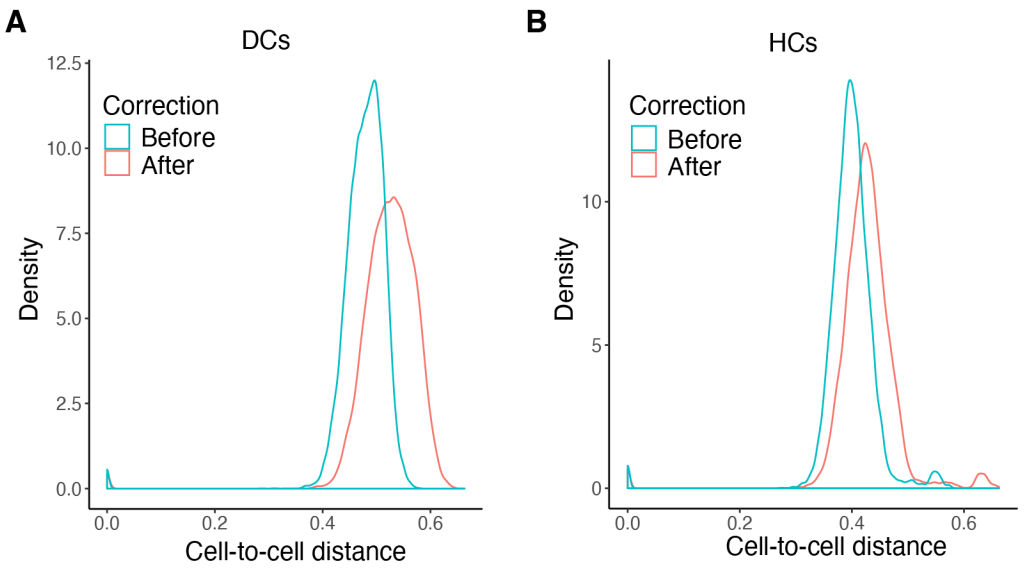
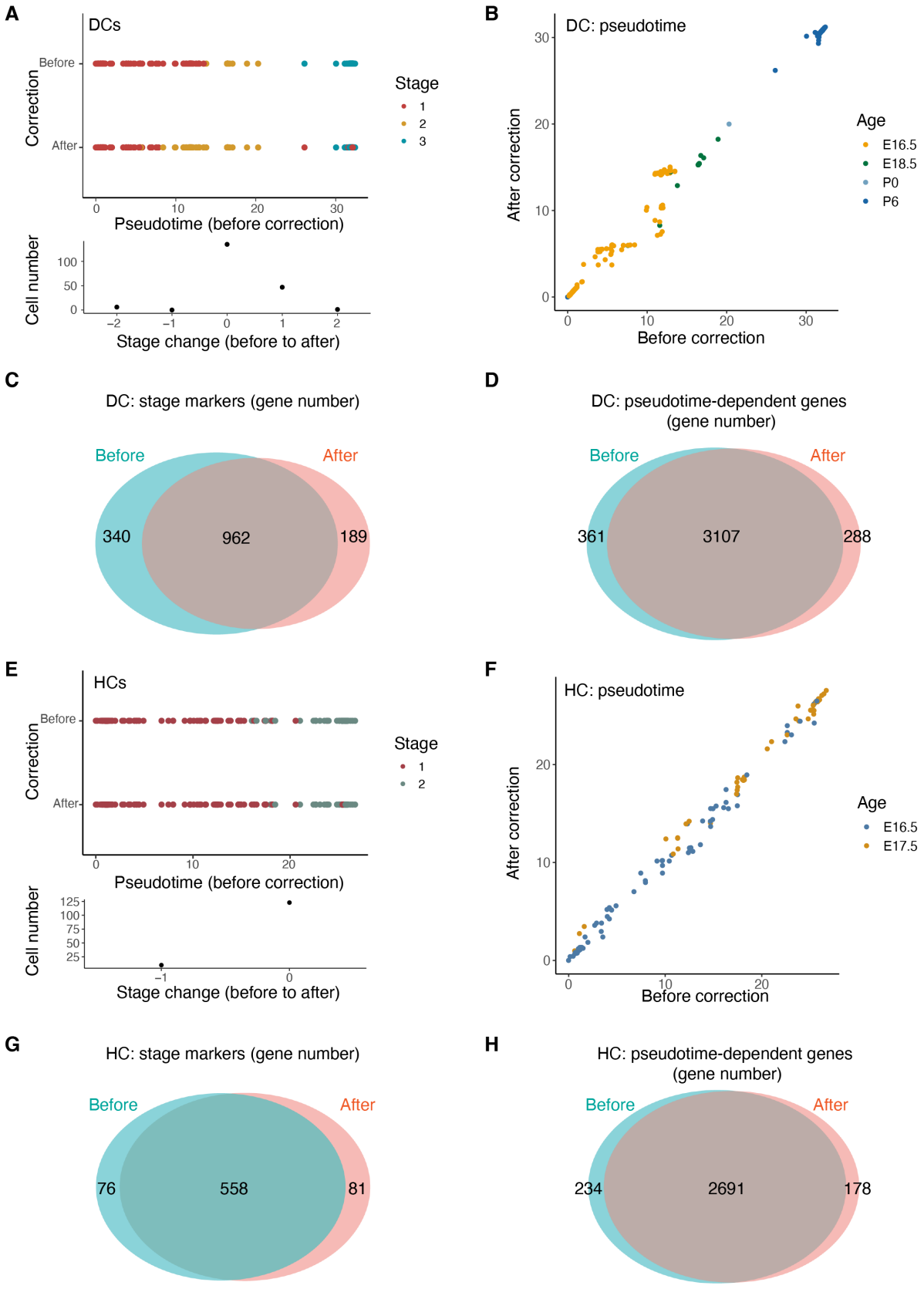


