# Effects of Index Misassignment on Multiplexing and Downstream Analysis

Learn why it happens and best practices to reduce the impact of index hopping.

## Introduction

Improvements in next-generation sequencing (NGS) technology have greatly increased sequencing speed and data output, resulting in the massive sample throughput of current sequencing platforms. Ten years ago, the Genome Analyzer was capable of generating up to 1 Gb of sequence data per run. Today, the NovaSeq™ 6000 System, built on the same core technology, is capable of generating up to 2 Tb of data in two days, which represents a > 2000× increase in capacity.[1]

A key to utilizing this increased capacity is multiplexing, which adds unique sequences, called indexes, to each DNA fragment during library preparation. This allows large numbers of libraries to be pooled and sequenced simultaneously during a single sequencing run. Gains in throughput from multiplexing come with an added layer of complexity, as sequencing reads from pooled libraries need to be identified and sorted computationally in a process called demultiplexing before final data analysis (Figure 1).

Index misassignment between multiplexed libraries is a known issue that has impacted NGS technologies from the time sample multiplexing was developed.[2] This white paper describes the mechanisms by which index hopping may occur, how Illumina measures index hopping, and best practices for mitigating the impact of index hopping on sequencing data quality.

## Mechanisms of index misassignment

### Molecular recombination of indexes, ie, "index hopping"

The development of exclusion amplification (ExAmp) chemistry and patterned flow cell technology was a significant advance in NGS technology that resulted in increased data output, reduced costs, and faster run times. This has enabled a broad range of applications including the $1000 Genome.[3] However, this clustering method used with patterned flow cells has been observed to result

