TECHNICAL NOTE

swiftbiosci.com

INCREASE SPECIFICITY OF DETECTING LOW FREQUENCY ALLELES WITH MOLECULAR IDENTIFIERS

Reliably detecting low frequency alleles from next-generation sequencing (NGS) has been a challenge, partly due to the lack of available methods to differentiate true variants from the background noise. This lack of specificity leads to high errors due to false positives and false negatives, and can ultimately lead to miscalculating the relative allele fraction of detected variants. Unique molecular identifiers can be used to distinguish between low frequency mutations and artifacts generated by either polymerase errors during PCR amplification or sequencing errors. Incorporating molecular identifiers into DNA libraries helps increase specificity to detect low frequency alleles. In addition, molecular identifiers provide more accurate deduplication; thereby significantly improving data retention and substantially greater genomic complexity, which is necessary for more comprehensive sample analysis.

Swift's molecular identifiers (MIDs) are proprietary technology developed exclusively for use with all Accel-NGS® 2S DNA Library Kits, and have been optimized with hybridization capture of enriched targets from tumor or cell-free DNA (cfDNA). This Technical Note describes modifications to the standard Accel-NGS 2S Kits to incorporate MIDs in DNA libraries for higher data retention and sensitivity in low frequency variant calling and offers important considerations.

Overview of Accel-NGS MID Technology

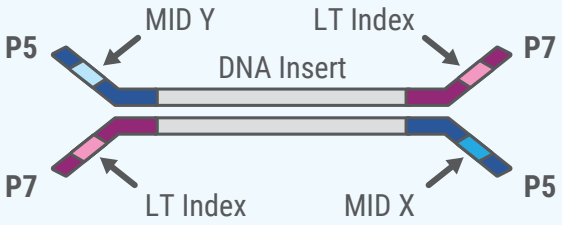


Figure 1: *Swift MID technology works in conjunction with Accel-NGS 2S chemistry. The P5 adapter sequence is TruSeq® HT with a 9-base random N sequence at the index 2 position. The P7 adapter sequence is TruSeq LT with a 6-base standard index at the index 1 position. Swift MIDs are strand-specific. Each double-stranded DNA (dsDNA) substrate receives two independent P5 MID adapters "X" and "Y". Due to the polishing reaction, both MID X and Y families will have identical termini.*

Accel-NGS 2S MID Validation

Library	Total Index 2 (I5) Reads	Unique MIDs	Mean MID Copy #	MID Copy Uniformity	Maximum Copy Number
PCR-free	5,498,922	262,104	20.98	98%	464
9 PCR Cycles	4,848,690	262,036	18.50	99%	447

Table 1: *Libraries including MID index sequences were prepared with Accel-NGS 2S Kits.*

- A 9-base random N sequence has 262,144 theoretical MID sequence combinations.
- Both PCR-free and amplified 2S libraries attained close to theoretical representation of MID combinations with high copy uniformity (% MIDs that are > 20% of the mean MID copy number) to minimize the likelihood of the same MID tagging > 1 DNA fragment at any target locus.
- The maximum copy number column indicates that no MID sequence was present at significantly greater than 20x mean copy number.

Implications of Strand Specificity Using Accel-NGS MID Technology

- Swift MID libraries enable strand-specific consensus sequence (SSCS) analysis — a novel approach in which the PCR duplicates of each MID family are bioinformatically grouped to generate an independent consensus sequence.[1]
- SSCS formation removes sequencer-derived errors, which account for > 99% of all artifacts.[1] This analysis also eliminates artifacts arising from PCR errors during library amplification.
- Because of the independent attachment of MIDs during ligation, Swift libraries cannot call a duplex consensus sequence via duplex sequencing to trace each strand of a dsDNA duplex back to one of the two strands of the original dsDNA molecule.
- It has been demonstrated that a high percentage of recovered reads provided by duplex sequencing (i.e. 80-90%) are single-stranded molecules whose sister-strand has been lost.[2] This phenomenon renders duplex sequencing largely intractable and costly.[3]