Figure 6. Index hopping correction increases the detection of single cell heterogeneity. Density plots showing the distribution of cell-to-cell distance among individual DCs (A) and HCs (B) before and after correction. Kolmogorov-Smirnov test: $p < 2.2 \times 10^{-16}$ for cell-to-cell distance before correction vs. after correction.

Next, we investigated how index hopping correction impacted the analysis of developmental stage classification. Our scRNA-seq dataset included single cells spanning a wide range of developmental timepoints, from embryonic to postnatal skin. Focusing on DCs, we subjected uncorrected and corrected data to Seurat clustering identifying three cell clusters (Fig. S7A, E). Consistent with the cells' chronological age, each cluster represented a distinct stage of skin development (Fig. S7B, F). We then used Monocle trajectory analysis to plot the single DCs along a developmental pseudotime before (Fig. S7C, D) and after (Fig. S7G, H) index hopping correction. Comparing the developmental stage classification of each DC before and after correction, we observed that 79.9% of DC cells retained their stage classification whereas 20.1% of cells were reclassified after correction (Fig. 7A). Consistent with stage reclassifications, there was a reordering of the DC pseudotime trajectory after

index hopping correction (Fig. 7B), suggesting that index hopping misassignments had impacted the results of both cell clustering and state hierarchy analyses.


A
DCs
Correction
Before
After
Pseudotime (before correction)
Cell number
Stage change (before to after)
Stage
1
2
3
B
DC: pseudotime
After correction
Before correction
Age
E16.5
E18.5
P0
P6
C
DC: stage markers (gene number)
Before
After
340
962
189
D
DC: pseudotime-dependent genes (gene number)
Before
After
361
3107
288
E
HCs
Correction
Before
After
Pseudotime (before correction)
Cell number
Stage change (before to after)
Stage
1
2
F
HC: pseudotime
After correction
Before correction
Age
E16.5
E17.5
G
HC: stage markers (gene number)
Before
After
76
558
81
H
HC: pseudotime-dependent genes (gene number)
Before
After
234
2691
178

344

345 **Figure 7. Index hopping correction impacts the single-cell analysis of cell developmental stages.**
346 **(A)** and **(E)** Upper: Scatter plots showing the cell developmental stages of individual DCs (A) and HCs (E)
347 assigned by Seurat clustering before and after index hopping correction. Lower: Histogram showing the
348 frequency of cell stage assignment changes from before correction to after correction.
349 **(B)** and **(F)** Scatter plots showing the pseudotime order of individual DCs (B) and HCs (F) identified by
350 Monocle trajectory analysis before and after correction. Colors indicate the age of mouse source for
351 individual cells.
352 **(C)** and **(G)** Venn diagrams comparing differentially expressed marker genes for any developmental stage
353 of DCs (C) and HCs (G) identified by Seurat before and after correction.
354 **(D)** and **(H)** Venn diagrams comparing pseudotime-dependent genes identified by Monocle for DCs (D) and
355 HCs (H) before and after correction.
356
357 We performed a similar pre and post correction comparison for two age groups of HCs
358 (E16.5 and E17.5). Clustering identified two distinct cell clusters before (Fig. S8A, B)
359 and after (Fig. S8E, F) correction, and trajectory analysis revealed a potential
360 developmental trajectory before (Fig. S8C, D) and after (Fig. S8G, H) correction. Index
361 hopping correction also exhibited an analogous, albeit smaller, effect on the stage
362 classification and pseudotime order of some HCs (Fig. 7E, F).
363
364 Analysis of differentially expressed genes is often used to define transcript signatures
365 for cell clusters whereas identification of genes that are dynamically expressed over a
366 pseudotime trajectory can suggest biologically relevant genes. Accordingly, we explored
367 the impact of index hopping correction on these analyses. Using Seurat and Monocle,
368 we identified differentially expressed developmental stage marker genes as well as
369 genes that significantly changed over the developmental pseudotime for both DCs and
370 HCs. Comparing the aggregate set of stage marker genes and pseudotime-dependent
371 genes before and after index hopping correction, we observed a substantial overlap in
372 each gene set for DCs (Fig. 7C, D) and HCs (Fig. 7G, H). Nonetheless, index hopping

