Statistical modeling, estimation, and remediation of sample index hopping in multiplexed droplet-based single-cell RNA-seq data

Rick Farouni [1,2,*] and Hamed S. Najafabadi [1,2,*]

[1] Department of Human Genetics, McGill University, Montreal, QC, Canada.

[2] McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada.

[*] To whom correspondence should be addressed.

rick.farouni@mcgill.ca; hamed.najafabadi@mcgill.ca

Abstract

Sample index hopping can substantially confound the analysis of multiplexed sequencing data due to the resulting erroneous assignment of some, or even all, of the sequencing reads generated by a cDNA fragment in a given sample to other samples. In those target samples, the data cross-contamination artifact takes the form of phantom molecules, molecules that exist only in the data by virtue of read misassignment. Phantom molecules should be a cause of great concern in droplet-based single cell RNA-seq experiments since they can introduce both phantom cells and artifactual deferentially-expressed genes in downstream analyses. More importantly, even when the index hopping rate is very small, the fraction of phantom molecules in the entire dataset can be high due to the distributional properties of sequencing reads across samples. To our knowledge, current computational methods are unable to accurately estimate the underlying rate of index hopping nor adequately correct for the resultant misassignment in droplet-based single cell RNA-seq data. Here, we introduce a probabilistic model that formalizes the phenomenon of index hopping and allows the accurate estimation of its rate. Application of the proposed model to several multiplexed datasets suggests that the sample index hopping probability for a given read is approximately 0.008, an arguable low number, even though, counter-intuitively, it can give rise to a large fraction of phantom molecules - as high as 85% - in any given sample. We also present a model-based approach for inferring the true sample of origin of the reads that are affected by index hopping, thus allowing the purging of the majority of phantom molecules in the data. Using simulated and empirical data, we show that we can reassign reads to their true sample of origin and remove predicted phantom molecules through a principled probabilistic procedure that optimally minimizes the false false positive rate. Thus, even though sample index hopping often substantially compromises single-cell RNA-seq data, it is possible to accurately quantify, detect, and reassign the affected reads and remove the phantom molecules generated by index hopping. Code and reproducible analysis notebooks are available at https://github.com/csglab/phantom_purge.

1 Introduction and Background

Contributions. The contributions of this paper are fourfold. (1) A generative probabilistic model that formalizes in a mathematically rigorous manner the phenomenon of index hopping; (2) A statistical approach for inferring the sample index hopping rate in droplet-based scRNA-seq data at the level of individual reads; (3) A non-heuristic, model-based approach for inferring the true sample of origin of hopped sequencing reads (i.e. real molecules); (4) A data de-contamination procedure for purging phantom molecules which optimally minimizes the false positive rate of the predictions.

Paper outline. In this section (Section 1), we describe the phenomenon of sample index hopping and provide a brief overview of existing published research that aims to quantify its rate and to correct for its data corrupting effects. We also highlight the limitations of existing approaches and argue that the issue of index hopping is a significant problem in the field, one that no adequate solution yet exists. In Section 2, we present our proposed solution in depth. The resulting methods section forms the core of this paper and is subdivided into four subsections corresponding to the four contributions we stated at the onset of the manuscript. In Section 3, we apply and evaluate the approach on empirical data. Finally, in Section 4, we discuss both the significance and limitations of our proposed approach and suggest a few ideas and directions for future research. Although much of what we propose in this paper can be ported, modified, and applied to other protocols, we limit the scope of the paper to droplet-based scRNA-seq data, with libraries prepared using the 10x Genomics Single Cell Protocol and that are subsequently multiplexed for sequencing on Illumina machines.

1.1 Overview of Sample Index Hopping on Illumina's Sequencers

Patterned flowcells. Ever since the emergence of next-generation sequencing technologies in 2006, the speed and throughput at which we can perform whole-genome DNA and RNA sequencing have been steadily increasing. For example, in recent years Illumina has introduced a family of sequencing platforms (i.e. the HiSeq 4000 and NovaSeq 6000) characterized by substantial increased capacity, faster run times, and lower sequencing cost. The significant boost in sequencing throughput[1] first came about in 2015 when Illumina introduced the patterned flow cell technology on the HiSeq 4000 sequencers. In particular, there were two key innovations that led to the significantly improved performance. (1) A flow cell surface design that optimizes the prearranged spacing of nanowells, thus allowing the accurate imaging of billions of clusters of amplified cDNA fragments. (2) *The Exclusion Amplification* (ExAmp) chemistry that instantaneously amplifies a single cDNA fragment, effectively excluding and preventing other cDNA fragments from forming a cluster within the same nanowell. In 2017, Illumina launched the NovaSeq 6000 instrument with an updated

[1] As measured by the total number of nucleotide sequence reads generated in a single run of an experiment

flow cell design that further decreased the spacing between the nanowells, thus significantly increasing the density of generated clusters and, consequently, the amount of data that can be generated. Currently, a total of 2.5B single reads can be generated on each of the four lanes that comprise a single NovaSeq S4 flow cell, compared to 350M reads on each of eights lanes of HiSeq 4000.

Sample multiplex sequencing. However, given that for most sequencing studies, the number of reads required to achieve an adequate transcriptome coverage tends to be in the tens of millions, it then becomes necessary to resort to *sample multiplex sequencing* strategies to fully utilize the massive capacity and cost efficiency that the patterned flow cell with ExAmp chemistry technology can bring about. The process of sample multiplexing can be briefly summarized with the following few steps. First, a *sample barcode index* is added to each cDNA fragment during library preparation (to one end in *single-indexing* or to both ends in *dual-indexing*). Then, cDNA fragments from multiple samples are subsequently pooled together in the same lane to be cluster-amplified and sequenced. Finally, the generated sequenced reads are demultiplexed into their respective source samples using the sample barcode indices. Note that with current technology, a maximum of a 384 samples can be multiplexed using a 96-plex on each of the four lanes of a NovaSeq S4 flow cell.

Sample read misassignment. Unfortunately, sample multiplexing can cause a significant percentage of the demultiplexed sequenced reads to be misassigned to an incorrect sample barcode. Although sample read misassignments can arise due to several factors, one specific mechanism termed *sample index hopping* is the primary cause of read misassignments in patterned flow cells. Index hopping is believed to result from the presence of free-floating indexing primers that attach to the pooled cDNA fragments just before the exclusion amplification step that generates clusters on the flow cell (Illumina, 2017).

1.2 Impact and Significance of Index Hopping.

As we detail in Subsection 1.3, several studies have shown that the existence of substantial number of reads assigned to incorrect samples in an experiment can potentially lead to signal artifacts and spurious results. Indeed, given the large amount of public data that have been and are being generated using multiplexed sequencing, index hopping should be a great cause of concern. Here we list the three main unwanted consequences in the context of scRNA-seq sequencing experiments that sample index hopping can bring about through the introduction of phantom molecules.

1. *Mixing of transcriptomes*: when a given cell-barcode is observed in both the donor and target samples, index hopping would result in phantom molecules being introduced in the cell with the corresponding cell-barcode in the target sample. When this happens,

the rates of false positives and even false negatives can increase in downstream statistical analyses such as differential gene expression.

2. *Phantom cells*: when a given cell-barcode is observed only in the donor sample, but not the target, phantom molecules would also lead to phantom cells emerging in the target sample.

3. *Cell-barcode miss-classification*: An abundance of phantom molecules associated with a given cell-barcode would lead cell-barcoding algorithms (used to computationally determine which cell-barcodes originate from proper cells) to incorrectly classify an empty droplet as a cell and conversely to classify a proper cell as an empty droplet.

1.3 Discovery and Quantification of Index Hopping

The problem of index hopping was first identified as early as December 2016 in a blog post reporting read sample misassignment on HiSeq 4000 and HiSeqX platforms (Hadfield, 2016). Soon afterwards, Illumina (2017) released a white paper acknowledging that index switching does indeed occur and tends to be higher in machines that use a patterned flow cell, but maintaining that the phenomenon affects only <2 % of reads. However, around the same time, Sinha *et al.* (2017) reported that 5-10% of sequencing reads were incorrectly assigned a sample index in a multiplexed pool of plate-based scRNA-seq samples. More recently, using two plate-based scRNA-seq datasets, Griffiths *et al.* (2018) provided a lower estimate of the index swapping on the HiSeq 4000 at approximately 2.5%. Another study (Vodák *et al.*, 2018) conducted in the context of exome sequencing also reported a contamination (index hopping rate) ranging from 2% to 8%. With the aim of reconciling the conflicting hopping rate estimates that had been reported, Costello *et al.* (2018) used a non-redundant dual indexing adapters developed in-house and performed an exhaustive study across multiple libraries (whole genomes, exome, and stranded RNA) and sequencer models (HiSeq 4000, NovaSeq 6000 and HiSeqX) to determine the rate of index hopping. They observed a rate of 0.2% to 6% in all sequencing runs. More importantly, they showed that even in bulk RNA-seq libraries where the indexing hopping rate was as low as 0.32%, spurious results in downstream analysis can be generated. This finding mirrors the conclusions reached by the four papers referenced in the preceding paragraph including Illumina's white paper.

1.4 Mitigating the Effects of Sample Index Hopping.

Existing experimental solutions. Given that index hopping would occur in all sequencing platforms that uses patterned flow cells, one can resort to any of the following three experimental strategies to mitigate sample index hopping from affecting the integrity of experimental data.

1. *One sample per lane.* By avoiding multiplexing altogether and running one sample per lane, one can confine the sample indices in a given lane to be for one given sample only,

thus minimizing the risk of index hopping from occurring in the first place. However, sample read misassignments can still occur due to other causes.

2. *Unique-at-both-ends dual indexing library.* By utilizing two unique barcodes for each sample, sample misassignment would occur only when both sample indices hop. Given the low probability that such an event would occur, index hopping can be reduced and computationally mitigated by discarding reads with unexpected combinations post-hoc.

3. *Post-library prep treatment.* By using Illumina's Free Adapter Blocking Reagent, the $3'$ ends of the free adapters become blocked preventing their extension and thus reducing the rate of index hopping Illumina (2017).

Limitations of Experimental Strategies. Although running one sample per lane can be feasible on HiSeq platforms for some single cell RNA-seq libraries, it would be financially prohibitive on newer higher capacity NovaSeq platforms. As for unique-at-both-ends indexing, their use in plate-based single cell experiments where many samples (i.e. cells) need to be multiplexed together is problematic since dual indexing restricts the number of possible samples that can multiplexed together (see (Griffiths *et al.*, 2018) for details). More importantly, unique-at-both-ends dual-indexing is incompatible with single index droplet-based protocols such as the widely used 10x Genomics single cell protocol. Lastly, the blocking reagent Illumina provides is only compatible with a few bulk DNA and RNA protocols, but not for single cell protocols. Whether it can be incorporated in single cell droplet-based protocol is yet to be determined.

Computational Strategies. Given the limitations of existing experimental strategies to eliminate index hopping, the urgency to develop computational strategies instead has only intensified since the phenomenon was first discovered. Indeed, the first attempt to computationally correct for index hopping in multiplexed sequencing libraries was only recently published (Larsson *et al.*, 2018). However, the linear regression method the authors propose is applicable only to plate-based scRNA-seq samples, where a unique combination of a pair of sample barcodes in each well determines the identity of cells and can give rise to "crosshair" pattern in the data when index hopping occurs. It is this pattern that allows the estimation of fraction of hopped reads and upon which their proposed approach relies. Another limitation of their approach is that it uses only a subset of the data corresponding to genes whose specific cell expression is above a given threshold. However, Griffiths *et al.* (2018) developed a more accurate method for plate-based scRNA-seq that makes fewer assumptions and that uses data from all the genes. In the same paper, the authors also proposed a heuristic computational strategy for droplet-based scRNA-seq 10x Genomics experiments that excludes hopped reads without removing the corresponding cell libraries, but did not attempt to propose a modeling framework for estimating sample index hopping

in droplet-based data since unlike plate-based assays, droplet-based assays do not use a unique pair of sample barcodes but rather use one mate of the paired read for quantification and a second mate to carry the UMI and cell barcode tags, thus rendering the "crosshair" pattern approach inapplicable.

Limitations of existing computational strategies. The computational strategy proposed in Griffiths *et al.* (2018) is the only attempt we are aware of that is aimed at estimating the rate and mitigating the effects of index hopping in droplet-based scRNA-seq data. Nonetheless, the strategy the authors propose, similar to other computational approaches we mentioned in the previous paragraph, does not in fact estimate the index hopping rate of individual reads, but rather computes a proxy measure, *the swapped fraction* in the case of Griffiths *et al.* (2018), which they define as the fraction of molecules (i.e. those with an identical UMI, cell barcode, gene label) that are observed in more than one sample. As we show in this paper, the probability at which an individual sequencing read swaps the sample index is not the same as the probability that one out of all the PCR-duplicated reads from the same molecule swaps the sample index. As we show, even when the sample index hopping probability at the level of individual read is very low, the fraction of molecules that we observe to have swapped sample indices can vary greatly, even as far as comprising the vast majority of molecules in a given sample. As for correcting for the effects of index hopping, the proposed molecular exclusion heuristic the authors propose, which we discuss further in Section 4, suffers from high positive and negative rates, resulting in unnecessarily discarding molecules that otherwise should have been retained and conversely, in retaining molecules that should have been discarded. Furthermore, it ignores the extremely variable molecular counts (i.e. *library complexity*) across samples, which, as our statistical formulation of the sample index hopping shows, would lead to nonsensical results when samples with uneven complexities are multiplexed.

2 Methods

2.1 Generative Model Specification

Consider a droplet-based single-cell sequencing experiment where a total of S (ranging from 2 to 96) libraries are pooled together and multiplexed on the same lane of a patterned flow cell. Here we simply define a *sequencing library* as a collection of index-barcoded and PCR-amplified cDNA fragments which are purified from the mRNA of a particular sample tissue. In a single sequencing run, millions of short sequencing reads are generated (from 350M reads on a single lane of HiSeq 4000 to 2.5B reads on a single lane of a Novaseq 6000 S4 flowcell), each of which is annotated with barcodes representing the sample, cell, and molecule from which the read originated. If the reads are aligned to the genome, a read would also be annotated with the genomic location where it mapped to. For example, each

mapped read in a 10x Genomics Single Cell 3' v2 Gene Expression Library is thus annotated by four labels:

1. A sample barcode. In the case of a 10x Genomics single-indexed library, there are four indices for each sample. In what follows, we collapse all four indices into a single sample index for computational, mathematical, and practical considerations. However, as we show in Section 2.2.2, the model we propose allows us to recover the sample index hopping rate at the level of the barcodes.

2. A 16bp cell-barcode index randomly selected out of a set containing 737,280 possible combinations.

3. A 10bp Unique Molecular Identifier (UMI) index for which there are a total of $4^{10} = 1,048,576$ combinations. A *UMI collision* occurs when two or more UMIs possess the same sequence.

4. A mapped gene ID as provided by a transcriptome. For example, Ensemble 95 (Zerbino *et al.*, 2018) has annotations for 19,768 and 21,823 non-conjoined coding and noncoding genes, respectively.

That is, after sample demultiplexing and transcriptome alignment, each read gets associated with a *cell-umi-gene-sample* label. We note here that by discarding unmapped reads and reads that mapped to genomic regions other than exon bodies, we are further making the implicit assumptions that the large portion of data that we end up retaining contains enough information to determine the index hopping rate and that the data we discard is not characterized by a different mechanism underlying the sample barcode index hopping phenomenon.

In scRNA-seq data, a cell is identified by a unique cell barcode. However, only a fraction (1k-10k) out of the +100K observed cell barcodes correspond to actual cells. In each of these cells, we typically observe anywhere from a minimum of a 1000 (a threshold usually specified manually) up to approximately 100,000 unique molecules, or more depending on the average read coverage per cell. Furthermore, given that the UMI index is 10bp long, the number of possible UMIs is one order of magnitude larger than the number of UMIs typically observed in any one cell. Accordingly, a UMI collision in a single cell, when it does occur, would result in two reads from different molecules to be considered as originating from the same cDNA fragment (i.e. PCR duplicates), thus reducing the number of real molecules that can be observed in any one cell. To further reduce the chance of collisions, gene ID labels can be considered in order to limit the space of collisions to those UMIs that are mapped to the same region of the genome. Although these genomic regions, each consisting of the union of exons belonging to a given gene, vary in length, the localization of possible collisions to a small region of the genome greatly reduces the probability of UMI collisions.