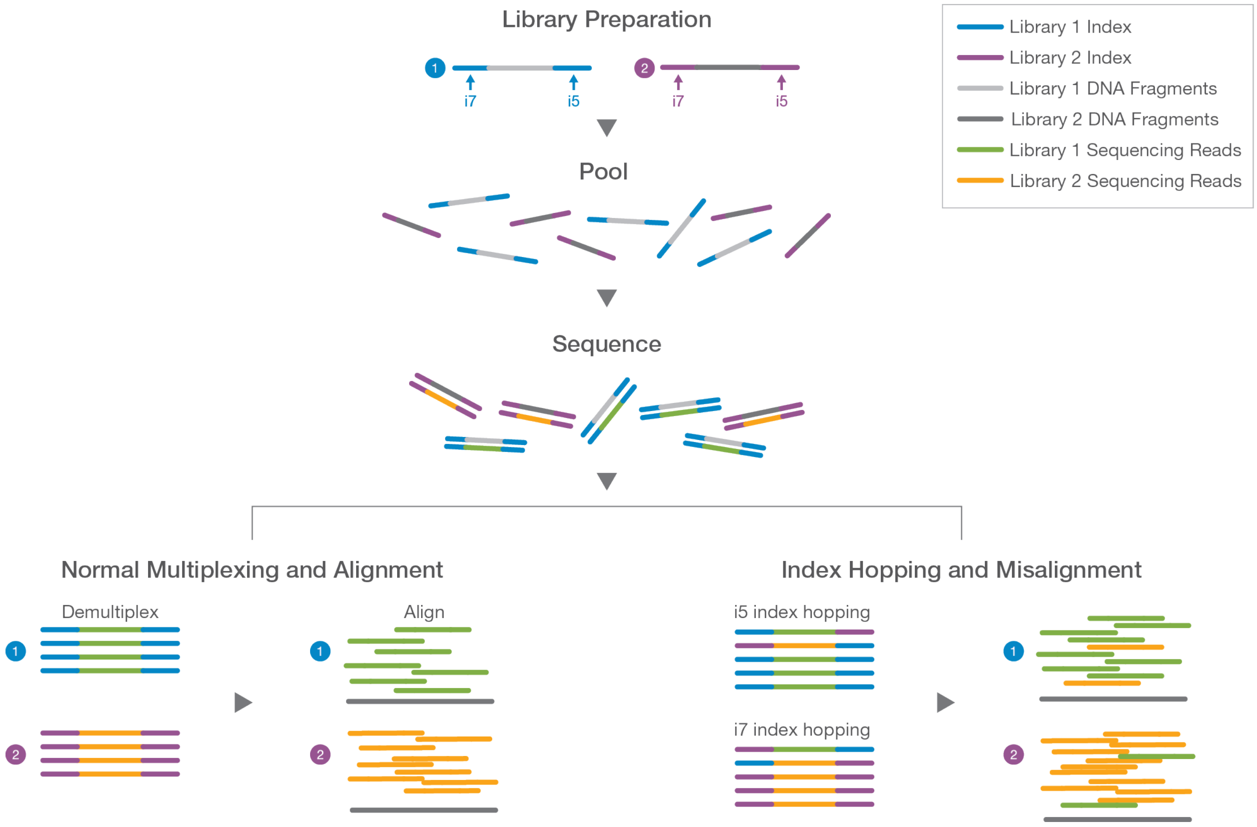

Figure 1: Overview of multiplexing and index hopping—Multiplexing enables pooling and sequencing of multiple libraries simultaneously during a single sequencing run through addition of unique index sequences to each DNA fragment during library preparation. Sequencing reads are sorted to their respective samples during demultiplexing, allowing for proper alignment. Index hopping causes incorrect assignment of sequencing reads and may lead to misalignment of reads or incorrect assumptions in downstream analysis.

in higher levels of index misassignment than traditional bridge amplification.[4] Index hopping is a specific cause of index misassignment that can result in incorrect assignment of libraries from the expected index to a different index in the pool, leading to misalignment and inaccurate sequencing results (Figure 1). Index hopping is the primary mechanism responsible for the observed increase in index misassignment in patterned flow cells.

## Contamination from free adapters/primers

After adapters are ligated to nucleic acid fragments, the products are cleaned up to remove any free, unligated adapters. Libraries can be cleaned up by a bead-based or gel purification step to remove free adapters or primers. Failure to remove free adapters or primers can lead to contamination of prepared libraries and may result in index hopping and misassignment. To demonstrate this possibility, adapters not present in a prepared library pool were spiked in at varying levels from 0–35% molar concentration relative to DNA input. Levels of index hopping increased in a linear fashion in correlation with increasing levels of adapter spike-in (Figure 2). These results highlight the importance of making sure that prepared libraries are clean before proceeding with a sequencing run.

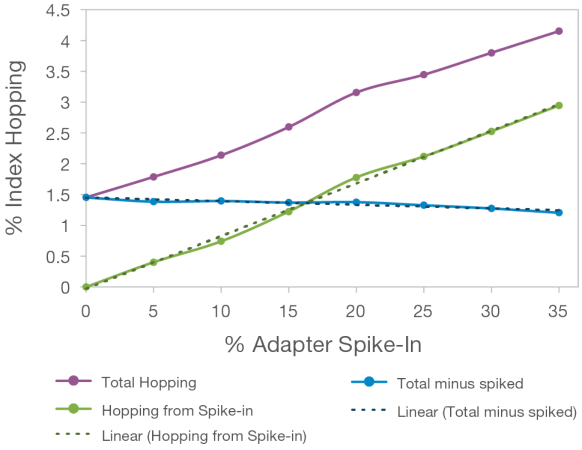

**Figure 2: Index hopping from free adapters**—Percent index hopping is plotted against levels of adapter spike-in. There is a positive, linear correlation between both total index hopping (purple line) and index hopping from spike-in (green line) and levels of added free adapter.

## Measuring index hopping

Library pooling experiments enable quantification of the level of index hopping. By using unique pairs of i5 and i7 index adapters, uniquely dual-indexed libraries are pooled, sequenced, and demultiplexed following a dual-indexed workflow. The percent index representation across all possible adapter combinations measures the level of index hopping at invalid combinations (Figure 3). For example, a value of 0.17% would correlate to ~1 index-hopping event per 600 correctly indexed pairs.

| | | i7 indexes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 701 | 702 | 703 | 704 | 705 | 706 | 707 | 708 |
| i5 indexes | 501 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| | 502 | 0.00% | 26.20% | 0.00% | 0.11% | 0.14% | 0.14% | 0.00% | 0.00% |
| | 503 | 0.00% | 0.17% | 0.00% | 0.10% | 23.41% | 0.12% | 0.00% | 0.00% |
| | 504 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| | 505 | 0.00% | 0.15% | 0.00% | 22.91% | 0.12% | 0.16% | 0.00% | 0.00% |
| | 506 | 0.00% | 0.14% | 0.00% | 0.12% | 0.12% | 23.37% | 0.00% | 0.00% |
| | 507 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| | 508 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

**Figure 3: Contamination matrix for unique indexes**—The percent index representation across all possible adapter combinations measures the level of index hopping. Overlap at valid and invalid combinations are shaded in green and red, respectively. Invalid index combinations are not preferentially impacted by index hopping.