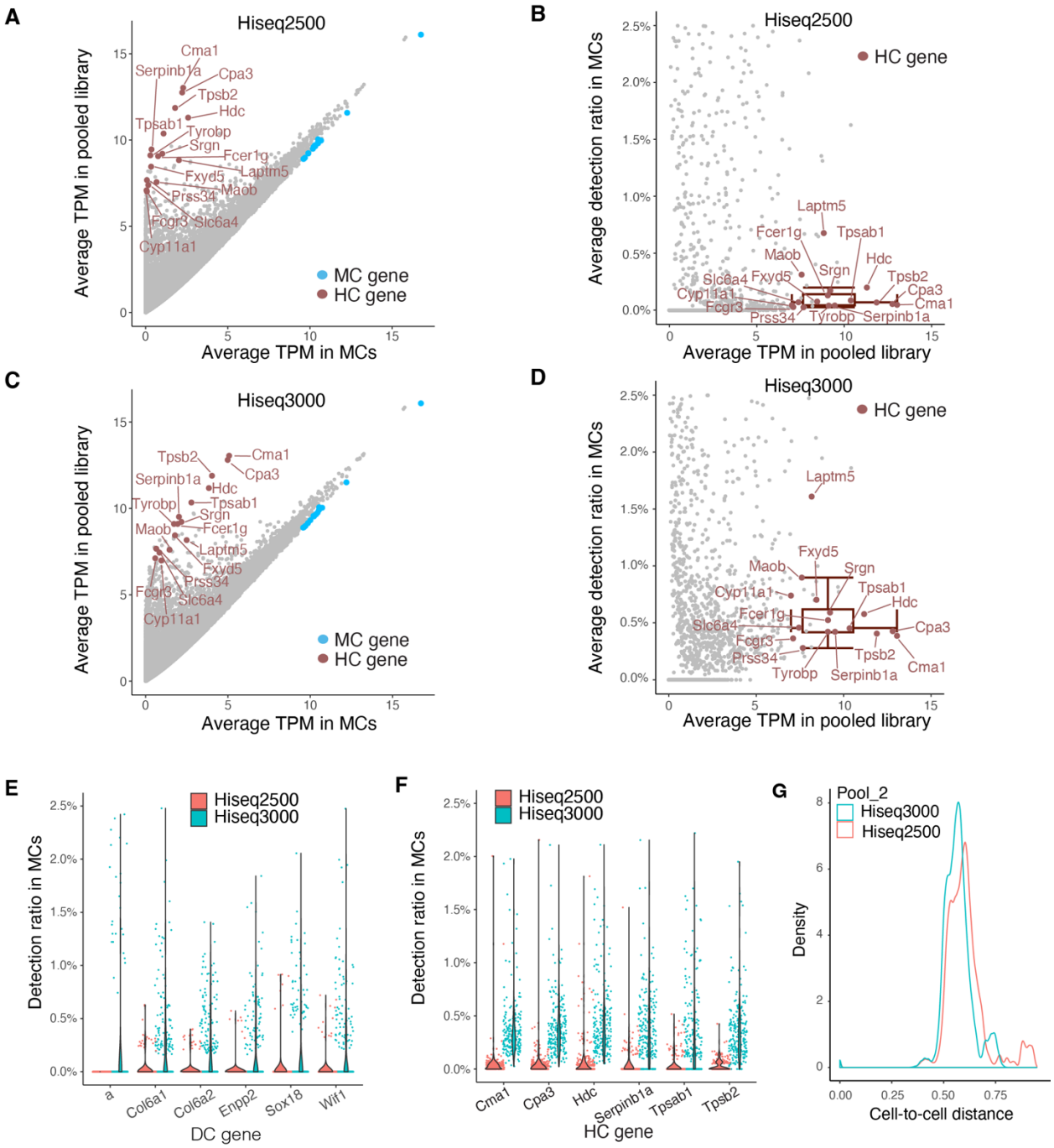
373 correction altered the composition of these gene lists in each case, once again
374 suggesting that index hopping misassignments can impact the gene composition of cell
375 state signatures. Collectively, these results demonstrate that index hopping measurably
376 influences the biological interpretation of scRNA-seq data.
377
378 **Hiseq2500 generates a lower index hopping rate compared to Hiseq3000**
379 As sequencing technology has advanced, the use of patterned flow cells has increased
380 throughput and efficiency. However, sequencing platforms utilizing a patterned flow cell
381 with Exclusion Amplification chemistry, such as the HiSeq3000 and NovaSeq, appear to
382 exhibit higher levels of index hopping compared to platforms without patterned flow cells
383 [3-5, 8, 9]. Our libraries were sequenced using the HiSeq3000 platform. To test if the
384 patterned flow cell technology impacted the degree of index hopping, we re-sequenced
385 one library (Pool #2) on the HiSeq2500, which has a similar sequencing depth but does
386 not employ a patterned flow cell. We compared the gene detection ratios in MCs
387 between the HiSeq2500 and HiSeq3000 platforms. As expected, when analyzing MCs,
388 the detection ratios of genes in the side population, including top HC markers, was
389 reduced on the HiSeq2500 compared to the HiSeq3000, despite the average library
390 TPM remaining unchanged between the two platforms (Fig. 8A-D). The detection ratio
391 of top DC and HC genes in individual MCs was also lower on the HiSeq2500 compared
392 to the HiSeq3000 (Fig. 8E, F). Interestingly, the detection ratio of non-self genes in MCs
393 on the HiSeq2500 was comparable to their HiSeq3000 post-correction levels (Fig. 8A-F,
394 5C-F, S5A'-H'). Similar trends were observed in HCs, where the detection ratios of non-
395 HC genes were lower on the HiSeq2500 compared to the HiSeq3000 (Fig. S9).
396 Furthermore, the cell distance among single cells in Pool #2 was greater on the