2.1.1 Modeling Sequencing Reads with a Mixture of Categorical Distributions

We start by observing that after cDNA molecules of a given transcript are amplified, they are subsequently fragmented into 300-400 bp fragments, each containing both a cell barcode and a UMI. During the sample index PCR amplification step, an Illumina Read 2 adapter is ligated to the fragments before they get amplified, resulting in a given number of sequencing reads, $n_l \in \mathbb{N}_+$, to have the same *cell-umi-gene-sample* label $l = 1 \dots L$ where $L \in \mathbb{N}_+$ denotes the total number of unique combinations of the labels. If none of the sequencing reads, $n = 1 \dots n_l$, belonging to label l are misassigned, then each read's observed sample index would correspond to the true sample index. When there is index hopping, we can model the read misassignment process as a mixture of S categorical distributions where each observed read sample index belongs to one sample. See Table 1 for an example toy dataset in which each row corresponds to a single sequencing read with label l.

In particular, let $s_{ln} \in \{1, \cdots, S\}$ denote the true sample index and $d_{ln} \in \{1, \cdots, S\}$ denote the observed sample index of the n'th read associated with label l. We can formulate the mixture model formally as a two-stage hierarchical sampling process, where we first draw the unobserved true sample index s_{ln} that gets assigned to read n with label l from a categorical distribution governed by a label-specific probability parameter vector $\boldsymbol{\varphi}_l$. Next we draw the observed sample index from another categorical distribution with a probability parameter vector $\boldsymbol{p}_{s_{ln}}$ whose value is conditional on the value of s_{ln}. In what follows, s_{ln} denotes the index of the element in the vector $\boldsymbol{s_{ln}}$ for which we observe a one (i.e $[\boldsymbol{s_{ln}} = 1]$).

$$
\begin{aligned}
\boldsymbol{s}_{ln} &\sim \text{Categorical}(\boldsymbol{\varphi}_l) \\
\boldsymbol{d}_{ln} | s_{ln} &\sim \text{Categorical}(\boldsymbol{p}_{s_{ln}}) \\
\text{where} \quad & \boldsymbol{\varphi}_l \in [0,1]^S, \|\boldsymbol{\varphi}_l\|_1 = 1 \\
& \boldsymbol{p}_{s_{ln}} \in [0,1]^S, \left\|\boldsymbol{p}_{s_{ln}}\right\|_1 = 1 \\
& l = 1 \dots L.; \; n = 1 \dots n_l. \\
& s_{ln} \in \{1, \cdots, S\}; S \in \mathbb{N}_+
\end{aligned}
\tag{1}
$$

The probability parameter vectors of all the S distributions can be stacked together in $\boldsymbol{P}$, the sample hopping probability matrix.

$$
\boldsymbol{P} = \begin{bmatrix} \boldsymbol{p}_1^\intercal \\ \boldsymbol{p}_2^\intercal \\ \vdots \\ \boldsymbol{p}_S \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1S} \\ p_{21} & p_{22} & \dots & p_{2S} \\ \vdots & \vdots & \ddots & \vdots \\ p_{S1} & p_{S2} & \dots & p_{SS} \end{bmatrix}
$$

When $\boldsymbol{P} = \mathbf{I}$, then the probability that a read keeps its sample index of origin is one, and consequently, reads do not hop over to other samples. That is, $d_{ln} = s_{ln}$ for all l and n.

Here p_{ij} denotes the probability that a read from sample i hops to sample j. The number of parameters in $\boldsymbol{P}$ equals $S \times (S-1)$.

Table 1: A toy dataset for three multiplexed samples (i.e. $S = 3$) in which each observation is a sequencing read with label l (i.e. a unique combination of c, u, and g, which denote the cell-barcode, UMI, and gene, respectively), a true source sample s, and an observed sample d_i, $i = 1, 2, 3$. In a typical experiment, a single lane can generate anywhere from 350 million to 2.5 billion sequencing reads.

label	l	s	d_1	d_2	d_3
$c_1u_2g_1$	1	1	0	1	0
$c_1u_1g_1$	2	2	1	0	0
$c_1u_3g_2$	3	2	1	0	0
$c_1u_3g_2$	3	2	1	0	0
$c_1u_3g_2$	3	2	0	1	0
$c_2u_1g_1$	4	1	0	1	0
$c_2u_1g_1$	4	1	0	0	1
$c_2u_1g_1$	4	1	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

2.1.2 Modeling Read Counts with a Mixture of Multinomial Distributions

In Model 1 there are a total of L parameter vectors $\boldsymbol{\varphi}_l$ each taking values on the probability simplex. Note that the n_l reads associated with label l can come from different source samples. This could happen for example when reads from two samples are assigned the same label simply by chance. However, such an outcome is extremely unlikely and by ruling it out we can greatly simplify the modeling framework if we constrain each of all the L parameter vectors $\boldsymbol{\varphi}_l$ to be the one vector (i.e. $\mathbf{1}$). That is, we assume that all the reads with a given *cell-umi-gene* label (i.e. PCR read duplicates of a unique cDNA molecule) can be indexed only with the same sample barcode. More formally, we formulate the assumption as follows.

Assumption 1 (Discrete Simplex Constraint Assumption)**.** Let $max(\boldsymbol{\varphi}_l) = 1$ for all l such that $\boldsymbol{\varphi}_l \in \Delta_1^S := \left\{ \left(\ \varphi_0, \ldots, \varphi_S \ \right) \mid \varphi_i \in \{0, 1\}, \sum_i \varphi_i = 1 \right\}$. That is, we restrict $\boldsymbol{\varphi}_l$ to belong to the discrete S-simplex, a set with cardinality S whose elements are vectors each consisting of a single 1 and zeros elsewhere.

Distribution of sample read counts. As a result of Assumption 1, the sample index s_{ln} becomes a constant random variable drawn from one of S degenerate categorical distributions such that all the n_l reads with label l are assigned the same sample index. That is, for each

$\boldsymbol{\varphi}_l$, all the $\boldsymbol{p}_{s_{ln}}$ vectors become the same for all the n_l reads (i.e. $\boldsymbol{p}_{s_l}$). Consequently, we can sum over the n_l categorical random variables to obtain a mixture of multinomials model for data $\{\boldsymbol{y}_l\}$, where $\boldsymbol{y}_l = (y_{l1}, \cdots, y_{ls}, \cdots, y_{lS})$ denotes the vector of read counts across S samples corresponding to an observed label l and an unobserved source (i.e. donor) sample $s_l \in \mathcal{S} := \{1, \cdots, S\}$.

Note here that the total number of mapped reads across all samples is given by $N = \|\boldsymbol{n} = (n_1, \cdots, n_L)\|_1$ where $\boldsymbol{n}$ is the L-dimensional vector of read count sums. In what follows, we partition $\boldsymbol{n}$ by the unique values (i.e. read count sums) that n_l can take, namely, $r \in \mathcal{R} := \{1, \ldots, R \mid r \in \mathbb{N}_+\}$, where $\|\mathcal{R}\| = I$ and R is the maximum value in $\mathcal{R}$. For each $r \in \mathcal{R}$ let $m_r \in \mathcal{M} := \{m_1, \ldots, m_R \mid m_r \in \mathbb{N}_+\}$ denote the corresponding number of times r is observed in $\boldsymbol{n}$ such that $\sum_{r=1}^{R} m_r = L$. We denote the empirical distribution of r also by $\mathcal{M}(.)$. Since the variable r counts the number of PCR duplicated reads - whether hopped or not - that a given unique molecule generates, we will refer to r as the *PCR duplication level*. Accordingly, we can write the mixture of multinomials model conditional on the *PCR duplication level* r as follows.

$$
\begin{aligned}
\boldsymbol{s}_l &\sim \text{Categorical}(\boldsymbol{\pi}_r) \\
\boldsymbol{y}_l \mid \boldsymbol{s}_l &\sim \text{Multinomial}(r, \boldsymbol{p}_{s_l}) \\
& l = 1, \ldots, m_r. \\
\text{where} \quad & \boldsymbol{y}_l \in \mathbb{N}_0^S,\ S \in \mathbb{N}_+, \\
& \|\boldsymbol{y}_l\|_1 = r, s_l \in s, \\
& \boldsymbol{\pi}_r \in [0,1]^S, \|\boldsymbol{\pi}_r\|_1 = 1, \\
& \boldsymbol{p}_s \in [0,1]^S, \|\boldsymbol{p}_s\|_1 = 1.
\end{aligned} \tag{2}
$$

where l now indexes the m_r observations with *PCR duplication level* r; the vector $\boldsymbol{p}_{s_l}$ denotes the s_l row of the $S \times S$ matrix of the sample read index-hopping probabilities; and the probability vector $\boldsymbol{\pi}_r$ represents the proportion of molecules across the S samples at *PCR duplication level* r.

The mixture model can be viewed more intuitively as a generative process in which we first sample a molecule $\boldsymbol{s}_l$ from a library sample s_l according to the categorical model then we amplify the molecule by generating r PCR read duplicates according to the multinomial model. The number of molecules originating from a given sample with a given PCR read duplicates r is determined by the parameter vector π_r whereas the number of PCR duplicated reads that end up hopping to other samples is determined by the parameter vector $\boldsymbol{p}_{s_l}$).

The molecular proportions complexity profile. Let $\boldsymbol{v}_r = \sum_l^{m_r} \boldsymbol{s}_l$ denote the molecule count across the S samples at level r and $\boldsymbol{v} = \sum_r^R \sum_l^{m_r} \boldsymbol{v}_r = \sum_r^R \sum_l^{m_r} \boldsymbol{s}_l$ the molecule count across the S samples for the entire data. Since $\mathrm{E}(\boldsymbol{v}_r) = m_r \times \boldsymbol{\pi}_r$, the expected total molecule count is therefore $\mathrm{E}(\boldsymbol{v}) = \sum_r^R m_r \times \boldsymbol{\pi}_r$. If $\boldsymbol{\pi}_r = \boldsymbol{\pi}$ for all $r = 1, \ldots, R$, then $\mathrm{E}(\boldsymbol{v}) = L \times \boldsymbol{\pi}$ would correspond to the *library complexities* of the S samples, where *Library*

complexity is defined as the expected number of unique molecules sampled with a finite number of sequencing reads generated in a given high-throughput sequencing run. It is important to note here that in a multiplexed experiment, several samples that usually vary in their molecular complexity are sequenced together. These samples would differ in the total number of unique transcripts each one has due to a host of factors, ranging from the presence of varying amounts of RNA that characterize different cell types (e.g. neuronal cells have low RNA content) to accidental errors in library preparation that could cause many cells to break up and lose their endogenous mRNA. That is, even if total number of available sequencing reads were budgeted evenly over the multiplexed samples, the number of unique molecules detected in the sequencing run could vary widely across the samples. We propose here that a more informative picture could be gained into the root cause of a sample's low library complexity, if we consider the molecule counts conditional on the *PCR duplication level* r or rather the set of molecular proportions $\Pi := \{\boldsymbol{\pi}_r\}_{r=1}^{R}$, which we term the *molecular proportions complexity profile*, in order to assess and identify potential problems across all the samples in a sequencing run. In Section 2.3.1 we derive Equation 6 for estimating Π from the observed proportions of total reads and in Section 3.3, we use *molecular proportions complexity profile* plots to identify problematic samples in empirical datasets.

The cardinality of the joint sample space. The conditional model's sample space is a joint space constructed by the Cartesian product of the sample space of the read count variable $\boldsymbol{y}$ and the sample space of the sample-of-origin latent variable $\boldsymbol{s}$.

$$\mathcal{S} \times \mathcal{Y}_r = \{\, (\boldsymbol{s}, \boldsymbol{y}) \mid \boldsymbol{s} \in \mathcal{S} \;\text{ and }\; \boldsymbol{y} \in \mathcal{Y}_r \,\}$$

The total number of possible read count outcomes conditional on *PCR duplication level* r is given by the cardinality of sample space of $\mathcal{Y}_r$. Since these outcomes correspond to the coefficients of r'th component of the Pascal's *S-Simplex* $\wedge_r^S$, the number of the outcomes (i.e. elementary events) are therefore given by the $(r+1)$'th *triangular number* $t_{(r+1)}$. Now, since the cardinality of $\mathcal{S}$ is determined by the number of samples S, the cardinality of $(\mathcal{S} \times \mathcal{Y}_r)$, which we denote by e_r, can be obtained as such.

$$\begin{aligned} e_r = S \times t_{(r+1)} &= S \times \binom{r + (S-1)}{(S-1)} \\ &= \frac{S \times (r + S - 1)!}{(S-1)!r!} \end{aligned}$$

The cardinality e_r increases very rapidly as S or r increases. Note that the number of possible outcomes across all possible values of r (i.e. the cardinality of the joint sample space of $(\mathcal{S} \times \mathcal{Y}_r \times \mathcal{R})$ and which we denote by e) is much larger since its cardinality is the sum of all outcomes across a range of r values (i.e. $e = \sum_r e_r$). For example, for $S = 8$ and $r = 1, \ldots, 100$, e is approximately 3 trillion!

Table 2: The set of all 9 possible outcomes generated by Model 2 conditional on $r = 1$ and $S = 3$. The probabilities of the outcomes are determined by the parameter p, the complement of the hopping rate, and the parameter vector π_r, representing the proportion of molecules across samples conditional on the read count sum r. Two possible probability assignments are shown: $\text{prob}^{(1)}$ corresponds to a model with $\pi_r = (0.6, 0.3, 0.1)$ and $p = 0.9$ (see Assumption 2), whereas $\text{prob}^{(2)}$ corresponds to a model with $\pi_r = (0.33, 0.33, 0.33)$ and $p = 0.9$. The variable f represents the number of phantom molecules. The proportion of phantoms, which in this case is 0.1, can be obtained by taking the expectation of f over the induced probabilities and dividing by total number of molecules.

outcome	s	y_1	y_2	y_3	$\text{prob}^{(1)}$	$\text{prob}^{(2)}$	f
{ 1, (1, 0, 0) }	1	1	0	0	0.540	0.300	0
{ 2, (1, 0, 0) }	2	1	0	0	0.015	0.017	1
{ 3, (1, 0, 0) }	3	1	0	0	0.005	0.017	1
{ 1, (0, 1, 0) }	1	0	1	0	0.030	0.017	1
{ 2, (0, 1, 0) }	2	0	1	0	0.270	0.300	0
{ 3, (0, 1, 0) }	3	0	1	0	0.005	0.017	1
{ 1, (0, 0, 1) }	1	0	0	1	0.030	0.017	1
{ 2, (0, 0, 1) }	2	0	0	1	0.015	0.017	1
{ 3, (0, 0, 1) }	3	0	0	1	0.090	0.300	0

When we consider the multinomial mixture model across a range of r values, it can be helpful to view the model in terms of how it divides and allocates the unit probability measure to a set of e outcomes, as determined by the parameter values that govern the distribution. To illustrate, we list in Table 2 the $e_1 = 9$ possible outcomes when we have $S = 3$ samples for data confined to *PCR duplication level* $r = 1$. In the table, we also list the probabilities assigned to the possible outcomes under two possible choices of model parameters. Notice that since s is latent, we only see collapsed marginal outcomes in the data we observe. That is, we see 3 outcomes rather than 9. We would like to highlight here an important point that is very relevant to subsequent sections; namely, if we perform a simple tally of read collisions in this data, we would fail to reveal the true fraction of phantom molecules (see Definition 2 below) since we cannot observe the true sample of origin $\boldsymbol{s}$ along with the observable outcome $\boldsymbol{y}$. Section 2.1.4 and accompanying Table 5 expand further upon this point.

2.1.3 Corollaries and Definitions

In addition to inducing the specification of Model 2, Assumption 1 entails the following two corollaries.

Corollary 1 (Label Collision Across Samples)**.** A given *cell-umi-gene* label combination has a zero probability of co-occurring with more than one sample barcode index. That is, reads annotated with the same label combination can only belong to one sample. This implication formalizes a key insight from Griffiths *et al.* (2018) that informed their molecular exclusion strategy to reduce the effects of index hopping in 10x Genomics experiments.

Corollary 2 (Number of Molecules)**.** The number of total molecules across all S samples equals the number of unique *cell-umi-gene* label combinations L. In other words, the number of molecules equals the number of rows in a merged data table of sample read counts, which has been fully joined by the combination of cell, umi, and gene key (Table 3).

Corollary 2 implies that since a distinct molecule is defined by a read (or multiple reads) with a unique *cell-umi-gene* label, any label collision across samples would result in a total number of *observed* molecules greater than L. Therefore, to avoid any potential naming confusions, we make the following three clarifying definitions.