The Power of MIDs for Data Retention

There are several types of duplicate molecules that exist in sequencing data — PCR duplicates, fragmentation duplicates, and complementary sister-strand duplicates derived from input dsDNA (optical or clustering duplicates that are eliminated based on the Illumina® flow cell position were not evaluated). Standard tools, such as Picard Mark Duplicates, cannot distinguish between these three classes of duplicates, each of which have a common aligned map position, so all three types will be removed prior to data analysis. However, use of MIDs allows for accurate identification of PCR from other duplicates. Although each share a common map position, PCR duplicates also share a common MID while sister-strand and fragmentation duplicates have unique MIDs. Since they can be separately identified and retained, it is expected that deduplication using MIDs would show an increase in data retention, particularly from cell-free DNA where fragmentation is less random than mechanical shearing. As shown in page 4, Table 2, we observed higher data retention from cfDNA samples when comparing MID vs. Picard deduplication. This effect will be less pronounced at lower depth of sequencing as the gain (at lower depth of sequencing) will be proportionally less.

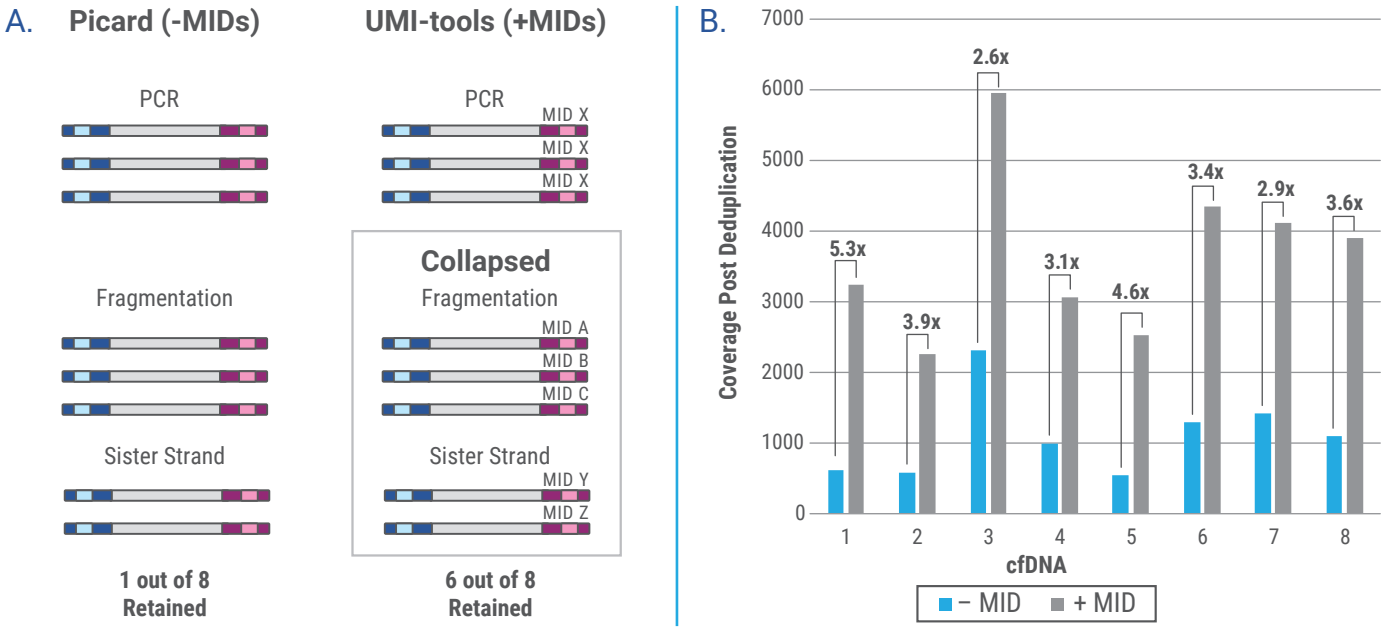


Figure 2: *A.) Left Side: Only 1x coverage was retained post-deduplication using the Picard tool in the absence of MIDs. A.) Right Side: MIDs showed an increase in data retention up to 6x coverage post deduplication. B.) cfDNA was isolated from 10 mL of blood using