## Impact of index hopping

The method of library preparation has been shown to contribute to levels of index hopping. In general, methods that only include ligation, such as the TruSeq™ DNA PCR-Free Library Prep Kit, generate libraries with higher levels of index hopping than methods that incorporate a subsequent PCR amplification step, such as the TruSeq Nano DNA Library Prep Kit (Figure 4). Libraries clustered on nonpatterned flow cells with traditional bridge amplification typically have lower rates of index hopping (≤ 1%) compared to libraries run on patterned flow cells using ExAmp cluster generation (≤ 2%). For example, analysis of a TruSeq PCR-Free library after cluster generation and sequencing shows lower levels of index hopping on a nonpatterned flow cell compared to a patterned flow cell (Figure 4).

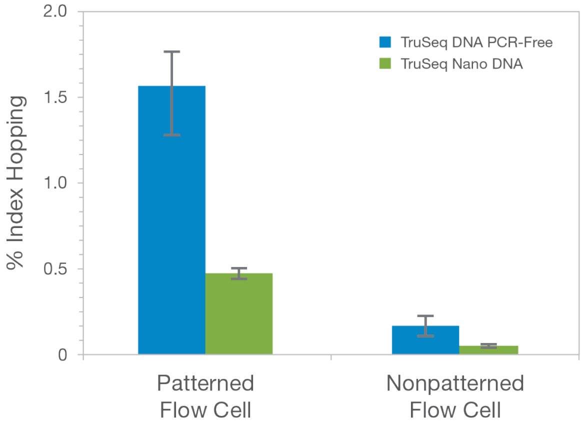

**Figure 4: Differences in rates of index hopping**—Levels of index hopping are higher with patterned versus nonpatterned flow cells, regardless of library prep method. Library prep methods with a PCR amplification step (eg, TruSeq Nano) show lower levels of index hopping compared to methods that include ligation only (eg, TruSeq DNA PCR-Free).

### Effect of index hopping on RNA sequencing experiments

To demonstrate the impact of typical levels of index hopping on RNA sequencing (RNA-Seq) of samples with very highly expressed markers, stranded mRNA libraries were prepared from total RNA samples from two different human tissues. The tissues were chosen such that one was highly enriched for expression of tissue-specific markers (liver), and the other had a more distributed expression profile not dominated by specific transcripts (brain).

Libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit following standard protocol. Samples were indexed with a unique index set, so that index hopping could be independently determined. Samples were sequenced either as pooled mixes of liver and brain or separate tissue pools, ie, liver pooled with liver or brain pooled with brain, in lanes as a 6 plex on the HiSeq™ 4000 System.

Sequencing data was demultiplexed and analyzed in BaseSpace™ Sequence Hub using the RNA Express App and the standard analysis pipeline. The percent index hopping was measured at 0.3–0.5% for the lanes analyzed. Fragments per kilobase million (FPKM) gene expression plots show detection of very highly expressed liver marker genes such as albumin (120,000–950,000 counts in liver) in the mixed tissue lane reads that are absent in the separately sequenced pooled brain reads as a consequence of index hopping (Figure 5, top panel). These liver markers observed in the pooled brain sample are at ~ 0.13% of the level observed in the liver sample. FPKM gene expression plots of replicates of the brain libraries sequenced in the presence of the liver tissue demonstrate the background noise is equivalent in replicates and not pulled out as differentially expressed (Figure 5, bottom panel). These results indicate that, to minimize the effect of index hopping, best practice is to pool similar samples together, so that dominant, very highly expressed transcripts will not lead to increased levels of index hopping. Storage of prepared libraries outside of recommended conditions (Table 1) has been demonstrated to increase rates of index hopping. Store individual libraries at –20° C; avoid storage at 4° C. Once pooled, sequence libraries as soon as possible or store at –20° C to mitigate index hopping.

## Best practices to reduce index hopping

In order to mitigate the effects of index hopping, specific recommendations dependent on the sequencing system, the library preparation workflow, and the application have been identified. These general guidelines and recommendations for reducing the impact of index hopping are provided (Table 1). Storage of prepared libraries outside of these recommended conditions has been demonstrated to increase rates of index hopping. Store individual libraries at –20° C; avoid storage at 4° C. Once pooled, sequence libraries as soon as possible or store at –20° C for up to one week to mitigate index hopping.

### Commercial solutions to reduce index hopping

Illumina offers both unique dual indexes and an enzymatic solution to minimize the effect of index hopping. The unique dual indexes eliminate hopped reads from downstream analysis, as unexpected combinations are assigned as undetermined and removed from the data. Ninety-six unique dual indexes are available for both TruSeq DNA and RNA workflows.