Definition 1 (Chimeric Observation)**.** A chimera refers to a hybrid creature composed of the parts of more than one animal. We use the analogy to refer to those read count *observations* for which $(y_{lj} \neq r)$ for all $j = 1, \ldots, S$, or in other words, to those *cell-umi-gene* labels where we observe reads from more than one sample. We can further classify chimeras by the number of collisions. A k-chimera has reads in k sample categories (e.g. $y = (4, 1, 0, 2, 0)$ is a 3-chimera when $S = 5$). In contrast, we refer to an observation with no collisions (i.e. a 1-chimera) as a non-chimera.

If there is no sample index misassignment whatsoever, then we expect that all the L observations to be non-chimeric. When the sample index hopping rate is large, we expect to see a correspondingly large proportion of chimeras, observations where there are collisions of labels across the samples.

Definition 2 (Phantom Molecule)**.** For a given observation with label *cell-umi-gene*, we term molecules observed in samples other than the true source sample as *phantom molecules*. We call them *phantom* since their existence is not real, due only to an artifact brought about by index hopping.

It is important to note that a sample index can hop even when we do not observe a label collision. Thus one or even all reads annotated with a given label may be actually hopping reads originating from the true source sample. In such a case, we are unable to determine the true source sample of these molecules. Nonetheless, we can estimate their expected observed proportions.

Definition 3 (Fugue Observation)**.** A disassociative fugue is a personality disorder characterized by unplanned travel, wandering, and even the establishment of a new identity. Similarly, for a given label, when all the reads hop over from an unobserved sample to establish a new identify in other samples, we term all such observations as *fugues* since we cannot know where they came from since we observe zero reads associated with the true sample of origin.

Given that we have no reason to believe that reads in a particular sample are characterized by a different chemical properties, we can simplify the problem by making the following assumption of uniformity.

Assumption 2 (One parameter to rule them all)**.** The probability of a read keeping its sample index is the same across samples (i.e. p) and the probability of a read switching the sample index is the same regardless of either its source or target sample (i.e. $p_h = \frac{(1-p)}{S-1}$).

Assumption 2 reduces the number of parameters in P from a total of $Q = (S-1) \times S$ to just a single parameter.

$$\boldsymbol{P} = \begin{bmatrix} p & p_h & \cdots & p_h \\ \vdots & \vdots & \ddots & \vdots \\ p_h & p_h & \cdots & p \end{bmatrix}$$

More importantly, Assumption 2 induces a symmetry on the probabilities we assign to the possible counts that lie within the same $m-$face of an $(S-1)-$simplex denoted by $\triangle_r^{S-1}$, which in turn constitutes the r'th component of the Pascal's Simplex denoted by $\wedge^S$.

Definition 4 (Sample Index Hopping Rate)**.** Whereas the term $p_h = \frac{(1-p)}{S-1}$ represents the probability that a sample index hops into a one particular target sample, the complement of p, namely $1-p$, refers to the quantity of interest, the Sample Index Hopping Rate, which we abbreviate as **SIHR** .

2.1.4 Toy Data Example

To illustrate the concepts and definitions in the preceding sections, consider the following toy example where we have three samples $S = 3$. The data would thus consist of three-dimensional discrete vector observations $\boldsymbol{y}_l \in \mathbb{N}_0^3$ sampled from a multinomial distribution with total sum of counts r, where $\|\boldsymbol{y}_l\|_1 = r$, and a probability parameter

$$\boldsymbol{p} \in \Delta^2 = \left\{ (p_1, \ldots, p_3) \in \mathbb{R}^3 \mid \sum_{s=1}^{3} p_s = 1 \text{ and } p_s \geq 0 \text{ for all } s \right\}$$

For concreteness, assume that the source sample s is known for each observation and consider what the data would look like when the sample index hopping rate (i.e. probability)

Table 3: Toy dataset of read counts and corresponding deduplicated counts (i.e. molecules). The data table shows 4 out of the 10 possible outcomes at read count sum level $r = 3$ for multiplexed data with $S = 3$ samples, see Figure 1. All 4 observations are associated with a unique *cell-umi-gene* label l each and an unobservable true sample of origin $s = 1$. For each read count $\boldsymbol{y}$, a vector of deduplicated read counts or molecules $\boldsymbol{x}$ is given. The chimeric number k denotes the number of molecules observed in each outcome and f denotes the number of phantom molecules. Although the chimeric vs. non-chimeric state of an observation can be directly seen from the data, the fugue vs. non-fugue state of an observation cannot be directly inferred, since it depends on knowing the true origin of the reads (i.e. the latent variable s). When there is no index hopping, the only outcomes possible are those classified as non-chimeric non-fugues.

					$\boldsymbol{y}$ (reads)			$\boldsymbol{x}$ (molecules)			
l	r	k	f	s	y_1	y_2	y_3	x_1	x_2	x_3	category
1	3	1	0	1	3	0	0	1	0	0	non-chimeric non-fugue
2	3	1	1	1	0	3	0	0	1	0	non-chimeric fugue
3	3	2	1	1	1	2	0	1	1	0	chimeric non-fugue
4	3	2	2	1	0	1	2	0	1	1	chimeric fugue
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

is not zero (e.g. $\boldsymbol{p}_1 = [.9, .05, .05]$) such that the probability concentration bleeds into outcomes with hopped reads (see Table 3). That is, we see that the same label we associate with a unique molecule co-occurs with more than one sample index. When there is no index hopping, for example, when **SIHR**$= 0$ (e.g. $\boldsymbol{p}_2 = [0, 1, 0]$), we observe reads only in one of the three samples (i.e. *non-chimeric non-fugues*). In Table 3, we show observations corresponding to only 4 out of the 10 possible outcomes at read count level $r = 3$. However, when r is large, the set of possible outcomes increases exponentially. When $S = 3$, probabilities of observing these outcomes are given by the corresponding terms of the trinomial expansion of $\boldsymbol{p}$.

To better visualize the distribution of possible outcomes, consider the case at the read count level $r = 3$ when reads hop from one sample only, that is, when the true sample of origin is $s = 1$ (See Figure 1). When $r = 3$ the number of coefficients and thus corresponding outcomes (i.e. elementary events) are given by the $(r+1)$'th *triangular number* $t = \binom{r+(S-1)}{(S-1)}$, which here equals $\frac{(3+1)\times(3+2)}{2} = 10$. That is, we group the 10 outcomes, which make up the sample space $\mathcal{Y}_r$ into a set of three categories: (1) Three non-chimeric outcomes, each lying on one of the three $0-$faces (vertices); (2) Six $2-$chimeric outcomes, two lying on each of the three 1-faces (edges); (3) One $3-$chimeric outcome lying in the middle of a single $2-$face (facet) as shown in Figure 1. In general, the number of $m-$faces of an $r-$simplex is

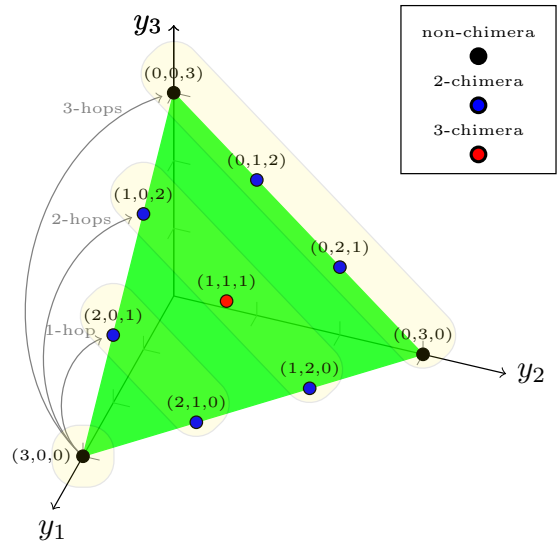


Figure 1: The 3rd component of Pascal's 3-Simplex denoted by $\wedge_3^3$. All 10 possible outcomes (at read count sum r=3) for three samples (S=3) are arranged on the simplex. The sample of origin is $s = 1$ as shown by the arrows. Outcomes can be classified according to the number of observed molecules (i.e. k-chimera) or the number of hopped reads. The four outcomes corresponding to all 3 reads hopping over are called *fugues*. The two fugue outcomes in the vertices are termed *non-chimeric fugues* whereas the other two on the edge are termed *chimeric fugues*

$u = \binom{r+1}{m+1}$. The unit probability measure is split over r outcomes instead of t outcomes. Out of the u $m-$faces, $t = \binom{r}{m}$ correspond to non-fugue outcomes whereas $f = (u - t)$ are fugues (i.e. those observations where all the reads hop from their source sample). That is, if in Figure 1, the true sample of origin is $s = 1$, then a 1-read hop would correspond to two possible outcomes: two 2-chimeric observations each giving rise to one true and one phantom molecule; A 2-read hop would correspond to three possible outcomes: two 2-chimeric observations each giving rise to one true and one phantom molecule and a single 3-chimeric observation giving rise to one true and two phantom molecules; A 3-read hop would would correspond to four possible outcomes: two 2-chimeric fugue observations each giving rise to two phantom molecules and a two non-chimeric fugue observations, each giving rise to one phantom molecule. Note however that when we consider hopping from all S samples, the number of possible outcomes e increases by a factor of S, so when $r = 3$, we have $e = 30$ possible outcomes.

The number of observations in data $\mathcal{Y}$ is typically high (in the hundreds of millions). Each observed vector of read counts can be categorized as a combination of *chimeric/non-chimeric*

and *fugue/non-fugue*. For each k-chimeric observation, we potentially have from one and up to $S-1$ *phantom molecules*. For example, the observation $\boldsymbol{y}_4$ has two *phantom molecule* and zero real molecules. If there is indeed no index hopping, then the number of *phantom molecules* would be zero whereas the number of real molecules would equal to L, the number of unique label combinations. Although the probability that all reads get misassigned to the same sample is negligibly low, the effect can be significant for non-chimeric labels for which there is only one or two reads (a group that makes up a large fraction of observations), as Table 2 shows.

2.1.5 Modeling the Distribution of k-chimeras

The distribution of k-chimeras is essentially the distribution of the total number of non-zero counts of a multinomial random variable. With regards to single cell data, determining whether a sample count is nonzero is equivalent to a deduplication process of obtaining molecule counts from read counts. Statistically, deduplication can be formulated as a thresholded latent variable model where each element of a potentially unobservable random vector $\boldsymbol{y}_l$ is thresholded into an observable Bernoulli random variable z_{ls} with an indicator function $\mathcal{I}$ that detects whether a molecule is observed. That is,

$$\begin{aligned} s_l &\sim \text{Categorical}(\boldsymbol{\pi}_r) \\ \boldsymbol{y}_l | s_l &\sim \text{Multinomial}(r, \boldsymbol{p}_{s_l}) \\ x_{li} &= \mathcal{I}(y_{li} > 0) \ \textit{for} \ l = 1, \ldots, m_r; i = 1, \ldots, S \end{aligned}$$

where $\boldsymbol{y}_l \in \mathbb{N}_0^S,\ \boldsymbol{p} \in [0,1]^S, \|\boldsymbol{p}\|_1 = 1,$ and $S \in \mathbb{N}$

The marginal distribution of each element of the multinomial observation $\boldsymbol{y}_l$ is binomial. That is,

$$y_{li} | s_l \sim \text{Binomial}(r, p^{[s_l = i]} p_h^{(1-[s_l=i])})$$

where the Iverson bracket notation is used. It follows that the distribution of x_{li} is a Bernoulli with a probability parameter that can be obtained by observing that the expectation of an event defined by an indicator function is equal to the probability of that event. That is,

$$x_{li} \sim \text{Bernoulli}(\zeta_i)$$

where

$$
\begin{aligned}
\zeta_i &= \mathrm{E}(x_{li}) = \mathrm{E}_s(\mathrm{E}_{x|s}(x_{li})) \\
&= \mathrm{E}_s(\mathrm{E}_{x|s}(\mathcal{I}(y_{li} > 0))) \\
&= \sum_{j=1}^{S} \mathbb{P}(y_{li} > 0 | s_l = j)\, \mathrm{P}(s_l = j) \\
&= 1 - \sum_{j=1}^{S} \mathbb{P}(y_{li} = 0 | s_l = j)\, \mathrm{P}(s_l = j) \\
&= 1 - \sum_{j=1}^{S} \left(1 - p^{[j=i]} p_h^{(1-[j=i])}\right)^r \pi_j^{(r)} \\
&= 1 - \left(\pi_i^{(r)}(1-p)^r + (1-\pi_i^{(r)})((1-p_h)^r)\right) \\
&= 1 - \pi_i^{(r)} - \left(\left(\frac{1}{1-p} - \frac{1}{S-1}\right)^r \left((1-p)^r - \pi_i^{(r)}\right)\right)
\end{aligned}
$$

Now given that the Bernoulli random variables (i.e. the elements of $\boldsymbol{x}_r$) can be treated as independent (for $r > S$) but not identically distributed, their sum, which indicates the category of the observation (i.e. $k-$chimera), can be given by the Poisson Binomial distribution. That is,

$$
k_r = \mathbf{1}^\intercal \boldsymbol{x}_r \sim PB(\mu_r = \sum_{i=1}^{S} \zeta_i)
$$

and a PMF given by the following recursive formula $\boldsymbol{\zeta}$

$$
\Pr(K = k) = \begin{cases} \prod_{i=1}^{S} (1 - \zeta_i) & k = 0 \\ \frac{1}{k} \sum_{i=1}^{k} (-1)^{i-1} \Pr(K = k - i) T(i) & k > 0 \\ where \;\; T(i) = \sum_{j=1}^{S} \left(\frac{\zeta_j}{1-\zeta_j}\right)^i & \end{cases}
$$

Note the independence assumption does not hold when $r < S$ since $\Pr(K = 0)$ is not zero even though it should be given that the r reads must be belong to at least one sample. Also note that the mean, μ_r, of the Poisson Binomial distribution is equal to the sum of the probabilities. In this case, the $\pi_i^{(r)}$'s do not affect μ_r since they cancel out. That said, they do affect the variance $\sigma_l^2 = \sum_{i=1}^{n} (1 - \zeta_i)\zeta_i$ where it is maximized when the ζ_i's, and

accordingly the $\pi_i^{(r)}$'s, are equal. In this case k_l can be reduced to the sum of dependent Bernoulli variables, which for $r > S$ can be approximated as the sum of independent and identically distributed Bernoulli variables. That is,

$$k_r \sim \text{Binomial}\left(S, 1 - \frac{1}{S}\Big((1-p)^r + (S-1)^{1-r}(S+p-2)^r\Big)\right)$$

The distribution of phantoms can be obtained from k_r by noting that for non-fugue observations, the number of phantoms equals $k_r - 1$. The expected fraction of phantoms at *PCR duplication level* r is therefore approximately as such.

$$\mathrm{E}(f_l) = S - 1 - (1-p)^r - (S-1)^{1-r}(S+p-2)^r \tag{3}$$

Figure 2 plots the functional relationship between the expected fraction of phantoms and the variables p, r, and S.

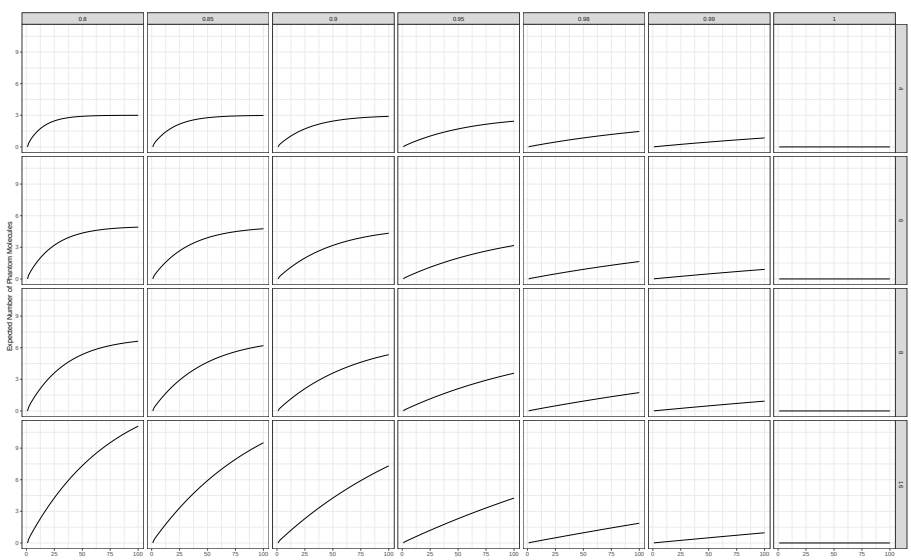


Figure 2: The expected number of phantoms for an observation with *PCR duplication level* r for a range of values of p and number of samples S

Distribution of non-chimeras For the case of $k = 1$, or non-chimeric observations, we can derive a closed form of the distribution by noting that a *non-chimera* is a count observation $\boldsymbol{y}_l$ for which $(y_{li} = r)$ for any $i \in \{1, \ldots, S\}$. We denote the event of observing a

non-chimera by a Bernoulli random variable $w_l := \mathcal{I}((y_{li} = r))$ with parameter p_w given by

$$\begin{aligned}
p_w &= \mathrm{E}(\mathcal{I}((y_{li} = r)) = \mathrm{P}(y_{lji} = r) \\
&= \sum_{j=1}^{S}\sum_{i=1}^{S} \mathrm{P}(y_{li} = r|(s_l = j))\,\mathrm{P}(s_l = j) \\
&= \sum_{j=1}^{S}\sum_{i=1}^{S} \left(p^{[j=i]} p_h^{(1-[j=i])}\right)^r \pi_j^{(r)} \\
&= \sum_{j=1}^{S} [p^r + (S-1) \times \left(\frac{1-p}{S-1}\right)^r] \pi_j^{(r)} \\
&= p^r + (S-1) \times \left(\frac{1-p}{S-1}\right)^r
\end{aligned}$$

As a result, the distribution of observing a *non-chimera* is given by

$$w_l|r \sim \mathrm{Bernoulli}\Big(p^r + (S-1) \times \left(\frac{1-p}{S-1}\right)^r\Big)$$

where $p_{f_0}(r) = (S-1) \times \left(\frac{1-p}{S-1}\right)^r$ is the probability of observing a non-chimeric fugue observation with r reads and p^r is the probability of observing a non-chimeric non-fugue observation with r reads.