the Qiagen® QIAamp® Circulating Nucleic Acid Kit. Libraries were prepared using Accel-NGS 2S Hyb Kits with MIDs and enriched for oncology-related genes and hotspots with the Agilent ClearSeq Comprehensive Cancer Panel that covers a 790 kb target containing 151 genes. Captured libraries were sequenced on the Illumina HiSeq® platform to an average depth of 20,000X. Deep sequencing maximized the number of PCR duplicates for each molecule used to generate a consensus sequence. On average, a 3.6x increase in data retention was observed. Sample to sample variation in coverage is due to different copy number variation in different human samples or pooling for sequencing. Deduplication was performed with either standard Picard tools (without MIDs) or with the addition of UMI-tools from Fulcrum Genomics (with MIDs).*

Errors Introduced During PCR Amplification and Sequencing

After ligation, the individually MID-labeled strands of each dsDNA duplex are PCR-amplified to create sequence families that share the same MID sequences (PCR duplicates). Even with the use of high fidelity proofreading polymerases for library amplification, which minimizes errors, events occurring in early PCR cycles can contribute to artifacts in sequencing data. These errors can be removed using Swift MID technology (Figure 3, left side).

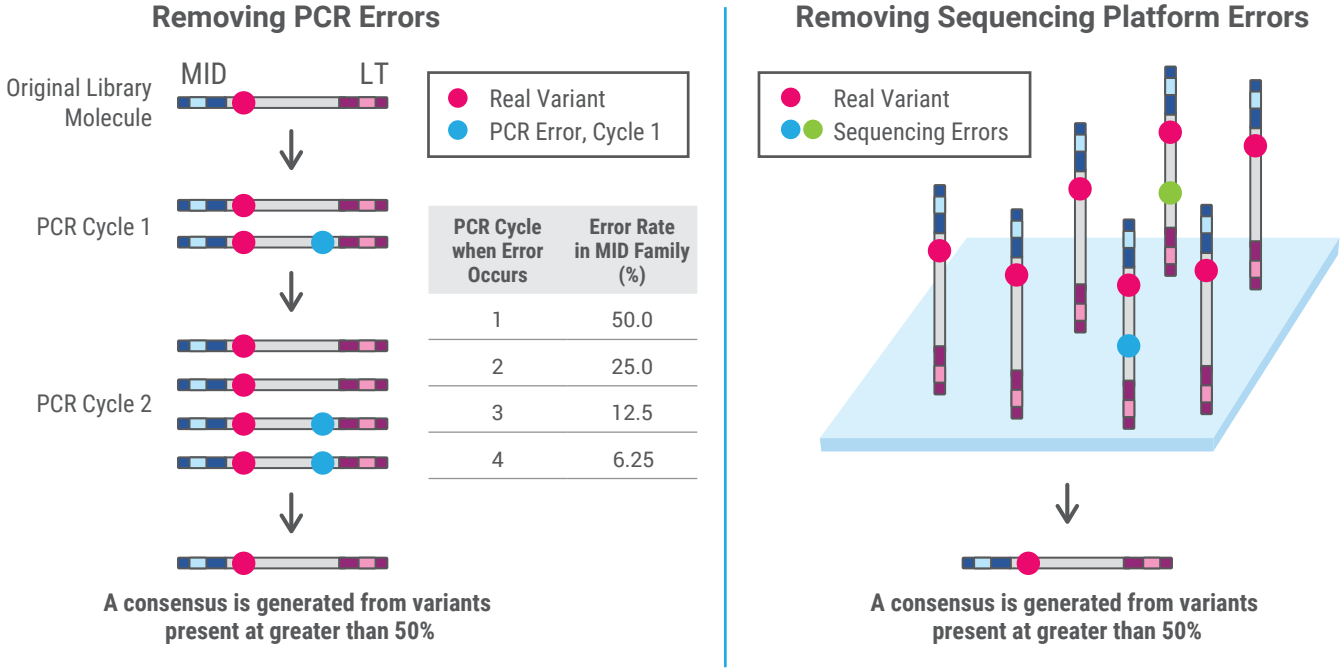


Figure 3: *Starting with a sister-strand, molecules containing the same MID can be used to generate a consensus sequence that retains true variants but removes artifacts generated by errors during PCR amplification and sequencing.*