397 HiSeq2500 than on the HiSeq3000 (Fig. 8G). These findings confirm that patterned flow

398 cell technologies are more susceptible to index hopping [3-5, 8, 9], and that our

399 correction strategy compensates for this vulnerability.

400



401

402 **Figure 8. Index hopping is less pronounced in Hiseq2500 platform compared to Hiseq3000.**

(A-D) Scatter plots showing the average TPM (A, C) and detection ratio (B, D) without index hopping correction of individual genes in single MCs in library pool #2 sequenced on Hiseq2500 (A, B) and Hiseq3000 (C, D) platforms. TPMs are displayed on a log2 scale (log2 (average TPM + 1)). Colors indicate the top marker genes for different cell types. Box plots depict the median and interquartile range ± SD of average detection ratio for the color-labeled HC genes.

(E) and **(F)** Violin plots showing the detection ratios in MCs of the top 6 DC (E) and HC (F) marker genes using sequencing data from Hiseq2500 and Hiseq3000 platforms without index hopping correction.

(G) Density plot showing the distribution of cell-to-cell distance among single cells in library pool #2 sequenced on Hiseq2500 and Hiseq3000. Kolmogorov-Smirnov test: $p < 2.2 \times 10^{-16}$ for cell-to-cell distance from Hiseq2500 vs. Hiseq3000.

DISCUSSION

In this study, we developed an approach to estimate and correct for index hopping in multiplexed scRNA-seq data. We performed plate-based single cell sequencing of four distinct cell types sorted from mouse skin with a focus on MCs as the most abundant cell type in the libraries. To estimate the rate of read misassignments, we analyzed differentially expressed genes across each cell type. Top genes expected to be expressed exclusively in their respective cell types were designated as marker genes, and the presence of non-self marker genes in a particular cell was interpreted as contamination from index hopping. We calculated the gene detection ratio for each gene in MCs and estimated the index hopping rate based on the ratio of non-self genes. Since the detection ratio of non-self genes in MCs was consistent regardless of their overall detection levels in the library, we assumed a fixed index hopping rate across all samples in the libraries. To correct for index hopping, we reduced the TPM for each gene by the index hopping rate multiplied by the transcript's average expression across the library. The correction significantly reduced read misassignments, as evidenced by the decreased detection ratio of non-self genes. Comparing data obtained from

