illumina®

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. 10 years ago, the Genome Analyzer was capable of generating up to 1 Gb of sequence data per run. Today, the NovaSeq™ Series of Systems, 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

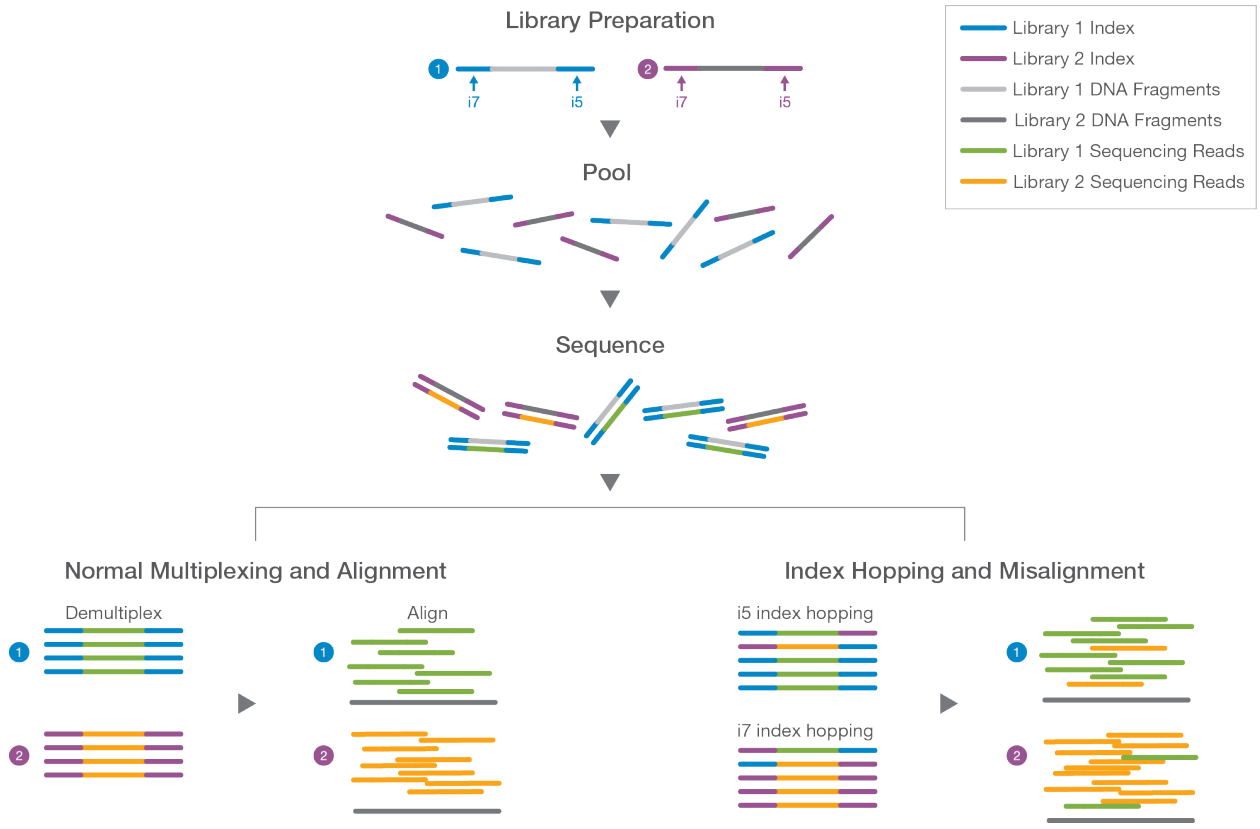

Figure 1: Overview of Multiplexing and Index Hopping— Multiplexing enables pooling and sequencing of multiple libraries simultaneously during a single sequencing run through addition of unique index sequences to each DNA fragment during library preparation. Sequencing reads are sorted to their respective 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.

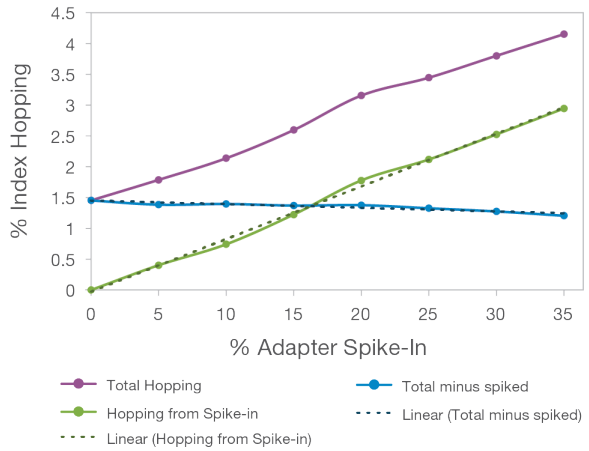

Figure 2: Index Hopping from Free Adapters — Percent index hopping is plotted against levels of adapter spike-in. There is a positive, linear correlation between both total index hopping (red line) and index hopping from spike-in (yellow line) and levels of added free adapter.