Free adapters in a library contribute to increased levels of index hopping by hybridizing and acting as a primer to produce index hopped strands (Figure 2). In addition to unique dual indexes, an enzymatic solution is available that reduces the level of free adapters in libraries. The Free Adapter Blocking Reagent blocks the 3′ end of the free adapters, and prevents extension. This post-library prep treatment reduces the rate of index hopping.

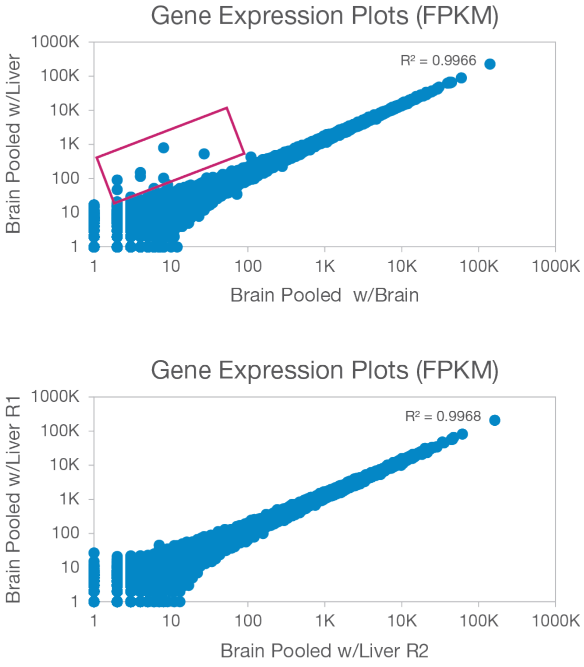

Figure 5: Impact of index hopping on RNA-Seq analysis—FPKM expression plots compare replicate samples of total RNA libraries of liver and brain tissue when sequenced separately or in a 6-plex pool on the HiSeq 4000 System. Detection of very highly expressed liver marker genes in pooled brain (red box) indicates occurrence of index hopping. The lower plot shows the negligible impact on replicate expression profiling of the mixed lane replicates. The coefficient of determination ($R^2$) for each plot is shown.

Table 1: Best practices for reducing index hopping

| Mitigation/Recommendation | Benefit/Outcome |
|---|---|
| Prepare dual indexed libraries with unique indexes[a] | Converts index hopped reads to undetermined |
| Sequence one 30× human genome per lane[b] | Avoids pooling and index hopping |
| Remove adapters (cleanup, spin columns, etc)[c] | Reduces levels of index hopping |
| Store prepared libraries at recommended temperature of –20° C[c] | Reduces levels of index hopping |
| Pool similar RNA-Seq samples together | Reduces contamination between high and low-expressors |

a. Available on all HiSeq Systems including the HiSeq X series of systems.
b. Only available on the HiSeq X series of sequencing systems.
c. See TruSeq Sample Preparation Best Practices and Troubleshooting Guide.

## Summary

Multiplexing represents both a major advance and a necessity in NGS technology, which enables significant increases in sample throughput. However, with multiplexing, the potential for index hopping is present regardless of the library prep method or sequencing system used. Index hopping may result in assignment of sequencing reads to the wrong index during demultiplexing, leading to misalignment and a potential negative impact on data quality. Evaluation of index hopping has shown that, for most applications, the impact on downstream analysis will be minimal. With the release of the IDT for Illumina Unique Dual Indexes and the Illumina Free Adapter Blocking Reagent, researchers have the tools to reduce the level of index hopping in their experiments and the ability to exclude residual index hopped reads from their downstream analyses.

## Ordering information

| Product | Catalog No. |
|---|---|
| IDT for Illumina-TruSeq DNA UD Indexes (24 indexes, 96 samples) | 20020590 |
| IDT for Illumina-TruSeq RNA UD Indexes (24 indexes, 96 samples) | 20020591 |
| IDT for Illumina-TruSeq DNA UD Indexes (96 indexes, 96 samples) | 20022370 |
| IDT for Illumina-TruSeq RNA UD Indexes (96 indexes, 96 samples) | 20022371 |
| Illumina Free Adapter Blocking Reagent (12 reactions) | 20024144 |
| Illumina Free Adapter Blocking Reagent (48 reactions) | 20024145 |

## References

1. Illumina. An Introduction to Next-Generation Sequencing Technology. 2016. Accessed April 2017.

2. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 2012:2513–2524.

3. Illumina. HiSeq X Series of Sequencing Systems. 2016. Accessed April 2017.

4. Illumina. Illumina Sequencing Technology. 2010. Accessed April 2017.

Pub. No. 770-2017-004-D QB 5746