430 Hiseq3000 and Hiseq2500 sequencing platforms confirmed that index hopping was
431 more pronounced in platforms employing patterned flow cells (e.g., Hiseq3000/4000/X,
432 NextSeq 1000/2000, NovaSeq X/X Plus, and NovaSeq 6000). Furthermore, we
433 evaluated the impact of index hopping correction on biological analysis. Our findings
434 revealed that correction of index hopping mitigated artificial cell-cell similarities,
435 impacted the results of cell clustering and state hierarchy analyses, and influenced the
436 gene composition of cell state signatures. In summary, we propose a simple and
437 practical methodology for estimating and correcting index hopping that appears to
438 improve the accuracy of transcriptome data and the cell biology inferred from its
439 analysis.
440
441 By strategically pooling cells with distinct transcriptomes, we were able to effectively
442 identify and characterize the extent of index hopping. We propose that this method can
443 be adapted to detect index hopping in multiplex sequencing experiments involving both
444 heterogeneous and homogeneous libraries. In the case of a homogeneous cell
445 population, one approach is to introduce single cells from a different cell type as "spike-
446 ins" during the library preparation process. This deliberate addition of distinct cells will
447 introduce non-self genes into the otherwise homogenous cell library allowing for the
448 generation of a side population of non-self genes if index hopping events occur. By
449 leveraging this spike-in strategy, researchers can effectively assess the presence and
450 impact of index hopping contamination, thus enhancing the quality and reliability of
451 scRNA-seq data in various experimental settings.
452

453 Our approach to index hopping has certain limitations. An Illumina white paper
454 demonstrated a correlation between the concentration of free index primers and the rate
455 of index hopping [10]. Variations in the amounts of cDNA and free index primers can
456 produce concentration inconsistencies when multiplexing and pooling libraries. To
457 mitigate this variability, we normalized single-cell cDNA concentrations prior to
458 tagmentation and indexing. However, multiplexed libraries with varied cDNA
459 concentrations per sample may result in differing index hopping rate across libraries.
460 Consequently, estimating a universal index hopping rate for correction may not be
461 possible. Therefore, we recommend normalizing cDNA concentration prior to indexing
462 when possible. In addition, our approach was applied to dual indexed multiplexed
463 libraries. Although unique indexing has been proposed as a strategy to address index
464 hopping, this issue can still arise, especially due to the amplification of free index
465 primers on sequencing platforms with patterned flow cells, as evidenced by the droplet-
466 based scRNA-seq studies [4]. Finally, our comparative analyses of scRNA-seq data
467 before and after correction revealed substantial effects on biological analyses. It is
468 crucial to note that the impact of correction on biological analysis may hinge on the
469 specific index hopping rate within each dataset. Balancing the decision to correct for
470 index hopping is paramount, as over-correction can also introduce false results,
471 particularly in analyses that are sensitive to minor changes in gene expression, such as
472 genes with low expression levels. Each experimental scenario requires a thoughtful
473 assessment to determine the most appropriate course of action when considering index
474 hopping correction.
475

476 In conclusion, our comprehensive analysis of index hopping in plate-based deep
477 scRNA-seq data suggests that having cells with distinct transcriptomes, due to intrinsic
478 diversity or the inclusion of cell "spike-ins" during library preparation, can serve as a
479 valuable strategy to address index hopping data contamination. To compensate for
480 index hopping, we developed a simplified method that enables both the estimation of
481 the index hopping rate and the subsequent correction of read misassignments.
482 However, it is crucial to consider the specific biological questions being addressed by
483 the dataset when making the decision to apply index hopping correction. By employing
484 our approach, researchers can gain a better understanding of the index hopping
485 phenomenon and its impact on scRNA-seq data. This knowledge can help improve the
486 accuracy and reliability of downstream analyses, ultimately leading to more robust and
487 meaningful biological insights.
488

489 ## MATERIALS AND METHODS

490 **Animals**

491 *Sox2*$^{GFP/+}$ [17] and *Gfi1*$^{GFP/+}$ [14] mice were derived from breeders purchased from Jax
492 (#017592 and #016162, respectively). Mice were housed and bred on an outcrossed
493 Swiss Webster background in a pathogen-free facility at NIH, Bethesda, MD. Genotyping
494 of mice was performed by allele-specific PCR on DNA extracted from tail tissue. All
495 experiments were performed in accordance with institutional guidelines according to
496 Institutional Animal Care and Use Committee-approved protocols.