2.2 Estimation of the Sample Index Hopping Rate

In the preceding subsection, we were able, with the aid of two simplifying assumptions, to achieve our first goal of building a generative model that probabilistically describes the phenomenon index hopping of multiplexed sample reads. Now we turn to our second goal of validating the model and estimating the index hopping rate from empirical experimental data. However, given the large number of observations that are commonly encountered in practice, we proceed to simplify the problem analytically by reducing the mixture model to a more tractable single parameter model for the distribution of non-chimeric observations. In what follows, we derive the distribution of the sum of non-chimeric observations and show, after making a third simplifying assumption, that the distribution's mean function corresponds to the solution of a differential equation governing a negative growth process. Then, we show how the index hopping rate can be estimated by formulating the problem as generalized linear regression model for binomial counts with a log link function that corresponds to the solution of the differential equation describing negative growth.

2.2.1 Distribution of Sum of non-Chimeras

Given that p_w is the same for all observations with the same *PCR duplication level* , we can sum all the m_r Bernoulli random variables in *PCR duplication level* r to obtain a Binomial distribution over the number of *non-chimeras* z_r conditional on r. That is, for a given r we have

$$z_r = \sum_{l=1}^{m_r} w_l \sim \text{Binomial}\left(m_r,\ p^r + (S-1) \times \left(\frac{1-p}{S-1}\right)^r\right)$$

The joint sampling distribution of the chimeras at all *PCR duplication level* values, concatenated as a vector $\boldsymbol{z}$, can be decomposed as follows

$$\begin{aligned} \mathrm{P}(\boldsymbol{z}|\theta) &= \prod_{r=1}^{R} \mathrm{P}(z_r|m_r, f(p,r,S)) \\ &= \prod_{r=1}^{R} \text{Binomial}\left(z_r|m_r,\ p^r + (S-1) \times \left(\frac{1-p}{S-1}\right)^r\right) \end{aligned} \tag{4}$$

Assumption 3 (The Negligible Contribution of the Probability of Observing non-Chimeric Fugues)**.** Given that the *the sample index hopping rate* we observe in experimental data tends to be very small, $SIHR < 0.05$, the contribution of the second term (i.e. the probability of non-Chimeric fugues) in the parameter of Model 4 is discernible only when $r \leq 2$.

Assumption 3 allows us to simplify Model 4 as follows.

$$\begin{aligned} \mathrm{P}(\boldsymbol{z}|\theta) &= \prod_{r=1}^{R} \mathrm{P}(z_r|m_r, \mu(p,r)) \\ &= \prod_{r=1}^{R} \text{Binomial}\big(z_r|m_r,\ p^r\big) \end{aligned}$$

Note that the mean function $\mu(r) = p^r$ can be expressed as such.

$$\mu(r) = p^r = e^{r\ log(p)} = e^{\beta r}$$

which corresponds to the integral curve solution of the differential equation

$$\frac{d\mu(r)}{d\ r} = \beta\mu(r)$$

This is a negative growth curve since $\beta = log(p) \in (-\infty, 0)$. We can estimate β by formulating the problem as a generalized linear regression for binomial counts with a systematic component $\eta = \beta r$ and a log link function $g(\mu) = log(\mu)$ such that

$$\mu = g^{-1}(\eta) = e^{\eta}$$

Note that the standard link function for binomial counts is the *logit* $g(\mu) = log(\frac{\mu}{1-\mu})$. However, it corresponds to the solution of another differential equation, namely

$$\frac{d\mu(r)}{dr} = \beta\ \mu(r)\big(1 - \mu(r)\big)$$

Although the log link function is not symmetric, η is defined only on the negative real line. Furthermore, the *log* link renders the parameters interpretable since β is basically just the log of p the parameter of interest.

The relationship between the number of non-chimeras z_r and the sample index hopping rate $(1-p)$ at a given *PCR duplication level* r and with an m_r number of observations can be formulated as a *generalized linear regression* model with a *log* link function as follows.

$$\begin{aligned} z_r &\sim \text{Binomial}\big(z_r|\ N = m_r,\ \mu = e^{\beta r}\big) \\ \textit{where} \quad z_r, m_r &\in \mathbb{N}_+;\ \beta = log(p) \in \mathbb{R}_-;\ r = 2, \ldots R. \end{aligned} \tag{5}$$

An estimate of the sample index hopping rate can be obtained from the regression coefficient as such

$$\begin{aligned} \textbf{SIHR} &= 1 - \hat{p} \\ &= 1 - exp(\hat{\beta}) \end{aligned}$$

We can use the estimate of the sample index hopping rate to obtain the expected number of hopped reads at each *PCR duplication level*, namely ($m_r \times r \times$ **SIHR**). We can also plug it into Equation 3 to compute the expected Fraction of Phantom Molecules (**FPM**) for the entire dataset. That is

$$\begin{aligned} \textbf{FPM} &= \mathrm{E}_m(\mathrm{E}(f_r)) \\ &= \frac{\sum_{r=1}^{R} m_r \times \mathrm{E}(f_r)}{\sum_{r=1}^{R} m_r} \\ &= \frac{\sum_{r=1}^{R} m_r \times \Big(S - 1 + (1-\hat{p})^r + (S-1)^{r-1}(S + \hat{p} - 2)^r\Big)}{\sum_{r=1}^{R} m_r} \\ &= \boldsymbol{m}_p^{\intercal}\ \boldsymbol{f} \end{aligned}$$

Here $\boldsymbol{m}_p = \frac{\boldsymbol{m}}{\|\boldsymbol{m}\|_1}$ where $\boldsymbol{m} = (m_1, \ldots, m_R)$ and $\boldsymbol{f} = (\mathrm{E}(f_1), \ldots, \mathrm{E}(f_R))$.

We would like to note that we can do away with Assumption 3 and estimate p in Model 4 using a likelihood optimization procedure. However, the optimization algorithm might not converge or it can get stuck at a local optimum. Furthermore, compared to the GLM approach, it is not straightforward to obtain standard errors or any measure of uncertainty by optimizing the likelihood. We provide code in the paper's GitHub repository showing the likelihood optimization approach and its overall similar results to the GLM approach.

2.2.2 Estimating the Sample Barcode Index Hopping Rate

A sample in a 10X Genomics experiment is actually labelled by, not one, but four different barcodes. By considering only the sample labels instead of the sample barcode, the statistical approach outlined in this paper models sample index hopping across samples rather than across individual barcodes. Nonetheless, we can still estimate the sample barcode index hopping rate by noting that when we treat all reads of a given sample as having the same sample index, we are effectively collapsing a $4 \times S$-dimensional multinomial vector of barcode sample index hopping probabilities $\boldsymbol{p}_b$ into an S-dimensional vector of sample index hopping probabilities $\boldsymbol{p}$. That is, since the read counts are modeled with a multinomial, we can fuse the four barcode count events within a sample by summing their read counts and corresponding probabilities. Accordingly, the relationship between the two parameters of interest p_b and p can be expressed as such.

$$\begin{aligned} p &= p_b + \frac{3(1-p_b)}{4S-1} \\ &= p_b\Big(1 - \frac{3}{4S-1}\Big) + \frac{3}{4S-1} \end{aligned}$$

where p_b is probability that a given sample barcode index does not hop. Accordingly, we can obtain an estimate of the sample barcode index hopping rate by expressing p_b in terms of p and S.

$$\textbf{SBIHR} = 1 - p_b = (1-p)\Big(\frac{4S-1}{4S-4}\Big)$$

Notice that the multiplicative factor is at its maximum (i.e. 1.75) when $S = 2$. As S increases however, it approaches 1 and the two terms become approximately equal.

2.3 Inferring the True Sample of Origin

In the multinomial mixture model (Model 2), we denote the true sample of origin of a sample index associated with a read count by the latent variable $\boldsymbol{s_l}$. One important advantage of formulating the process of sample index hopping as a probabilistic model is that we can now derive, via Bayes theorem, the posterior probability distribution of the unobservable random variable $\boldsymbol{s_l}$ and quantify our residual uncertainty about the true sample of origin given the observed read count and the set of parameters p and the molecular proportions $\Pi := \{\boldsymbol{\pi}_r\}_{r=1}^{R}$ that govern the mixture model.

To derive the posterior distribution of s, we would need the plug-in estimates for the p and Π. In Subsection 2.2, we showed how to estimate p. In what follows, we derive a formula that allows us to obtain estimates of the molecular proportion parameter vectors Π from the distribution of read counts conditional on the PCR duplication level.

2.3.1 Estimating the Molecular Proportions Complexity Profile

Let $\boldsymbol{u}_r = (u_{r1}, \dots u_{rS})(= \sum_l^{m_r} \boldsymbol{s}_l)$ represent the unobserved molecule (i.e. UMIs) counts at *PCR duplication level* r across S samples. The distribution of $\boldsymbol{u}_r$ can be obtained by noting that sum of m_r identically distributed categorical random variables with parameter vector $\boldsymbol{\pi}_r$ is a multinomial random variable with the same parameter vector. That is,

$$\boldsymbol{u}_r = \sum_l^{m_r} \boldsymbol{s}_l \sim \text{Multinomial}(m_r, \boldsymbol{\pi}_r)$$

Now, the expected molecule count across samples in the entire data is given as such.

$$\text{E} \sum_{r \in \mathcal{R}} \boldsymbol{u}_r = \sum_{r \in \mathcal{R}} (m_r \times \boldsymbol{\pi}_r)$$

However, we do not observe $\boldsymbol{s}_l$, and the expected total number of observed molecules in the data can be much greater due to the presence of phantom molecules brought about by sample index hopping.

Let $\boldsymbol{v}_r = \sum_l^{m_r} \boldsymbol{y}_l$ denote the total read count at *PCR duplication level* r. In what follows, it helps to think of the m_r observations as members of S partitions, where the size of partition s is given by π_{rs}, the proportion of molecules in Sample s at *PCR duplication level* r. Now since

$$\begin{aligned} \text{E}(\boldsymbol{v}_r) = \text{E}(\sum_l^{m_r} \boldsymbol{y}_l) &= \sum_l^{m_r} \text{E}_s(\text{E}_{y|s}(\boldsymbol{y}_l)) \\ &= (m_r \times r) \sum_{s=1}^{S} \boldsymbol{p}_s \pi_{rs} \end{aligned}$$

therefore

$$\text{E}(v_{rs}) = (m_r \times r)\Big(\frac{\pi_{rs}(S \times p - 1) + (1 - p)}{S - 1}\Big)$$

and so we can obtain an estimate for π_{rs} from the observed proportion of read counts $\bar{v}_{rs}$ for sample s observations at *PCR duplication level* r. That is,

$$\hat{\pi}_{rs} = \frac{\bar{v}_{rs}(S - 1) + (p - 1)}{(S \times p - 1)} \tag{6}$$

Notice that when there is no sample index hopping (i.e. $p = 1$), the proportion of molecules equals the proportion of reads, that is $\hat{\pi}_{rs} = \bar{v}_{rs}$. Also, note that for the denominator to be positive, the proportion of reads $\bar{v}_{rs} > \frac{1-p}{S-1}$ for all $s = 1, \dots, S$. In empirical data, when this relationship is violated, we can set $\hat{\pi}_{rs} = 10^{-6}$ and renormalize $\hat{\boldsymbol{\pi}}_r$ accordingly.

2.3.2 The Posterior Distribution of the True Sample of Origin

The posterior distribution can be obtained as follows.

$$\begin{aligned}
\mathbb{P}(\boldsymbol{s_l} \mid \boldsymbol{y}_l;\ r, \hat{p}, \hat{\boldsymbol{\pi}}_r) &= \frac{\mathbb{P}(\boldsymbol{y}_l \mid \boldsymbol{s_l};\ r, \hat{p})\mathbb{P}(\boldsymbol{s_l}; \hat{\boldsymbol{\pi}}_r)}{\mathbb{P}(\boldsymbol{y}_l)} \\
&= \frac{\text{Multinomial}(\boldsymbol{y}_l \mid \boldsymbol{s_l}; r, \hat{\boldsymbol{p}}_{s_l})\,\text{Categorical}(\boldsymbol{s_l}; \hat{\boldsymbol{\pi}}_r)}{\sum_{s=1}^{S} \text{Multinomial}(\boldsymbol{y}_l \mid \boldsymbol{s_l}; r, \hat{\boldsymbol{p}}_{s_l})\,\text{Categorical}(\boldsymbol{s_l}; \hat{\boldsymbol{\pi}}_r)} \\
&= \frac{\left(\hat{\boldsymbol{p}}_{s_l}^{\boldsymbol{y}_l}\right)\left(\hat{\boldsymbol{\pi}}_r^{\boldsymbol{s}_l}\right)}{\sum_{s=1}^{S}\left(\hat{\boldsymbol{p}}_{s_l}^{\boldsymbol{y}_l}\right)\left(\hat{\boldsymbol{\pi}}_r^{\boldsymbol{s}_l}\right)}
\end{aligned}$$

which for each element, simplifies to

$$\mathbb{P}(\boldsymbol{s_l} = s \mid \boldsymbol{y}_l;\ r, \hat{p}, \hat{\boldsymbol{\pi}}_r) = \frac{\left(\frac{S-1}{\frac{1}{\hat{p}}-1}\right)^{y_{ls}} \hat{\pi}_{rs}}{\sum_{s=1}^{S}\left(\frac{S-1}{\frac{1}{\hat{p}}-1}\right)^{y_{ls}} \hat{\pi}_{rs}}$$

where the plug-in estimates $\hat{\pi}_{rs}$ is the proportion of molecules in sample s at PCR duplication level r as computed by Equation 6 and $\hat{p}$ is the complement of the sample index hopping rate as estimated by Model 5.

We label the sample with the maximum posterior probability as the true sample of origin. That is

$$s_l^{(t)} = argmax(\mathbb{P}(\boldsymbol{s_l} \mid \boldsymbol{y}_l; r, \hat{p}, \hat{\boldsymbol{\pi}}_r)$$

The corresponding posterior probability of $s_l^{(t)}$ is simply the maximum over the posterior probability vector, that is

$$q_l|\boldsymbol{y}_l = max(\mathbb{P}(\boldsymbol{s_l} \mid \boldsymbol{y}_l; r, \hat{p}, \hat{\boldsymbol{\pi}}_r)$$

When $\hat{\boldsymbol{\pi}}_r$ is uniform, the maximum $q_l|\boldsymbol{y}_l$ can have duplicated values. However, such an outcome is extremely unlikely given the high variability of the molecular proportions that characterize empirical data and therefore will not be considered.

2.4 Purging Phantom Molecules

For each of the L observations, we label the molecule corresponding to $s_l^{(t)}$ as a real molecule and all the others (with nonzero reads) as phantom molecules. Such a procedure achieves the minimum possible number of false negatives at the expense of false positives. However, it could be well the case that sacrificing a potential real molecule is worth more than retaining

a phantom molecule in the dataset. In scientific applications, most often, a weak signal is preferable to an artifactual signal. That said, to achieve the minimum possible number of false positives, it might be the case that the entire data would need to be discarded. An alternative approach attempts to find the optimal trade-off that minimizes both the false positive and false negative rates.