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).

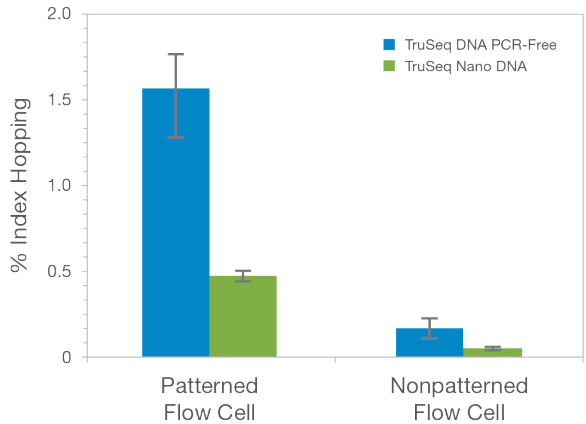

Figure 4: Differences in Rates of Index Hopping — Levels of index hopping are higher with patterned versus nonpatterned flow cells, regardless of library prep method. Library prep methods with a PCR amplification step (eg, TruSeq Nano) show lower levels of index hopping compared to methods that include 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 the 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.

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 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.

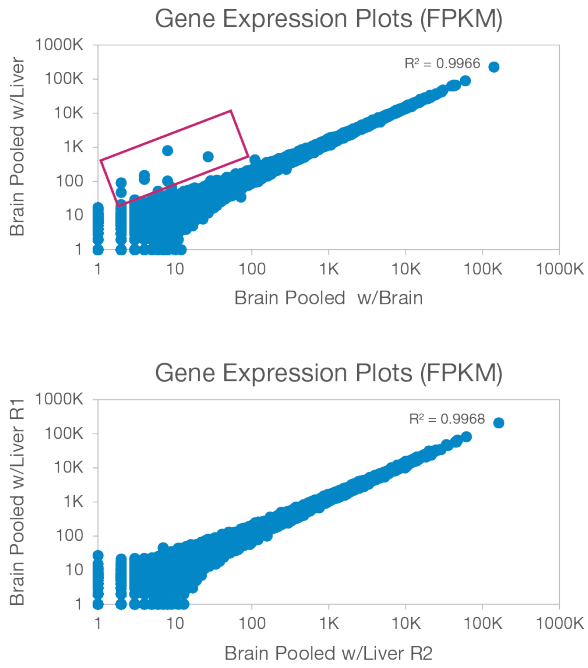

Figure 5: Impact of Index Hopping on RNA-Seq Analysis — FPKM expression plots compare replicate samples of total RNA libraries of liver and brain tissue when sequenced separately or in a 6-plex pool on the HiSeq 4000 System. Detection of very highly expressed liver marker genes in pooled brain (red box)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

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. Not supported on the HiSeq X series of sequencing systems.
b. Only available on the HiSeq X series of sequencing systems.
c. See TruSeq Sample Preparation Best Practices and Troubleshooting Guide.

Pooling Guidelines for Dual-Indexed Sequencing

TruSeq High-Throughput (HT) Library Prep Kits contain either a DNA adapter plate (DAP) or RNA adapter plate (RAP) depending on the kit. An adapter plate is a 96-well plate containing 96 uniquely indexed adapter combinations, designed for manual or automated preparation of up to 96 uniquely indexed libraries. Illumina has determined optimal pooling guidelines for multiplexing 12 8–plex combinations (Table 2) or 16 6–plex combinations (Table 3) that maximize use of the adapter plate and can be used to mitigate or identify index hopping. These unique index combinations result in