497 **Cell dissociation and sorting**

498 Postnatal day 0 (P0) *Sox2*$^{GFP/+}$ and *Gfi1*$^{GFP/+}$ mice were euthanized, and the skin was
499 harvested, removing the subcutaneous tissue. For embryonic day 18.5 (E18.5) cells, the
500 pregnant *Sox2*$^{GFP/+}$ female mice were euthanized, and the embryos were harvested and
501 euthanized. The skin was harvested from the embryos. The harvested E18.5 or P0 skin
502 was floated in dispase (Corning, #354235) at room temperature for 1 hour to separate the
503 epidermal sheet. The epidermis was chopped into small pieces, and the epidermal cells
504 were dissociated in 0.05% Trypsin at 37 °C for 15 minutes, followed by the addition of
505 DMEM containing 10% FBS. The cells were then sequentially filtered through 70 μm and
506 40 μm cell strainers. Subsequently, the cells were centrifuged at 300 x g for 5 minutes,
507 and the supernatant was discarded. The cell pellets were resuspended and processed
508 for cell sorting using the Beckman MoFlo Astrios or BD influx cell sorter. DAPI was added
509 immediately before cell sorting, and single DAPI^{-}GFP^{+} cells were sorted into 96-well
510 plates pre-loaded with lysis buffer (lysis buffer prepared following the smart-seq2 method
511 [18]). Single DAPI^{-}GFP^{-} cells from P0 epidermis were sorted as controls.
512
513 For embryonic day 16.5 (E16.5) and embryonic day 17.5 (E17.5) cells, pregnant
514 *Sox2*$^{GFP/+}$ (E16.5) and *Gfi1*$^{GFP/+}$ (E16.5 and E17.5) female mice were euthanized, and
515 the embryos were subsequently harvested and euthanized. The skin from the embryos
516 was then harvested, chopped into small pieces, and treated with 0.05% Trypsin at 37 °C
517 for 15 minutes. This was followed by the same cell dissociation procedure used for P0
518 and E18.5 skin, without separating the epidermis.
519
520 Postnatal day 6 (P6) *Sox2*$^{GFP/+}$ and *Gfi1*$^{GFP/+}$ mice were euthanized. The harvested skin
521 was then incubated in 0.15% trypsin at 37°C for 30 minutes. The epidermis was

522 obtained by scraping with a scalpel in 0.05% trypsin. Epidermal cells were then
523 processed using the same method employed for neonatal epidermis.

524 **scRNA-seq**

525 Libraries from isolated single cells were generated following the Smart-seq2 protocol [18]
526 with the following modifications. Prior to tagmentation, the concentration of single-cell
527 cDNA libraries was determined using the Quant-iT PicoGreen dsDNA Assay (Thermo
528 Fisher Scientific). Subsequently, the cDNA libraries were relocated to new plates with
529 normalized concentration. Libraries from each set of 384 cells with unique barcode
530 combinations were pooled together, resulting in each pooled library containing cells from
531 all ages and both reporters. Sequencing was performed on the Hiseq3000 sequencer
532 (Illumina), resulting in three pooled libraries. Additionally, the second pooled libraries
533 underwent repeated sequencing on the Hiseq2500 sequencer (Illumina), achieving a
534 sequencing depth comparable to that of the Hiseq3000. The sequencing generated
535 paired-end reads of either 125bp or 100bp. Subsequently, reads of 125bp were trimmed
536 to 100bp for further processing. The sequencing reads were mapped to the mouse mm10
537 reference genome using STAR 2.5.2b, and gene expression was quantified using RSEM
538 1.3.0. The gene expression matrix with transcript per million values was used for further
539 analysis. Log_2-scaled TPM values (log_2(TPM+1)) were used in all figures where TPM was
540 shown, unless otherwise specified (log_2TPM in Figure 5G, Figure 5H, and Figure S6). For
541 non-self genes in MCs, their average detection ratios in MCs are considered as the index
542 hopping rate.

543 **Cell biology analysis**

544 Cell-to-cell distance was analyzed using the R package 'sincell' based on Spearman
545 correlation [16]. Cell clustering was analyzed using Seurat 2.3 [19]. Cell trajectory was
546 analyzed using Monocle 2 [20, 21].

547 **Statistical analysis**

548 For comparisons of gene detection in single cells between two different conditions (such
549 as before vs. after correction, Hiseq2500 vs. Hiseq3000), paired t-test was applied.