To optimally minimize the false positive rate, we would need to find the optimal cutoff value q^* below which, observations with $s_l^{(t)}$ that we labeled as real molecules are now relabeled as phantom – in other words, we discard them along with previously labeled phantom molecules. In order to determine q^*, we would need to work with the marginal posterior cumulative distribution function of q, which does not have a closed-form expression but which can be expressed as such.

$$\begin{aligned}\mathbb{P}(q) &= P(q|\boldsymbol{y})\mathbb{P}(\boldsymbol{y}) \\ &= \sum_{o\in\wedge^S} \mathbb{P}(q|\boldsymbol{y}=o)\mathbb{P}(\boldsymbol{y}=o) \\ &= \sum_{o\in\mathcal{O}} \sum_{r\in\{1,\ldots,I\}} \mathbb{P}(q|\boldsymbol{y}=o^{(r)})\mathbb{P}(\boldsymbol{y}=o^{(r)}|n=r)\mathbb{P}(n=r) \qquad (7)\end{aligned}$$

Here, $\mathcal{O}$ consists of all the outcomes $\boldsymbol{y}$ that correspond to the coefficients in r'th component of the Pascal's S-Simplex $\wedge_r^S$. Whereas $\mathbb{P}(\boldsymbol{y}=o^{(r)}|n)$ is given by the multinomial distribution can be computed numerically for the first dozen values of *PCR duplication level* r, the marginal distribution of r, $\mathbb{P}(n)$, can be approximated by the observed proportion of *PCR duplication levels* which we denoted by $M(.)$ in Model 2. However, since r can have large values in empirical data, we can work with the empirical distribution instead and approximate the marginal distribution of $P(\boldsymbol{y})$ by the observed relative frequency distribution of outcomes (e.g. the proportion of $\boldsymbol{y}=(1,2,0,0)$ observations in the data).

2.4.1 Determining the Optimal False Positive Rate

The approach outlined in the previous section can be thought of as a classification task in which we attempt to predict whether a given molecule is a phantom molecule, which we consequently discard, or a real molecule, in which case we retain. We would like to minimize both the error that a molecule we deem real is in fact a phantom (i.e. false positive) and the error that a molecule we deem phantom is in fact a real molecule (i.e. false negative). In a dataset consisting of L observations and M molecules, the maximum number of real molecules is L. In what follows, it helps to work with proportions relative to the number of observations L. Accordingly, we define $u=\frac{M-L}{L}$ as the *molecule inflation factor*, a lower limit on the fraction of molecules we can predict to be phantom (with respect to L). Although the initial total number of molecules in the entire data before index hopping is L, the upper limit of the number of real molecules we are able to predict as real is slightly less than L due to the existence of fugue observations in the dataset, each one of which

would lead to one extra true phantom molecule to be unaccounted for. We term these molecules as *fugue phantoms*. L therefore would equal the sum of real molecules and fugue phantoms. Accordingly, if g is the proportion of fugue observations in the dataset, which we can obtain from Equation 10 (see below), then there would be a total of $(u + g)$ true phantom molecules and a total of $(1 - g)$ true real molecules in the data. For example, for a dataset with $L = 100$ observation, a total number of molecules $M = 150$, and an estimated proportion of fugues $g = 0.05$, the *molecule inflation factor* u would equal 0.5 so that the total number of true phantom molecules is given by $(0.5 + 0.05) \times L = 55$ and the total number of real molecules is given by $(1 - 0.05) \times L = 95$.

In our classification task, we label the molecule with the maximum posterior probability (q) above or equal to a given threshold $q*$ as a real molecule while the remaining are all labelled as phantoms. Furthermore, all the molecules in observations whose q falls below the cutoff are also labeled as predicted phantoms even though a proportion of them are real molecules which we cannot confidently classify (i.e. false negatives). All molecules we label as phantom are eliminated from the dataset.

Table 4: The confusion matrix for classifying molecules as real or as phantom. Given a data table of read counts with L observations, the total number of molecules would be $(1 + u) \times L$ where u is the *molecule inflation factor*. The variable o^* denotes the proportion of observations that fall above threshold q^* and such that $o^* \times L$ would be the number of molecules we label as real. The constant g is the proportion of fugue observations estimated for that given dataset.

		Prediction		
		Phantom	**Real**	**Total**
Truth	**Phantom**	TN $=u + g-$FP	FP(qr^*)	u $+g$
	Real	FN $=$FP $+1 - o^* - g$	TP $= o^*$ - FP	1-g
	Total	u+1-o^*	o^*	1+u

In what follows, it would be easier to work with the complement of q, which we denote $qr = 1 - q$, which is the probability of a molecule originating from a sample other than the one with maximum posterior probability. For a given selected threshold value qr^*, we let $o^* = F_{qr}(qr^*) := \Pr(qr \leq qr^*)$ denote the proportion of observations whose corresponding predicted real molecules we retain. Here, F_{qr} is the empirical CDF of qr. Consequently, the fraction of predicted real molecules is o^* and the fraction of predicted phantom molecules

is $1 + u - o^*$, since the total fraction of molecules must sum up to $u + 1$ (see Table 4). Out of the (o^*) predicted real molecules, the proportion of false positives is given by the expectation of the probability of false assignment over the subset of the data o^*

$$\begin{aligned} FP(qr^*) &= \int_0^{qr^*} 1 - F_{qr}(qr) \ \mathrm{d}(qr). \\ &= \int_0^{qr^*} qr \ f_{qr}(qr) \ \mathrm{d}(qr). \end{aligned}$$

Given that the values of qr tend to concentrate close to zero, we provide the following additional transformed score that is easier to work with when plotting graphs.

$$qs = 160 + 10 \log_{10}(qr + 10^{-16})$$

Here qs ranges from 0 to 160 as qr ranges from 0 to 1.

Note that the empirical cumulative distribution function of qr is discrete (i.e. its ECDF increases by jump discontinuities only) even though it takes real values in $\mathbb{R}^{[0,1]}$. Accordingly, the threshold qr^* should be set such that very few potential real molecules are discarded for any marginal decrease in false positives. The trade-off can be made clearer to see when we plot the **FNR** against the **FPR**, where (**FPR**) $:= \frac{FP}{(u+g)}$ and **FNR** $:= \frac{FN}{(1-g)} = \frac{o^*+FP-g}{(1-g)}$. The optimal value for the threshold q^* is the one that simultaneously minimizes both the **FPR** and **FNR** (i.e. the coordinate that lies closest to the origin), but note that discarding lower quality data would lead to a small **FPR** but also a large **FNR**. We can use *Youden's* **J** statistic = 1- (**FPR** +**FNR**) to find the optimal cut-off value $q*$ that maximizes **J**. The cut-off given by *Youden's* **J** is optimal in the sense that it minimizes the probability of random guessing when we give equal weight to **FPR** and **FNR**, or equivalently to sensitivity and specificity (Youden, 1950).

2.4.2 Computing the Proportion of Fugue Observations (g)

To obtain the proportion of all fugues $\mathbf{g}$, we need to take into account the proportion of chimeric fugues $\mathbf{g_c}$ as well the proportion of non-chimeric fugues, which we denote by $\mathbf{g_0}$. We can compute $\mathbf{g_0}$ by multiplying the observed relative frequency proportion vector of the read count levels (i.e. $\overline{\boldsymbol{m}}$) by the vector of non-chimeric fugue probabilities, whose elements are given by $p_{f_0}(r)$ and which we derived in Section 2.1.5.

$$\begin{aligned} \mathbf{g_0} = \mathrm{E}_r \left(p_{f_0}(r) \right) &= \sum_{r=1}^{R} \overline{m}_r \times (p_{f_0}(r)) \\ &= \sum_{r=1}^{R} \overline{m}_r \times (S-1) \times \left(\frac{1-p}{S-1} \right)^r \\ &= \overline{\boldsymbol{m}}^{\intercal} \ \boldsymbol{p_{f_0}} \end{aligned} \tag{8}$$

The probability of observing chimeric fugues can be obtained as such.

$$p_{f_c}(r) = \sum_{\boldsymbol{y} \in \mathcal{T}_r} \binom{r}{\boldsymbol{y}} \times \left(\frac{1-p}{S-1}\right)^r \tag{9}$$

where the set $\mathcal{T}_r$ consists of outcomes corresponding to events where all the r reads hopped to two or more target samples: the outcomes (0, 2, 1) and (0, 1, 2) in Figure 1. For $r = 1$, the set $\mathcal{T}_r$ is empty; for $r = 2$, there are $\binom{S-1}{r} = \frac{(S-1)(S-2)}{2}$ such outcomes. For $r > 3$, it becomes increasing hard to enumerate and compute the probabilities for all the possible outcomes, especially when S is large as well. However, the value of $\left(\frac{1-p}{S-1}\right)^r$ becomes negligibly small for $r > 3$ given that p tends to be very close to 1. Furthermore, empirical data is characterized by a highly skewed distribution of *PCR duplication levels* such that the proportion of observations in the first few r values makes up a substantial fraction of all observations. Therefore even if $p_{f_c}(r)$ was not negligible for $r > 3$, it would ultimately be weighted down by the respective $\overline{m}_r$. Under this assumption, the proportion of fugue observations can therefore be accurately approximated as such.

$$\begin{aligned}
\mathbf{g} &= \mathbf{g_0} + \mathbf{g_c} \\
&= \sum_{r=1}^{R} \overline{m}_r \times \big(p_{f_0}(r) + p_{f_c}(r)\big) \\
&\approx \overline{m}_1 (S-1)\big(\frac{1-p}{S-1}\big) \\
&\quad + \ \overline{m}_2 \Big((S-1)\big(\frac{1-p}{S-1}\big)^2 + 2\frac{(S-1)(S-2)}{2}\big(\frac{1-p}{S-1}\big)^2\Big) \\
&\quad + \cdots \\
&= \overline{m}_1(1-p) + \overline{m}_2(1-p)^2 + \overline{m}_3(1-p)^3 + \cdots
\end{aligned} \tag{10}$$

For example when $p = .99$, $\overline{m}_1 = 0.20$, $\overline{m}_2 = 0.15$, and $\overline{m}_3 = 0.1$ $\mathbf{g} = 0.0020151$. Or in other words, only 0.2 % of observations are fugues. If all the data is restricted to have *PCR duplication level* $r = 1$ (i.e. $\overline{m}_1 = 1$), then $\mathbf{g} = (1-p)$ would be the maximum possible proportion of fugue observations.

3 Application on Empirical Data

3.1 Description of Datasets

We applied the proposed model on three publicly available 10X Genomics scRNA-seq datasets: (1) a control non-multiplexed dataset in which each sample was sequenced on a separate lane of HiSeq 2500; (2) a multiplexed dataset sequenced on HiSeq 4000; and (3) a multiplexed dataset sequenced on NovaSeq 6000. The HiSeq 2500 and HiSeq 4000 datasets consist of 8 libraries of mouse epithelial cells, which have also been used previously by Griffiths *et al.* (2018) for analysis of sample index hopping. The two datasets were downloaded from the authors' host server using the *get_ data.sh* script available on our paper's GitHub repo (for a more detailed description of the data, please refer to the original publication, Bach *et al.* (2017)). The third dataset was obtained from the Tabula Muris project's repository. It consisted of 16 libraries (i.e. the P7 libraries) of mouse tissue samples, which were pooled and multiplexed on two lanes of an S2 flowcell in a single NovaSeq 6000 sequencing run (Consortium, 2018). BAM files for the 16 samples were downloaded from the NIH's SRA data repository (SRA accession number SRP131661) and converted back to FASTQ files using the 10X Genomics *bamtofastq* utility. Cell Ranger 3.0.0 was run with the default options on each set of samples multiplexed on the same lane.

3.2 Description of Data Preprocessing and Analysis

Here we provide a rough summary of the data analysis steps we performed on the datasets. We also provide the code implementing our proposed approach and a set of reproducible R Markdown notebooks detailing the data analysis steps on the paper's GitHub repository.

Step 1: creating a joined read counts table. In order to create the joined data table of gene-mapped read counts across samples, we loaded from the *molecule_ info.h5* file of each sample the data corresponding to the following four fields: *cell-barcode*, *gene*, *umi*, and *reads*. We then joined the data from all samples into a single data table that is keyed by a unique *cell barcode-gene-umi* combination ID. In addition to the unique label ID, each row of the data table contained read counts across all the S samples. Given that the number of observations in the joined read counts data table tends to be in the hundreds of millions, we then collapsed observations with similar outcomes since working with a data table of outcome tallies drastically reduces the number of computations required. We also added a few grouping variables to the joined data table of outcome tallies that are needed for subsequent analysis steps. The resulting data table contained the following additional variables: an outcome character variable (e.g. "(1,0,1,1)") to group the observations into unique outcomes, a count variable (n) denoting the number of times a particular outcome was observed in the data, a *PCR duplication level* variable (r), and a ($k_chimera$) variable that counts the number of molecules we observe in each outcome.

Step 2: estimating the sample index hopping rate. Using the introduced grouping variable, we tallied the number of observations that are chimeric and those that are non-chimeric at each *PCR duplication level* r. The resulting data table is then used as input to the GLM Model 5 in order to estimate, from the proportion of chimeric observations, the sample index hopping rate.

Step 3: inferring the true sample of origin and reassigning reads. To compute the posterior probability of the true sample index (the latent variable s) for each unique observed outcome, we first summarized from the data the conditional and marginal distributions of reads. The summaries are represented by a vector of *PCR duplication level* proportions $\overline{\boldsymbol{m}}$, which we used to compute the empirical marginal distribution of q, and a set of I across-samples read count proportion vectors from which we estimated the proportion of molecules across samples $\hat{\boldsymbol{\pi}}_r$, one at each *PCR duplication level* r, that in turn we plugged in to calculate the conditional posterior probability $q|\boldsymbol{y}$. For each outcome, the posterior probabilities of the possible true samples of origin are computed and the index of the sample with the maximum posterior probability along with posterior probability itself is added to the original joined read count table. We then used the predicted true sample of origin and its associated posterior probability to reassign reads to thier predicted sample of origin.

Step 4: computing an optimal cutoff for purging predicted phantom molecules In order to remove predicted phantom molecules from the data while minimizing the rate of false positives and false negatives, we computed *Youden's* **J** statistic and used the corresponding $q*$ cutoff to retain predicted real molecules. Before purging the data, we first deduplicated the read counts to obtain a table of molecule (i.e. UMI) counts. After filtering, we collapsed the molecule counts over gene labels to produce a gene-by-cell umi-count expression matrices for all the samples sequenced in the same lane.

Step 5: identifying RNA-containing cells before and after purging We ran the cell-calling algorithm *EmptyDrops* (Lun *et al.*, 2019) to distinguish cells from empty droplets. We used the set of called cells in each sample to filter out background cell-barcodes from the deduplicated purged UMI count matrices. The filtered matrices were then saved to file in order to be used for downstream statistical analyses. By comparing the results of the cell-calling algorithm from both purged and unpurged data, we were able to examine the extent of contamination of phantom molecules on data that would have otherwise not been purged.

3.3 Results

3.3.1 Molecular Proportions Complexity Profiles

Figure 3 depicts the molecular proportions complexity profile plots for all four datasets. The same three samples (B1, B2, and D2) in both the Hiseq 2500 and Hiseq 4000 datasets show molecular proportion curves indicative of low library complexity. Indeed even though the number of reads is approximately equal across the 8 samples, the number of molecules differ drastically, by an order of 100 or more, and are lowest in these three samples (see analyses notebooks). Furthermore, whereas the proportion curve of Sample D2 peaks at a moderate PCR duplicate level then tapers off, those of B1 and B2 start picking up and maintain a steady proportion for a wide range of high duplicate level values, up to $r = 1706$ and $r = 3744$, for Hiseq 2500 and Hiseq 4000, respectively. These trends are indicative of problematic samples; incidentally, Samples B1, B2 are the two samples that our analysis identifies to have the largest fraction of phantom molecules. As for the Novaseq 6000 datasets, the curves for Samples $P7_1$ and $P7_8$ indicate low complexity libraries. This is also confirmed by the analysis since both samples turned out to have the two lowest number of molecules but the two highest fractions of phantom molecules among the 16 multiplexed samples.

3.3.2 Estimates of the Sample Index Hopping Rate

The model fits are summarized in Table 5 with estimates of p along with the 99 percent confidence intervals listed for the four datasets. The table also lists the derived estimates of the sample index hopping rates (SIHR) and the sample barcode index hopping rates (SBIHR). As the figures show, the non-multiplexed (HiSeq2500) dataset has a much lower (SIHR) than the 3 multiplexed datasets, whose estimates show high consistency. Notice how a slight increase in the sample index hopping rate translates into a sizable increase in the molecule inflation factor and therefore in the proportion of phantom molecules in the dataset. Although the model was fit on a subset of the data for r ranging between $r = 2$ and $r = 25$, the estimates of the sample index hopping rate are little influenced by the small fraction of observations found at $r > 25$. The small dots on the top of Figure 3 represent an estimate on the lower limit of the proportion of *nonmissing* observations at every PCR duplication level r.