Table 2: Pooling Guidelines for 8-Plex Combinations

1		2		3		4		5		6	
Adapter Pair	Coordinate	Adapter Pair	Coordinate	Adapter Pair	Coordinate	Adapter Pair	Coordinate	Adapter Pair	Coordinate	Adapter Pair	Coordinate
D501–D705	A5	D502–D706	B6	D503–D701	C1	D505–D702	E2	D506–D704	F4	D507–D703	G3
D502–D704	B4	D501–D702	A2	D505–D703	E3	D503–D706	C6	D507–D705	G5	D506–D701	F1
D503–D703	C3	D505–D705	E5	D506–D706	F6	D507–D701	G1	D504–D702	D2	D508–D704	H4
D505–D701	E1	D503–D704	C4	D507–D702	G2	D506–D703	F3	D508–D706	H6	D504–D705	D5
D506–D710	F10	D507–D712	G12	D504–D707	D7	D508–D708	H8	D501–D709	A9	D502–D711	B11
D507–D709	G9	D506–D708	F8	D508–D711	H11	D504–D712	D12	D502–D710	B10	D501–D707	A7
D504–D711	D11	D508–D710	H10	D501–D712	A12	D502–D707	B7	D503–D708	C8	D505–D709	E9
D508–D707	H7	D504–D709	D9	D502–D708	B8	D501–D711	A11	D505–D712	E12	D503–D710	C10
7		**8**		**9**		**10**		**11**		**12**	
Adapter Pair	Coordinate	Adapter Pair	Coordinate	Adapter Pair	Coordinate	Adapter Pair	Coordinate	Adapter Pair	Coordinate	Adapter Pair	Coordinate
D501–D710	A10	D502–D712	B12	D503–D707	C7	D505–D708	E8	D506–D709	F9	D507–D711	G11
D502–D709	B9	D501–D708	A8	D505–D711	E11	D503–D712	C12	D507–D710	G10	D506–D707	F7
D503–D711	C11	D505–D710	E10	D506–D712	F12	D507–D707	G7	D504–D708	D8	D508–D709	H9
D505–D707	E7	D503–D709	C9	D507–D708	G8	D506–D711	F11	D508–D712	H12	D504–D710	D10
D506–D705	F5	D507–D706	G6	D504–D701	D1	D508–D702	H2	D501–D704	A4	D502–D703	B3
D507–D704	G4	D506–D702	F2	D508–D703	H3	D504–D706	D6	D502–D705	B5	D501–D701	A1
D504–D703	D3	D508–D705	H5	D501–D706	A6	D502–D701	B1	D503–D702	C2	D505–D704	E4
D508–D701	H1	D504–D704	D4	D502–D702	B2	D501–D703	A3	D505–D706	E6	D503–D705	C5

Table 3: Pooling Guidelines for 6-Plex Combinations

1		2		3		4		5		6		7		8	
Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well
D501–D705	A5	D501–D710	A10	D502–D704	B4	D502–D709	B9	D503–D703	C3	D503–D711	C11	D505–D701	E1	D505–D707	E7
D502–D706	B6	D502–D712	B12	D501–D702	A2	D501–D708	A8	D505–D705	E5	D505–D710	E10	D503–D704	C4	D503–D709	C9
D503–D701	C1	D503–D707	C7	D505–D703	E3	D505–D711	E11	D506–D706	F6	D506–D712	F12	D507–D702	G2	D507–D708	G8
D505–D702	E2	D505–D708	E8	D503–D706	C6	D503–D712	C12	D507–D701	G1	D507–D707	G7	D506–D703	F3	D506–D711	F11
D506–D704	F4	D506–D709	F9	D507–D705	G5	D507–D710	G10	D504–D702	D2	D504–D708	D8	D508–D706	H6	D508–D712	H12
D507–D703	G3	D507–D711	G11	D506–D701	F1	D506–D707	F7	D508–D704	H4	D508–D709	H9	D504–D705	D5	D504–D710	D10
9		**10**		**11**		**12**		**13**		**14**		**15**		**16**	
Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well	Adapter Pair	Well
D506–D710	F10	D506–D705	F5	D507–D709	G9	D507–D704	G4	D504–D711	D11	D504–D703	D3	D508–D707	H7	D508–D701	H1
D507–D712	G12	D507–D706	G6	D506–D708	F8	D506–D702	F2	D508–D710	H10	D508–D705	H5	D504–D709	D9	D504–D704	D4
D504–D707	D7	D504–D701	D1	D508–D711	H11	D508–D703	H3	D501–D712	A12	D501–D706	A6	D502–D708	B8	D502–D702	B2
D508–D708	H8	D508–D702	H2	D504–D712	D12	D504–D706	D6	D502–D707	B7	D502–D701	B1	D501–D711	A11	D501–D703	A3
D501–D709	A9	D501–D704	A4	D502–D710	B10	D502–D705	B5	D503–D708	C8	D503–D702	C2	D505–D712	E12	D505–D706	E6
D502–D711	B11	D502–D703	B3	D501–D707	A7	D501–D701	A1	D505–D709	E9	D505–D704	E4	D503–D710	C10	D503–D705	C5

filtering of misassigned reads during secondary analysis. Misassigned reads will be flagged as "unaligned reads" and can be excluded from alignment. Using unique dual index combinations (Tables 2, 3) is a best practice to make sure that reads with incorrect indexes do not impact variant calling or assignment of gene expression counts.

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. While a permanent solution for index hopping is under development, this white paper provides guidelines and best practices to minimize index hopping.

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.

Illumina, Inc. • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.

illumina®