550

551 **DATA AND CODE AVAILABILITY**

552 The scRNA-seq data generated in this study can be accessed in the NCBI Gene
553 Expression Omnibus (GEO) under the accession number GSE213104. All analyses
554 were performed in R using established packages. Scripts utilized in this study can be
555 provided upon request.

556

557 **ACKNOWLEDGEMENTS**

558 We thank Drs. Livak Ferenc and Teresa Hawley from the CCR Flow Cytometry Core, as
559 well as Drs. James Simone, Jeff Lay and Kevin Tinsley from the NIAMS Flow Cytometry
560 Section, for their assistance with cell sorting. We acknowledge Dr. Valery Bliskovsky
561 from the CCR Genomics Core for his help with single-cell cDNA library preparation, and
562 Faiza Naz from the NIMAS Genomic Technology Section for her assistance with
563 sequencing. This work utilized the computational resources of the NIH HPC Biowulf
564 cluster (https://hpc.nih.gov). This research was supported by the NIAMS intramural
565 research program (ZIA AR041221, to IB). Its contents are solely the responsibility of the
566 authors and do not necessarily represent the official views of the NIH.

567

AUTHOR CONTRIBUTIONS

Conceptualization: IB; Methodology: LM, SD, MCK; Data curation: LM, SD; Resources: IB, SD; Formal analysis: LM, JJ, JH, SC; Investigation: LM, LC, SB, YS; Writing - original draft preparation: LM; Writing - review and editing: all authors; Supervision: IB, HK; Funding acquisition: IB.

REFERENCES

1. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155(1):27-38. Epub 2013/10/01. doi: 10.1016/j.cell.2013.09.006. PubMed PMID: 24074859; PubMed Central PMCID: PMCPMC3969849.
2. Li Q, Zhao X, Zhang W, Wang L, Wang J, Xu D, et al. Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS platform. BMC Genomics. 2019;20(1):215. Epub 2019/03/15. doi: 10.1186/s12864-019-5569-5. PubMed PMID: 30866797; PubMed Central PMCID: PMCPMC6416933.
3. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics. 2018;19(1):30. Epub 2018/01/10. doi: 10.1186/s12864-017-4428-5. PubMed PMID: 29310587; PubMed Central PMCID: PMCPMC5759201.
4. Farouni R, Djambazian H, Ferri LE, Ragoussis J, Najafabadi HS. Model-based analysis of sample index hopping reveals its widespread artifacts in multiplexed single-cell RNA-sequencing. Nat Commun. 2020;11(1):2704. Epub 2020/06/03. doi: 10.1038/s41467-020-16522-z. PubMed PMID: 32483174; PubMed Central PMCID: PMCPMC7264361.
5. Griffiths JA, Richard AC, Bach K, Lun ATL, Marioni JC. Detection and removal of barcode swapping in single-cell RNA-seq data. Nat Commun. 2018;9(1):2667. Epub 2018/07/12. doi: 10.1038/s41467-018-05083-x. PubMed PMID: 29991676; PubMed Central PMCID: PMCPMC6039488.
6. Larsson AJM, Stanley G, Sinha R, Weissman IL, Sandberg R. Computational correction of index switching in multiplexed sequencing libraries. Nat Methods. 2018;15(5):305-7. Epub 2018/04/28. doi: 10.1038/nmeth.4666. PubMed PMID: 29702636.
7. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, et al. Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. BioRxiv. 2017.
8. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics. 2018;19(1):332. Epub 2018/05/10. doi: 10.1186/s12864-018-4703-0. PubMed PMID: 29739332; PubMed Central PMCID: PMCPMC5941783.