As Figure 4 shows, the GLM fits the four datasets rather well, especially for r values for which there is almost negligible missingness. Also, notice that the observed chimeric proportions show a downward deviation from the predicted mean trend at high *PCR duplication levels*, especially for those top 1 percent of observations that are characterized by noisiness and sparsity. In fact, the downward trend tends to occur right after the cumulative proportion of molecules in the dataset begins to plateau. The downward trend we observe in the proportion of chimeras seems to indicate a slight decrease in hopping rate that kicks in at high PCR duplication levels. The downward trend in the proportion of chimeras that

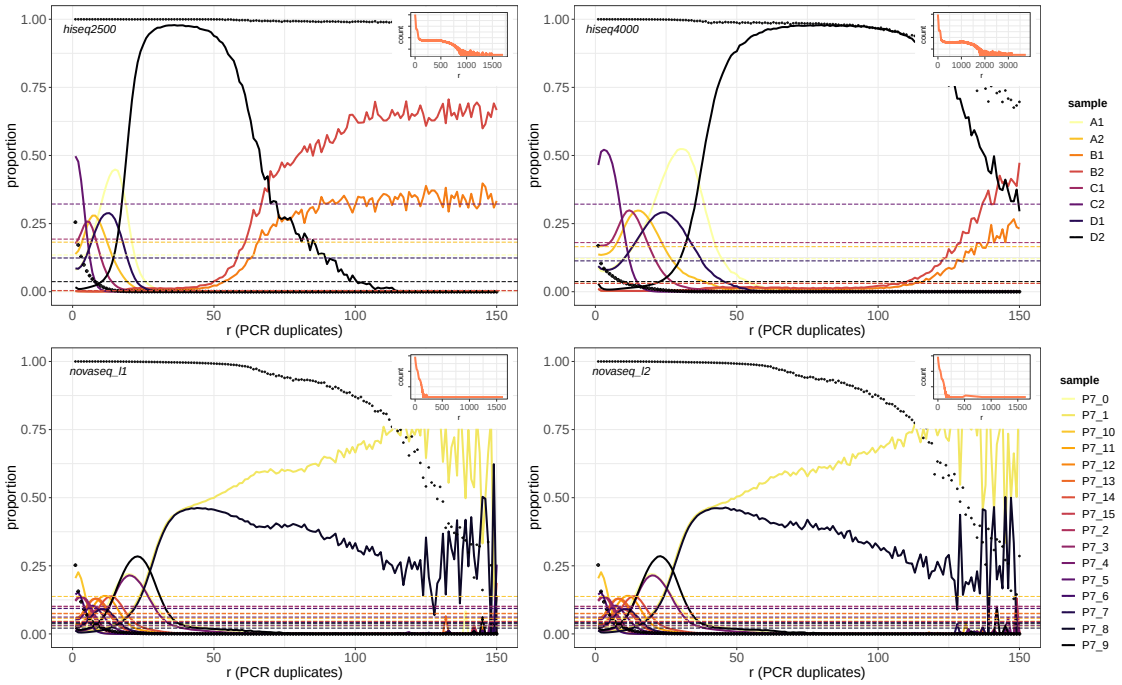


Figure 3: The molecular proportions complexity profile plots for the four datasets. The proportion of library complexity (i.e. number of molecules) across the samples is represented by the horizontal dotted lines. The proportion of observations falling at each PCR duplication level r is represented by a set of dark crossed circles at the bottom.The top line (black dots) represents a lower limit on nonmissingness (the probability associated with outcomes observed). Whereas the plots show r up to a maximum of 150, the corresponding subplots in the upper right corners show the distribution of the number of observations for the entire observed range of r.

we observe to deviate from from the model predicted trend could potentially be explained by sample dropout at high r values. That is, when the molecules of only a subset of the samples are PCR duplicated at high levels, the multinomial model we specified can no longer assume that we have S samples since sample indices cannot hop to nonexistent sample molecules. Accordingly, we expect the curve to differ when we have S samples compared to when we have fewer. Another aspect of the observed chimeric proportions becomes apparent at very high r values. The observed trend becomes noisy and fluctuating. This can be explained by the the high proportion of missingness at those levels. That is, there are finite observations at a given *PCR duplication level* r, and therefore we will observe only a few of the possible outcomes, especially when $r > 25$, where the number of potential outcomes are in the millions. When this happens, the variability of the mean estimate would increase as

Table 5: Estimates of the sample index hopping rate **SIHR** for one non-multiplexed (HiSeq2500) and three multiplexed sample libraries from two experiments sequenced on two different platforms. In the Novaseq dataset, the same sample libraries were multiplexed on two lanes. $\hat{p}$ is the estimate of the model parameter (i.e. the complement of the hopping probability); $m.e.$ is the margin of error corresponding to a 99% confidence interval; u is the molecule inflation factor, a summary measure that represents a lower limit on the number of phantom molecules in the data; g is the proportion of fugue observations; and PF (Phantom Fraction) is the fraction of phantoms in the dataset.

dataset	$\hat{p}$	**SIHR**	**SBIHR**	$m.e.$	u	g	PF
HiSeq2500	0.99989	0.00011	0.00012	$\pm 1.11 \times 10^{-6}$	0.00060	0.00003	0.00062
HiSeq4000	0.99130	0.00870	0.00963	$\pm 9.96 \times 10^{-6}$	0.08169	0.00148	0.07688
NovaSeqL1	0.99179	0.00821	0.00862	$\pm 4.73 \times 10^{-6}$	0.04131	0.00208	0.04167
NovaSeqL2	0.99168	0.00832	0.00873	$\pm 5.25 \times 10^{-6}$	0.04187	0.00211	0.04222

the number of observations at a given r decreases.

3.3.3 The Optimal Cutoff for Minimizing the FPR

The amount of phantom molecules in a given dataset is given by $(u + g)$. Given this information, there are three courses of action we can take. (1) We keep the data as it is. By doing so, we are basically labeling all the phantom molecules as real, which would correspond to $(u + g)$ false positives and an FPR=1, as given by the *asis* rows of Table 6. That is, we are in effect incorrectly classifying all true phantom molecules as real. (2) We purge the phantom molecules by reassigning the reads to the sample with largest posterior probability as described in Subsection 2.3. By doing so, we can decrease the number of false positives, and consequently the false positive rate, to values given by the *none* rows in Table 6. (3) We decrease the FPR further at the cost of a slight marginal increase in the FNR by choosing a cutoff, as described in Subsection 2.4.1, at which predicted real molecules with low q values are classified as phantom instead. Doing so would bring the FPR and FNR to values given by the *optim* rows in Table 6.

Figure 5 shows the ECDF of qs with the corresponding FPR, FNR, and J values superimposed on the plots. As can be seen, the optimal cutoff qs^* that maximizes J occurs at the point when a small marginal decrease in FPR would be offset by a large increase in FPR. In Table 6, the corresponding cutoff values of the posterior probability q below which predicted real molecules are deemed phantom and discarded are shown along with the preceding cutoff value, the one with a higher false positive rate.

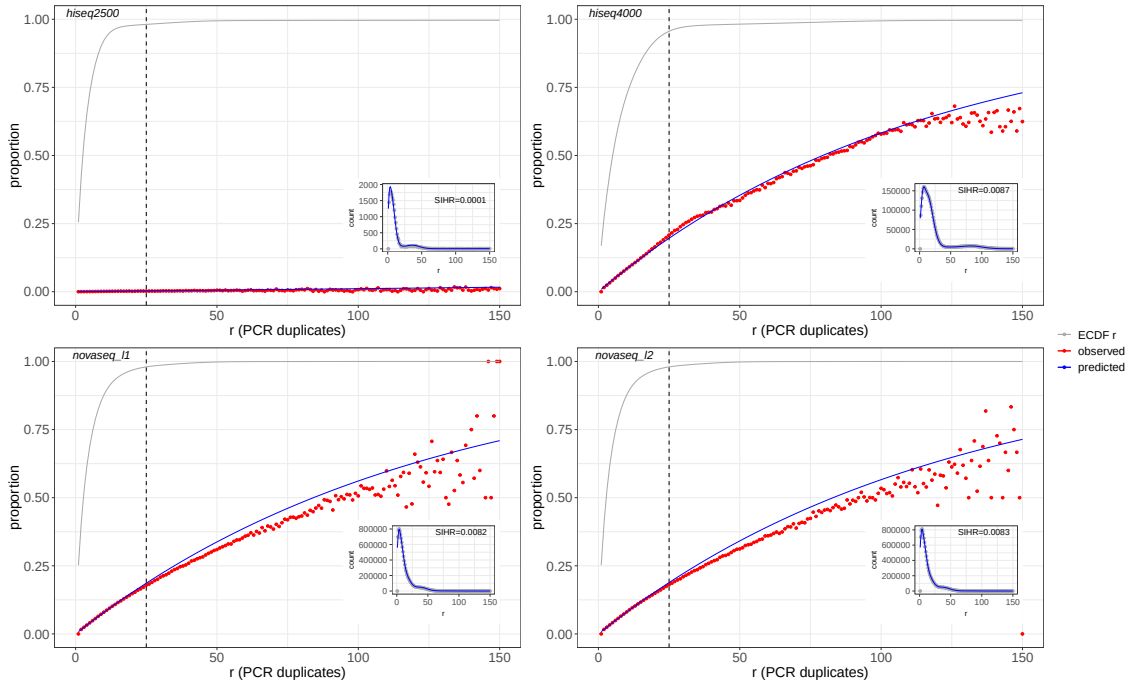


Figure 4: GLM predicted (blue) vs observed (red) proportion of chimeras. The subplots show the fit on the absolute count of chimeras. The vertical dotted line represents the maximum r threshold for fitting the model. The ECDF of r (grey) shows the cumulative proportion observations that are less than or equal to r.

3.3.4 The Effects and Extent of Contamination by Phantom Molecules

If we had chosen not to control the false positive rate with setting a threshold, or in other words if we had set $o^* = 1$, then the fraction of molecules (in multiples of L) we predict as phantom for the entire dataset would be given by u, out of which true phantoms are given by TN and false phantoms by FN. However, the corresponding fractions for individual samples can depart drastically from these marginal measures for the entire dataset as a whole since library complexity varies widely across samples and some samples can consist mostly of empty cells. Indeed this is the case for the Hiseq 4000 dataset. As Table 7 shows, the proportion of phantom molecules for this dataset ranges from 0 .012 to 0.864 whereas it ranges from 0.013 to 0.175 in both of the two Novaseq 6000 datasets (see Supplementary Tables 11 and 12). In contrast, in the Hiseq 2500 dataset, which had not been multiplexed and whose hopping rate we estimated to be extremely small, we see that the range of predicted proportion of phantom molecules ranges from 0.0002 to .02. Note that fraction of reads to molecules (i.e. FRM) tends to be a good indicator of which samples would have a high proportion of phantoms and also which would have a high proportion of false predicted

Table 6: Table showing the optimal cutoff values q, corresponding outcome, inferred sample of origin s, False Positive Rate *FPR*. The cutoff preceding the optimal cutoff and the minimum value of q are also provided for each of the four datasets.

dataset	cutoff	outcome	q	s	FPR	FP	TP	FN	TN
HiSeq2500	optim	0,0,0,0,1,0,0,2	0.99966	8	0.0261	0.0000	0.9946	0.0054	0.0006
HiSeq2500	below	0,0,0,0,0,1,0,2	0.99936	8	0.0261	0.0000	0.9946	0.0054	0.0006
HiSeq2500	none	1,0,0,0,0,0,1,0	0.50174	7	0.0565	0.0000	1.0000	0.0000	0.0006
HiSeq2500	asis				1.0000	0.0006	0.9994	0.0000	0.0000
HiSeq4000	optim	0,0,0,1,1,0,0,1	0.92359	5	0.0131	0.0011	0.9946	0.0039	0.0821
HiSeq4000	below	0,1,2,0,0,0,1,0	0.92281	3	0.0131	0.0011	0.9946	0.0039	0.0821
HiSeq4000	none	0,1,0,0,1,0,1,0	0.43691	5	0.0194	0.0016	0.9984	0.0001	0.0815
HiSeq4000	asis				1.0000	0.0832	0.9183	0.0000	0.0000
NovaSeqL1	optim	0,0,0,0,0,0,0,0,2,0,0,0,0,0,2,0	0.96084	9	0.0447	0.0019	0.9923	0.0056	0.0415
NovaSeqL1	below	0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0	0.95818	9	0.0447	0.0019	0.9924	0.0055	0.0415
NovaSeqL1	none	0,0,0,0,1,0,1,0,1,1,0,0,1,0,0,0	0.27418	10	0.0613	0.0027	0.9973	0.0006	0.0407
NovaSeqL1	asis				1.0000	0.0434	0.9587	0.0000	0.0000
NovaSeqL2	optim	0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0	0.95800	9	0.0448	0.0020	0.9922	0.0056	0.0420
NovaSeqL2	below	0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0	0.95737	15	0.0479	0.0021	0.9953	0.0025	0.0419
NovaSeqL2	none	0,0,0,1,0,0,0,0,1,1,0,1,0,1,0,1	0.23670	10	0.0614	0.0027	0.9973	0.0006	0.0413
NovaSeqL2	asis				1.0000	0.0440	0.9581	0.0000	0.0000

phantoms (i.e. real molecules which we incorrectly classified as phantoms).

Given that a substantial percentage of mapped molecules in the dataset originate from empty droplets, one can argue that what matters in the final analysis is whether there turns out to be a substantial degree of phantom contamination in the subset of molecules originating from non-empty droplets. To assess the extent and effects of contamination by phantom molecules in this subset, we used the EmptyDrops (Lun *et al.*, 2019) algorithm to call cell-barcodes in order to determine whether a cell-barcode originated from a cell or an empty droplet. For each dataset, we ran EmptyDrops the first time on the unpurged data and then a second time of purged data. To compare the results of the two runs, all barcodes classified as either cell or empty by the algorithm were divided into three groups each as shown in Tables 8,9 and 10. In the tables, the *Consensus* columns show the number of cell-barcodes that maintain the same classification as cell or empty, across the two runs. The *Phantom* columns show the number of cell-barcodes whose associated molecules were all predicted as phantom molecules (true and false phantoms). The *Re-classified* columns show the number of barcodes that have been reclassified after running the cell-calling algorithm of purged data. For example, we see that for the B2 sample in the Hiseq 4000 dataset, 1 barcode that was previously classified as background, was re-classified as cell after purging and 655 cell-barcodes that were classified as cell before purging were re-classified as empty after purging. Accordingly, if we had called cells without first purging the dataset of phantom molecules, we would have believed that we have called a total of 940 (= 240+654+14) cells in sample B2, whereas re-running cell-calling on purged data would have produced no more

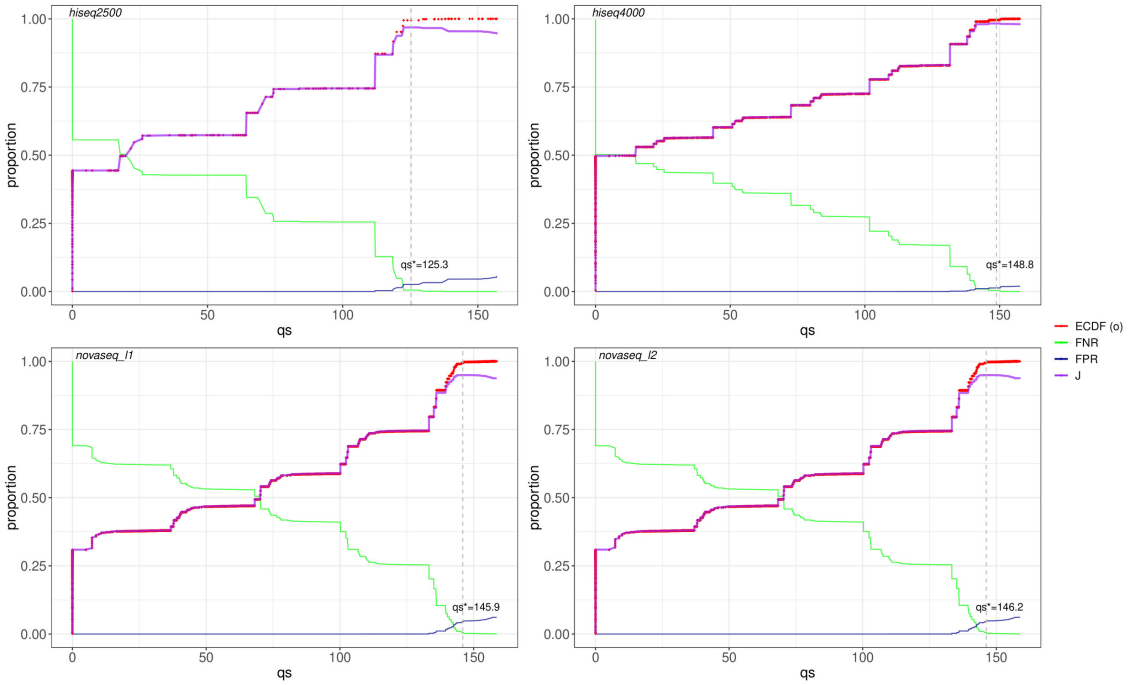


Figure 5: The empirical cumulative distribution function of qs mapping the false assignment probability of a inferred real molecule to a quantile o, representing the cumulative proportion of inferred real molecules whose false assignment probability is less or equal to a given value qs^*. Overlayed on the plots are the cumulative marginal false positive and false negative rates together with *Youden's* **J** = 1- (**FPR** +**FNR**), whose argmax corresponds to the optimal cutoff (vertical dashed line).

than 14 cells. As the tables show, there are two effects that phantom molecules bring about: (1) the emergence of phantom cells and more commonly (2) leading cell-calling algorithms to miss-classify empty droplets as cells.

For the Novaseq 6000 datasets, the effects of contamination are less severe than those see in the Hiseq 4000 dataset. Even though we observe phantom cells only in Sample $P7_8$ (incidentally the sample with the highest FRM in both lanes), for many samples the number of barcodes that are reclassified as empty can make a substantial fraction of total cells (e.g. Sample $P7_01$ with 225 barcodes). Note that we opted to treat the two lanes separately for the cell-calling stage of the analysis to evaluate the consistency of the proposed approach on two datasets that only differ by flowcell sequencing lane. For downstream analysis, data from the two lanes should be combined before cell-calling is performed.

Table 7: Extent of contamination by sample for two Hiseq datasets (a non-multiplexed Hiseq 2500 dataset and mutliplexed Hiseq 4000 dataset). The number of molecules is given in multiples of million. The proportion of total molecules predicted as real and therefore retained is given by the *real* column. The proportion of total molecules that predicted as real but discarded since their q values fall below the optimal threshold is given by the *false phantoms* column. The proportion of molecules predicted as phantom is given by the *phantoms* column. The two proportions columns give show the discrepancy between library size and library complexity across the samples. The last column measures the extent of PCR Amplification bias as given by *FRM* (i.e. the Fraction of total number of Reads over total number of Molecules (i.e. library size divided by library complexity).

Hiseq 2500 Dataset

		Breakdown			**Proportions**		
sample	molecules	real	false phantoms	phantoms	molecules	reads	FRM
A1	5.94	0.9986	0.0000	0.0014	0.135	0.134	6.1
A2	8.02	0.9997	0.0000	0.0003	0.182	0.135	4.5
B1	0.15	0.8288	0.1507	0.0205	0.003	0.126	228.8
B2	0.20	0.8414	0.1451	0.0135	0.005	0.132	177.8
C1	8.51	0.9997	0.0000	0.0003	0.193	0.118	3.7
C2	14.19	0.9998	0.0001	0.0002	0.322	0.122	2.3
D1	5.45	0.9995	0.0001	0.0004	0.124	0.112	5.5
D2	1.63	0.8843	0.1141	0.0016	0.037	0.121	20.1

Hiseq 4000 Dataset

		Breakdown			**Proportions**		
sample	molecules	real	false phantoms	phantoms	molecules	reads	FRM
A1	7.03	0.9662	0.0001	0.0337	0.121	0.131	11.6
A2	9.59	0.9775	0.0003	0.0222	0.165	0.131	8.6
B1	1.82	0.0803	0.0554	0.8643	0.031	0.128	44.1
B2	1.75	0.1143	0.0617	0.8239	0.030	0.140	50.1
C1	10.46	0.9780	0.0005	0.0215	0.180	0.122	7.3
C2	18.65	0.9874	0.0006	0.0120	0.321	0.121	4.0
D1	6.56	0.9684	0.0002	0.0314	0.113	0.107	10.2
D2	2.16	0.8787	0.0006	0.1207	0.037	0.121	35.1

Table 8: Effect of contamination on cell calling.

Hiseq 2500 Dataset						
	Phantom		**Consensus**		**Reclassified as**	
sample	cell	empty	cell	empty	cell	empty
A1	0	1332	241	123938	0	0
A2	0	770	375	139139	0	0
B1*	25	7021	184	20300	75	11
B2	28	9075	584	21350	45	41
C1	0	808	585	145437	0	0
C2	0	668	1096	195839	0	0
D1	0	778	455	99200	0	0
D2	0	20632	135	41721	0	1

Hiseq 4000 Dataset						
	Phantom		**Consensus**		**Reclassified**	
sample	cell	empty	cell	empty	cell	empty
A1	0	12478	401	138115	14	139
A2	0	11203	565	156896	15	105
B1*	270	82361	34	27478	159	930
B2	240	78203	14	32296	1	655
C1	0	11750	730	168480	0	120
C2	0	10087	1320	232402	5	154
D1	0	11611	661	119470	16	64
D2	0	15912	165	82287	7	29

4 Discussion

4.1 Significance of Proposed Approach

As we have mentioned earlier, the strategy Griffiths *et al.* (2018) propose does not in fact estimate the index hopping rate of individual reads, but rather computes a proxy measure, *the swapped fraction*, which they define as the fraction of molecules (i.e. those with an identical UMI, cell barcode, gene label) that are observed in more than one sample. In this paper, these molecules would correspond to *phantom molecules* and the swapped fraction would correspond to *the fraction of phantom molecules* but only in *non-fugue* observations (see definitions in Section 2.1.3). Incidentally, by overlooking the possibility that a given fraction of observations consists entirely of phantom molecules, the molecule exclusion strategy they propose can potentially retain a high proportion of false positives

molecules. Furthermore, by assigning the true sample of origin as the sample with the largest read fraction while discarding all molecules that have a read fraction below a given threshold (default being ≤ 0.8), their strategy would further result in high false positive and high negative rates since neither the information implicit in the absolute read count distribution nor the extremely variable distribution of molecular proportions across samples and *PCR duplication levels* are considered. For example, even if an observation such as $\boldsymbol{y} = (1, 4, 0, 0)$ has the highest read fraction at $s = 2$, it would still be an unlikely event if for instance we have 70 times more molecules in Sample 1 than in Sample 2, that is, $\pi_5 = (0.7, 0.01, 0.19, 0.1)$. This exact situation was encountered in real data as described in Section 3. In particular, note that, for the Hiseq 4000 dataset, q is higher for outcome $y^{(opt)} = (0, 0, 0, 1, 1, 0, 0, 1)$ than it is for $y^{(below)} = (0, 1, 2, 0, 0, 0, 1, 0)$ even though in the latter, the inferred true sample of origin $s = 3$ (Sample B1) has two reads whereas the three molecules in the former has one read each, including the inferred true sample of origin $s = 5$. This might seem counter-intuitive at first sight, but once we consider the proportion of molecules, it becomes apparent why it makes sense to retain the molecule corresponding to $s = 5$ but discard the one corresponding to $s = 3$. For the Hiseq 4000 data, we have $\hat{\pi}_r = (0.075, 0.136, 0.004, 0.005, 0.169, 0.517, 0.082, 0.011)$ for $r = 3$ for $r = 4$. Given that the lowest proportions are for $s = 3, 4, 8$ in that order, and highest for $s = 5$, it does make sense indeed to expect that the two reads $y^{(opt)}$ originated from $s = 5$. In contrast, for $y^{(below)}$, it is not really clear that the hopped reads originated from $s = 3$ given the sample has the lowest proportion of molecules. For the Hiseq 2500 data, we have $\hat{\pi}_r = (0.096, 0.173, 0.001, 0.001, 0.216, 0.409, 0.093, 0.011)$ for $r = 3$, which explains why we end up discarding the molecule $s = 8$ in $y^{(below)}$ but not in $y^{(opt)}$ since there are half the number of molecules in $s = 5$ than in $s = 6$, making it more likely that the read indeed hopped from $s = 8$. We have a similar situation for the Novaseq datasets where the sample with the lowest proportion of molecules (i.e. 0.005) corresponds to $s = 15$ (i.e. Sample $P7_8$), making it more likely that the two hopped reads in $y^{(opt)}$ of L1, the one hopped read in in $y^{(below)}$ of L1, and the one hopped read in $y^{(opt)}$ of L2 actually originated from $s = 9$. This also makes $y^{(below)}$ of L2 more likely to be a fugue non-chimeric observation, such that the one read in $s = 15$ would in reality have had hopped from a sample with a larger number of molecules.

Indeed, the model we propose aims at capturing the distribution patterns in the count data in order to both estimate the index hopping rate at the level of individual reads and quantify its effect on the observed fraction of phantom molecules. Given that our model is probabilistic, we can instead discard molecules not based on their observed read fraction, but rather on a corresponding posterior probability that is a function of not only the distribution of the read counts, but also the model's estimated sample index hopping rate, the library sizes and the expression profiles of the multiplexed samples - as given by molecular proportions complexity profile (i.e. the distribution of molecular proportions conditional on the *PCR duplication levels*). Furthermore, the probabilistic modeling approach we delineated in this paper provides an estimate of the fundamental quantity of interest -*the sample barcode*

index hopping rate- and connects it to its manifestation as a data contamination artifact as measured by *the fraction of phantom molecules*. As we have shown (see Equation 3), the fraction of phantom molecules is determined not only by the sample index hopping rate, but also by the number of multiplexed samples, as well as by their library size, coverage, and molecular complexity profiles. As such, this complex relationship, which we have attempted to capture in our proposed model, can potentially provide an explanation for the large variance in the index hopping rate estimates (0.2% - 8%) that have been reported in the literature (Larsson *et al.*, 2018; Griffiths *et al.*, 2018; Costello *et al.*, 2018; Vodák *et al.*, 2018).

4.2 Limitations and Future Directions

The work presented in this paper has the following limitations.

The non-negligible probability of a *cell-umi-gene* label collision across samples when the number of samples is large. As the number of multiplexed samples increases, Corollary 1 would no longer hold since the probability of observing, in more than one sample, a given *cell-umi-gene* label combination becomes non-negligible. That said, in single cell experiments, multiplexing more than 16 samples on a single lane is not commonly done given the smaller library size and lower genomic coverage that follows as a result. Furthermore, the adoption of a longer UMI index in the latest 10X Genomics Single Cell 3' v3 assay would further reduce the probability of potential collisions, thus rendering this concern less of a problem.

The sensitivity of the FPR-minimizing cutoff on the ECDF of q. As we have already mentioned, the marginal distribution of q given by Equation 7 has no closed-form nor is it feasible to compute the exact probabilities for all its possible outcomes even when r is not large. As r increases, we tend to observe in the data an increasingly fewer proportion of all possible outcomes and subsequently we would expect the resulting ECDF to deviate from the theoretical distribution. That said, the deviations would only affect the classification measures and the determination of the cutoff o^* whether optimally or not, but they would not affect the read reassignment procedure and the initial purging that retains all predicted real molecules (i.e. when we set $o^* = 1$) since the marginal distribution is not involved at this step.

Memory requirements. For data generated on the Hiseq 4000, the analysis workflow does not require computational requirements beyond what is found on a regular desktop with 32G of RAM. However, for Novaseq 6000 data with a large number of multiplexed samples (e.g. $S > 8$), more memory would be needed - up to 150G if not more. In particular, the first step in the workflow consisting of joining data from all samples into a datatable keyed by cell-barcode, UMI, and gene ID is the most memory intensive.

Incorporating the sample barcode indices in the model. In this work, we formulated a model for index hopping that treats all the four distinct sample barcodes present in a given sample as identical. Even though we showed in Subsection 2.2.2 that the sample barcode index hopping rate can be derived from the model nonetheless, we have not used information from the sample barcode indices to reassign reads or to purge phantom molecules. The model can potentially be expanded to incorporate the 4 sample barcodes, but such an extension would be accomplished at the cost of increased, if not prohibitive, computational and memory requirements since the workflow would need to start with the FASTQ files, not the output of the CellRanger pipeline and furthermore, the read count data table would need to be joined and keyed across 4 times the number of samples, potentially requiring more memory than would be available on some clusters. We would like to point out that the main reason we chose to formulate a model for index hopping at the level of samples and not individual sample barcodes is that in the end, sample index hopping between reads belonging to the same sample would not have any effect on downstream analyses since no phantom molecules would be generated when a read in a given sample swaps its sample barcode index with one of the other three barcode indices that are assigned to the same sample.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR) grant PJT-155966, Brain Canada Foundation, and the Alfred P. Sloan Foundation (grant FG-2018-10450) to H.S.N, as well as resource allocations from Compute Canada. H.S.N holds a CIHR Canada Research Chair.

Acknowledgements

The authors thank Alfredo Staffa, Haig Djambazian, and Ioannis Ragoussis for their comments and useful discussions.

References

Bach, K. *et al.* (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications*.

Consortium, T. T. M. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**(7727), 367–372.

Costello, M. *et al.* (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*.

Griffiths, J. A. *et al.* (2018). Detection and removal of barcode swapping in single-cell RNA-seq data. *Nature Communications*.

Hadfield, J. (2016). Index mis-assignment between samples on hiseq 4000 and x-ten. Accessed: 2018-11-30.

Illumina, I. (2017). Effects of index misassignment on multiplexing and downstream analysis. *URL: www.illumina.com*.

Larsson, A. J. *et al.* (2018). Computational correction of index switching in multiplexed sequencing libraries.

Lun, A. T. L. *et al.* (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, **20**(1), 63.

Sinha, R. *et al.* (2017). Index switching causes "spreading-of-signal" among multiplexed samples in illumina hiseq 4000 dna sequencing. *bioRxiv*.

Vodák, D. *et al.* (2018). Sample-Index Misassignment Impacts Tumour Exome Sequencing. *Scientific Reports*.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*.

Zerbino, D. R. *et al.* (2018). Ensembl 2018. *Nucleic Acids Research*.

Supplementary Material

Table 9: Effect of contamination on cell calling in Novaseq 6000 Lane 1 dataset.