606 9. Owens GL, Todesco M, Drummond EBM, Yeaman S, Rieseberg LH. A novel post hoc
607 method for detecting index switching finds no evidence for increased switching on the Illumina
608 HiSeq X. BioRxiv. 2017.
609 10. Illumina. Effects of index misassignment on multiplexing and downstream
610 analysis. 2017. Available from: https://www.illumina.com/content/dam/illumina-
611 marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf.
612 11. Ezpeleta J, Garcia Labari I, Villanova GV, Bulacio P, Lavista-Llanos S, Posner V, et al.
613 Robust and scalable barcoding for massively parallel long-read sequencing. Sci Rep.
614 2022;12(1):7619. Epub 20220510. doi: 10.1038/s41598-022-11656-0. PubMed PMID:
615 35538127; PubMed Central PMCID: PMCPMC9090787.
616 12. Xiao Y, Thoresen DT, Miao L, Williams JS, Wang C, Atit RP, et al. A Cascade of Wnt, Eda,
617 and Shh Signaling Is Essential for Touch Dome Merkel Cell Development. PLoS Genet.
618 2016;12(7):e1006150. Epub 2016/07/16. doi: 10.1371/journal.pgen.1006150. PubMed PMID:
619 27414798; PubMed Central PMCID: PMCPMC4944988.
620 13. Vassen L, Okayama T, Moroy T. Gfi1b:green fluorescent protein knock-in mice reveal a
621 dynamic expression pattern of Gfi1b during hematopoiesis that is largely complementary to
622 Gfi1. Blood. 2007;109(6):2356-64. Epub 2006/11/11. doi: 10.1182/blood-2006-06-030031.
623 PubMed PMID: 17095621.
624 14. Yucel R, Kosan C, Heyd F, Moroy T. Gfi1:green fluorescent protein knock-in mutant
625 reveals differential expression and autoregulation of the growth factor independence 1 (Gfi1)
626 gene during lymphocyte development. J Biol Chem. 2004;279(39):40906-17. Epub 2004/07/15.
627 doi: 10.1074/jbc.M400808200. PubMed PMID: 15252036.
628 15. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master
629 transcription factors and mediator establish super-enhancers at key cell identity genes. Cell.
630 2013;153(2):307-19. Epub 2013/04/16. doi: 10.1016/j.cell.2013.03.035. PubMed PMID:
631 23582322; PubMed Central PMCID: PMCPMC3653129.
632 16. Julia M, Telenti A, Rausell A. Sincell: an R/Bioconductor package for statistical
633 assessment of cell-state hierarchies from single-cell RNA-seq. Bioinformatics. 2015;31(20):3380-
634 2. Epub 2015/06/24. doi: 10.1093/bioinformatics/btv368. PubMed PMID: 26099264; PubMed
635 Central PMCID: PMCPMC4595899.
636 17. Andoniadou CL, Matsushima D, Mousavy Gharavy SN, Signore M, Mackintosh AI,
637 Schaeffer M, et al. Sox2(+) stem/progenitor cells in the adult mouse pituitary support organ
638 homeostasis and have tumor-inducing potential. Cell Stem Cell. 2013;13(4):433-45. Epub
639 2013/10/08. doi: 10.1016/j.stem.2013.07.004. PubMed PMID: 24094324.
640 18. Picelli S, Faridani OR, Bjorklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-
641 seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171-81. Epub 2014/01/05. doi:
642 10.1038/nprot.2014.006. PubMed PMID: 24385147.
643 19. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic
644 data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411-20.
645 Epub 20180402. doi: 10.1038/nbt.4096. PubMed PMID: 29608179; PubMed Central PMCID:
646 PMCPMC6700744.
647 20. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and
648 regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat

649 Biotechnol. 2014;32(4):381-6. Epub 20140323. doi: 10.1038/nbt.2859. PubMed PMID:
650 24658644; PubMed Central PMCID: PMCPMC4122333.
651 21. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding
652 resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979-82. Epub 20170821.
653 doi: 10.1038/nmeth.4402. PubMed PMID: 28825705; PubMed Central PMCID:
654 PMCPMC5764547.
655

656 ## SUPPORTING INFORMATION

657 Figures S1-S9:

658 **Figure S1.** Skin schematic showing cells labeled by GFP in Sox2GFP/+ and Gfi1GFP/+
659 reporter mice.

660 **Figure S2.** Index hopping can be detected when occurring among different cell types.

661 **Figure S3.** Top DC markers genes are detected in MCs with single index hopping
662 potential from DCs and the detection ratio is correlated with number of different cell types
663 with the potential for index hopping.

664 **Figure S4.** Rank ordered scatter plot showing distribution of gene detection ratios in MCs
665 across all genes from mm10 genome annotated by Ensembl.

666 **Figure S5.** Index hopping correction reduces the detection of non-self marker genes in
667 different cell types.

668 **Figure S6.** Index hopping correction reduces the detection of non-self marker genes.

669 **Figure S7.** Assigned developmental stages of single DCs are altered by index hoping
670 correction.

671 **Figure S8.** Assigned developmental stages of single HCs are altered by index hoping
672 correction.

673 **Figure S9.** Detection ratios of top MC marker genes in HCs show that index hopping is
674 reduced on Hiseq2500 compared to Hiseq3000 sequencing platforms.