Novaseq 6000 Lane 1 Dataset						
	Phantom		**Consensus**		**Reclassified as**	
sample	cell	empty	cell	empty	cell	empty
P7_0	0	61772	1108	245916	175	194
P7_1	0	73677	549	231823	104	223
P7_10	0	41615	2450	375417	18	10
P7_11	0	64228	3474	259464	1	33
P7_12	0	59249	3027	291809	17	28
P7_13	0	53451	3381	305484	13	26
P7_14	0	64676	2600	244136	13	28
P7_15	0	54118	2754	242667	10	8
P7_2	0	54021	4037	306265	12	21
P7_3	0	49839	3973	330019	26	26
P7_4	0	63917	927	250569	29	69
P7_5	0	47395	1969	316510	39	12
P7_6	0	45809	7267	304095	63	17
P7_7	0	63518	732	210350	21	11
P7_8	54	158555	1050	111944	27	95
P7_9	0	76342	2309	238230	5	91

Table 10: Effect of contamination on cell calling in Novaseq 6000 Lane 2 dataset

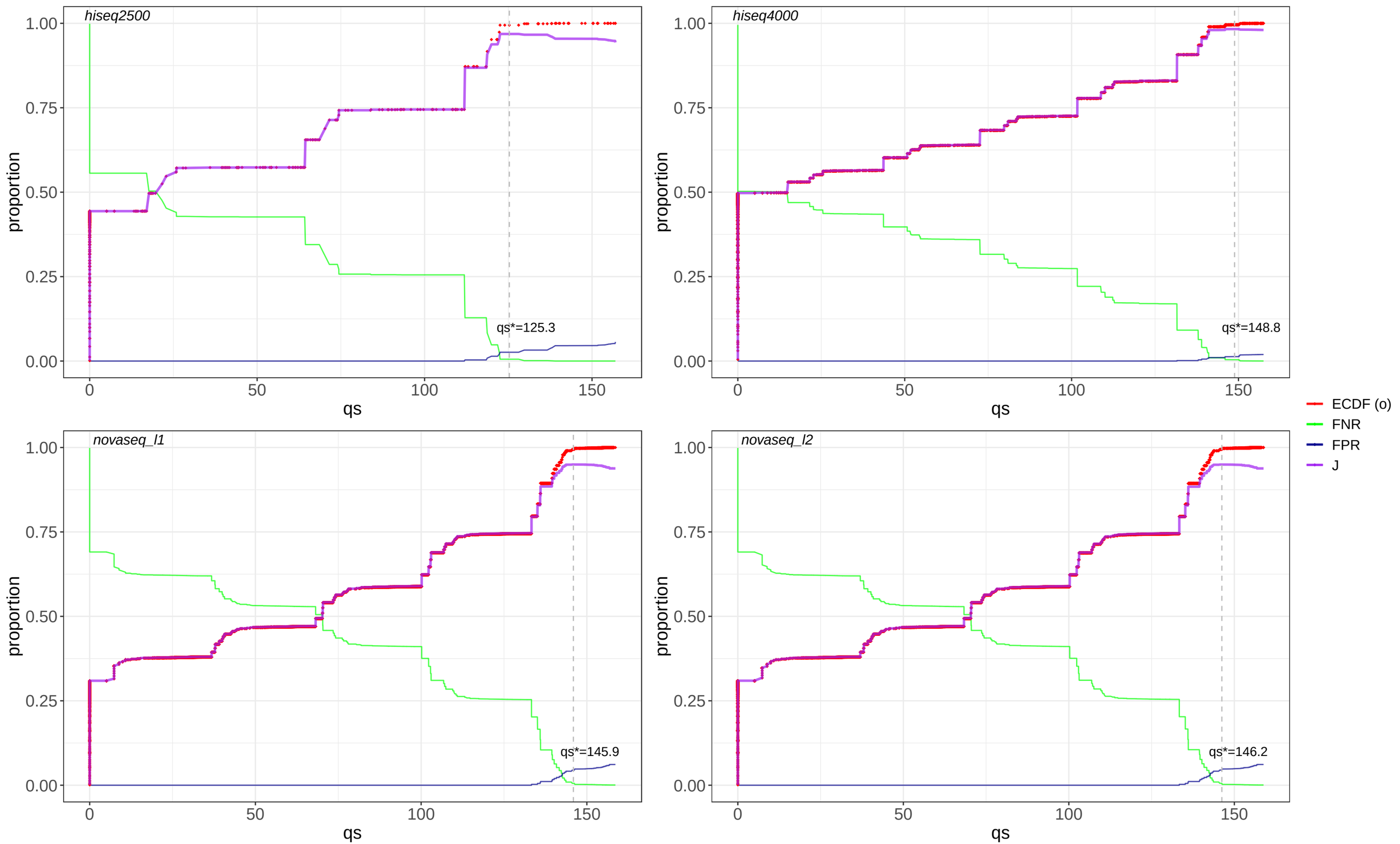
Novaseq 6000 Lane 2 Dataset						
	Phantom		**Consensus**		**Reclassified as**	
sample	cell	empty	cell	empty	cell	empty
P7_0	0	62321	944	246446	295	215
P7_1	0	73684	611	232297	48	225
P7_10	0	42025	2451	376555	21	12
P7_11	0	64495	3476	261000	0	38
P7_12	0	59620	3029	293206	15	33
P7_13	0	54120	3389	306709	7	27
P7_14	0	65386	2597	245290	15	33
P7_15	0	54121	2757	244201	6	8
P7_2	0	53777	4036	307338	11	22
P7_3	0	50220	3981	330875	23	26
P7_4	0	64176	826	251665	17	173
P7_5	0	47132	2021	317365	43	17
P7_6	0	46228	7266	305391	73	16
P7_7	0	63731	732	211400	25	10
P7_8	50	160092	1049	111937	28	109
P7_9	0	76811	2305	239238	13	95

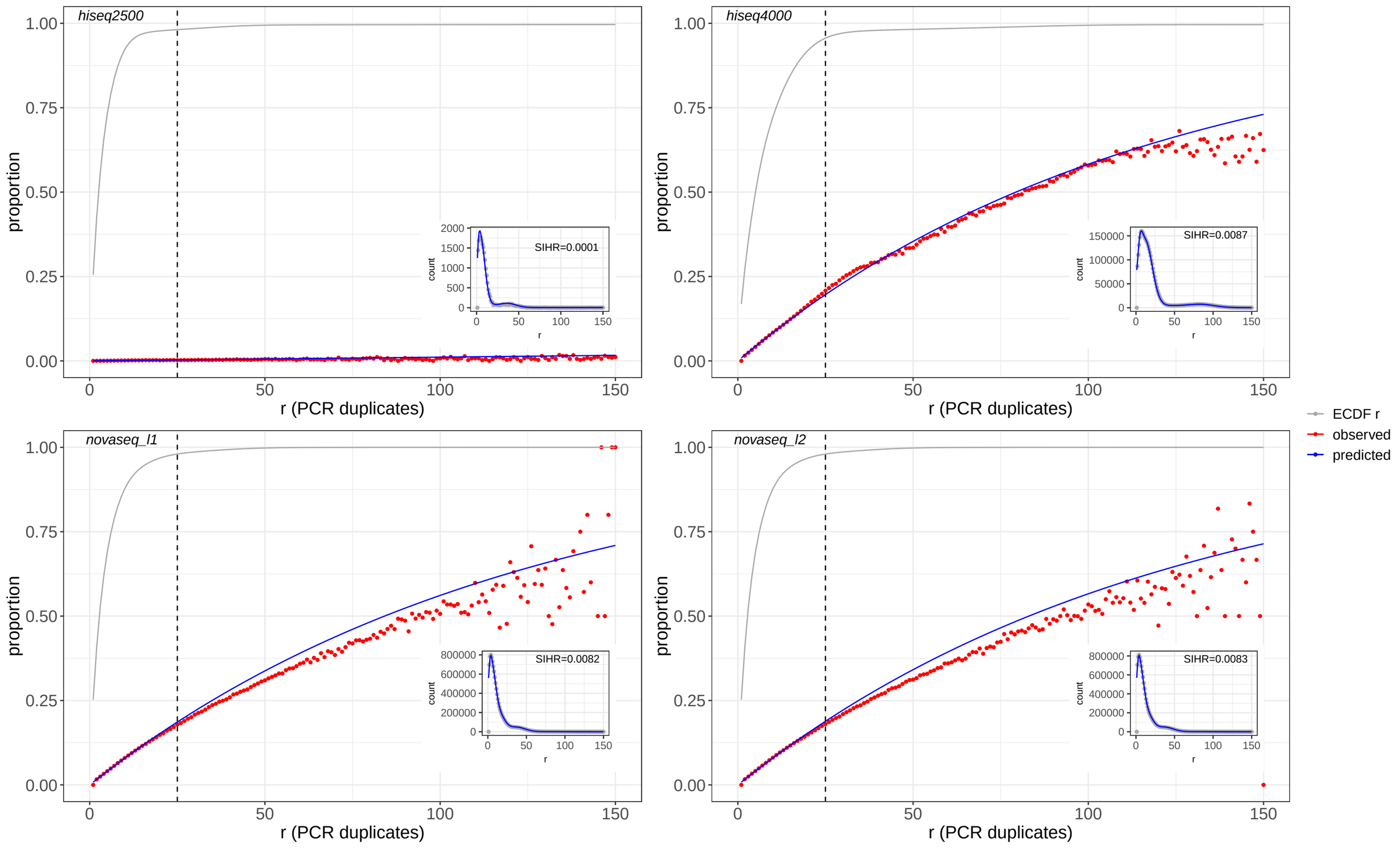
Table 11: Extent of contamination by sample for Novaseq 6000 Lane 1 dataset (the first lane of Tabula Muris experimental run). The number of molecules is given in multiples of million. The proportion of total molecules predicted as real and therefore retained is given by the *real* column. The proportion of total molecules that predicted as real but discarded since their q values fall below the optimal threshold is given by the *false phantoms* column. The proportion of molecules predicted as phantom is given by the *phantoms* column. The two proportions columns give show the discrepancy between library size and library complexity across the samples. The last column measures the extent of PCR Amplification bias as given by *FRM* (i.e. the Fraction of total number of Reads over total number of Molecules (i.e. library size divided by library complexity).

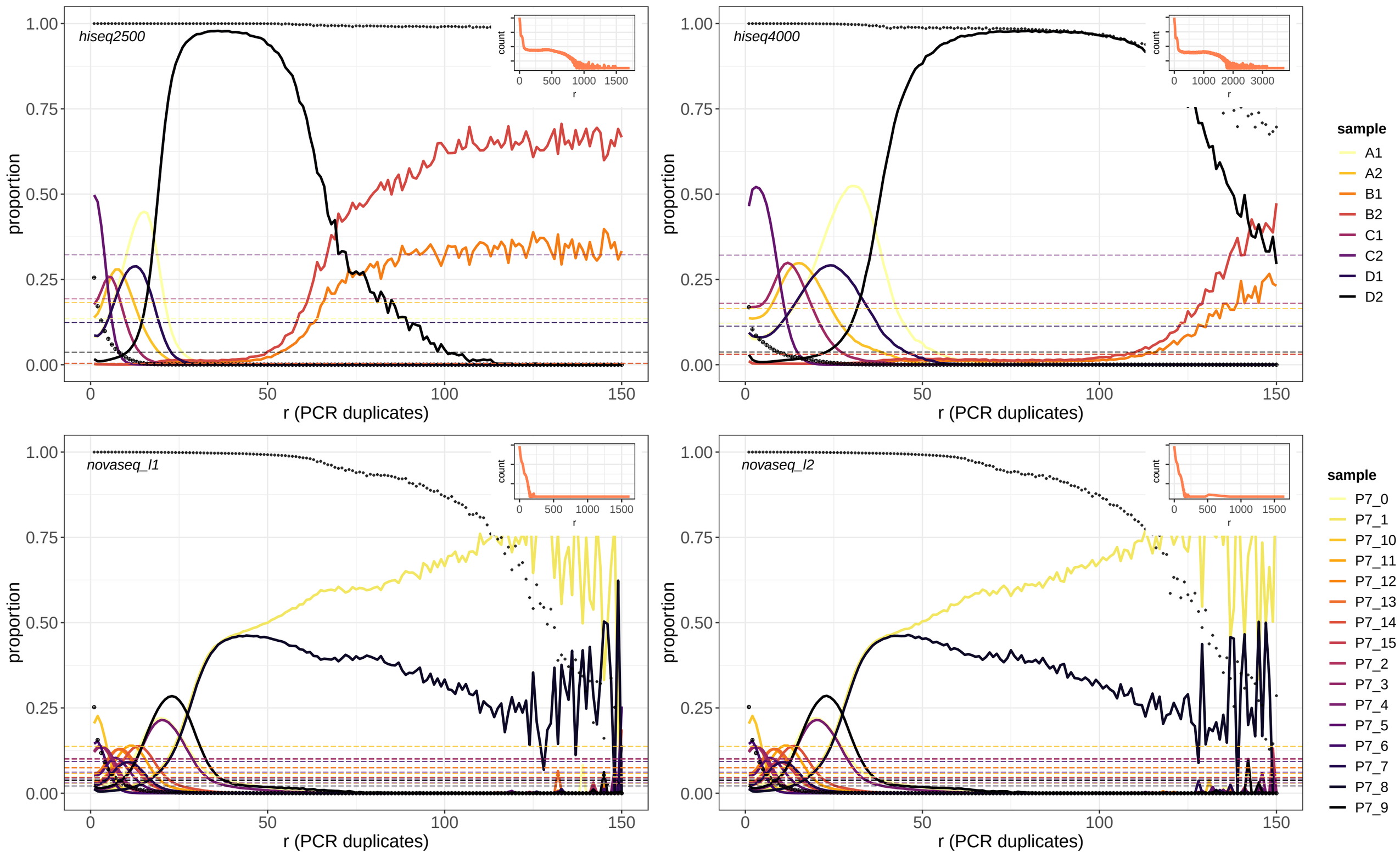
Novaseq 6000 Lane 1 Dataset							
		Breakdown			**Proportions**		
sample	molecules	real	false phantoms	phantoms	molecules	reads	FRM
P7_0	13.54	0.9316	0.0012	0.0672	0.048	0.074	7.9
P7_1	8.16	0.8653	0.0004	0.1343	0.029	0.081	14.5
P7_10	38.68	0.9827	0.0040	0.0132	0.138	0.064	2.4
P7_11	16.20	0.9549	0.0014	0.0437	0.058	0.068	6.1
P7_12	21.28	0.9663	0.0022	0.0315	0.076	0.070	4.8
P7_13	20.80	0.9681	0.0022	0.0298	0.074	0.071	4.9
P7_14	12.92	0.9470	0.0007	0.0523	0.046	0.061	6.8
P7_15	12.81	0.9600	0.0014	0.0386	0.046	0.048	5.4
P7_2	28.02	0.9748	0.0038	0.0214	0.100	0.057	2.9
P7_3	28.58	0.9754	0.0031	0.0214	0.102	0.061	3.1
P7_4	8.48	0.9101	0.0001	0.0898	0.030	0.059	10.1
P7_5	17.40	0.9678	0.0020	0.0302	0.062	0.054	4.5
P7_6	26.17	0.9770	0.0049	0.0181	0.093	0.042	2.3
P7_7	11.69	0.9512	0.0010	0.0478	0.042	0.046	5.7
P7_8	5.99	0.6818	0.1429	0.1753	0.021	0.072	17.3
P7_9	10.49	0.9146	0.0005	0.0849	0.037	0.073	10.1

Table 12: Extent of contamination by sample for Novaseq 6000 Lane 2 dataset (the second lane of Tabula Muris experimental run). The number of molecules is given in multiples of million. The proportion of total molecules predicted as real and therefore retained is given by the *real* column. The proportion of total molecules that predicted as real but discarded since their q values fall below the optimal threshold is given by the *false phantoms* column. The proportion of molecules predicted as phantom is given by the *phantoms* column. The two proportions columns give show the discrepancy between library size and library complexity across the samples. The last column measures the extent of PCR Amplification bias as given by *FRM* (i.e. the Fraction of total number of Reads over total number of Molecules (i.e. library size divided by library complexity).

Novaseq 6000 Lane 2 Dataset							
		Breakdown			**Proportions**		
sample	molecules	real	false phantoms	phantoms	molecules	reads	FRM
P7_0	13.59	0.9309	0.0012	0.0679	0.048	0.074	7.9
P7_1	8.20	0.8645	0.0004	0.1351	0.029	0.081	14.4
P7_10	38.78	0.9825	0.0041	0.0135	0.137	0.064	2.4
P7_11	16.25	0.9543	0.0014	0.0443	0.058	0.068	6.1
P7_12	21.34	0.9657	0.0022	0.0320	0.076	0.070	4.8
P7_13	20.85	0.9676	0.0022	0.0303	0.074	0.071	4.9
P7_14	12.96	0.9463	0.0007	0.0530	0.046	0.061	6.8
P7_15	12.84	0.9595	0.0014	0.0392	0.046	0.048	5.4
P7_2	28.11	0.9748	0.0035	0.0218	0.100	0.057	2.9
P7_3	28.67	0.9751	0.0032	0.0217	0.102	0.061	3.1
P7_4	8.51	0.9088	0.0001	0.0910	0.030	0.059	10.1
P7_5	17.43	0.9674	0.0020	0.0306	0.062	0.054	4.5
P7_6	26.25	0.9767	0.0049	0.0184	0.093	0.042	2.3
P7_7	11.72	0.9506	0.0010	0.0485	0.042	0.046	5.7
P7_8	6.03	0.6779	0.1459	0.1762	0.021	0.072	17.3
P7_9	10.53	0.9139	0.0005	0.0856	0.037	0.073	10.1


hiseq2500
hiseq4000
novaseq_l1
novaseq_l2
qs*=125.3
qs*=148.8
qs*=145.9
qs*=146.2
proportion
qs
0.00
0.25
0.50
0.75
1.00
0
50
100
150
ECDF (o)
FNR
FPR
J


hiseq2500
hiseq4000
novaseq_l1
novaseq_l2
proportion
r (PCR duplicates)
count
r
SIHR=0.0001
SIHR=0.0087
SIHR=0.0082
SIHR=0.0083
ECDF r
observed
predicted


hiseq2500
hiseq4000
novaseq_l1
novaseq_l2
proportion
r (PCR duplicates)
count
r
sample
A1
A2
B1
B2
C1
C2
D1
D2
P7_0
P7_1
P7_10
P7_11
P7_12
P7_13
P7_14
P7_15
P7_2
P7_3
P7_4
P7_5
P7_6
P7_7
P7_8
P7_9