- For polymerase errors occurring in the first cycle of PCR, the sequence artifact is expected to be present in ~50% of all MID-tagged family members.
- Errors introduced at later PCR cycles would be represented by frequencies less than 50%.
- Sequencing errors would also be expected to be present at a low frequency.

A true low allele frequency present in the original template is distinguished by its presence in greater than 50% of PCR duplicates for each MID family so the variant is retained in the SSCS. Therefore, identification of low frequency variants requires that at least three PCR duplicates are sequenced per MID family to enable SSCS analysis. Only alleles present in the majority of PCR duplicates are retained in the SSCS, any variant present in a minority of PCR duplicates is removed. Taking coverage into consideration, this corresponds to sequencing DNA libraries to **near saturation** (i.e., every unique molecule from the library is sequenced more than once).

Considerations for Low Frequency Allele Detection

To detect a true low allele frequency, there should be a balance between the target copy number of the input DNA, the uniformity of sequence coverage generated, and the lower threshold of detection.

- Sequencing to near saturation is more easily achieved from smaller targeted panels than larger targeted panels, which require a significantly higher volume of sequencing (e.g., small gene panel vs. exome). Therefore, if near-saturation is desired, it is highly recommended to use the smallest targeted panel possible to more easily achieve adequate coverage of your desired target sequences (see Figure 4 for sequencing results from an 0.8 Mb target panel).
- Additionally, higher input genome copy number is proportional to a higher library complexity and impacts the depth of sequencing required to achieve near saturation. However, genome copy number is also proportional to desired sensitivity (see Table 2), so a balance between the two must be considered.

- In order to ensure representation of DNA fragments carrying low frequency alleles within a sample, sufficient genome copy number is needed. Sample inputs with a predicted low frequency allele copy number of < 10 may not be present in the DNA sample due to Poisson distribution of DNA fragments in solution. Therefore, to avoid multiple sample replicates in the experimental design, and to ensure low copy number representation, minimally a copy number of 10 or greater should be considered (Table 2).

Input DNA (ng)	Genome Copy #	1% AF Copy #	0.5% AF Copy #	0.1% Copy #
1	330	3.3	1.6	< 1.0
10	3300	33.0	16.5	3.3
20	6600	66.0	33.0	6.6
40	13200	132.0	66.0	13.2

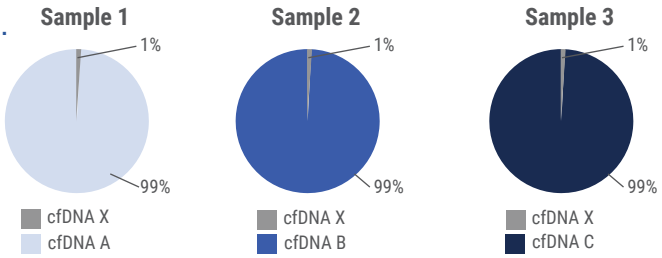
Table 2: *Theoretical maximum coverage and lowest allele frequency (AF) sensitivity for different DNA amounts and their corresponding genome copies.*

The Power of MIDs for Low Frequency Allele Detection

We performed an experiment in which cfDNA extracted from four individuals with unique genetic backgrounds was cross-spiked at a frequency of 1%. In addition, Coriell gDNA samples from different genetic backgrounds were cross-spiked at a frequency of 0.5% and 1%. MID-tagged libraries were then prepared and enriched with the IDT xGen® Pan-Cancer Panel that covers an 800 kb target containing 127 genes. The libraries were sequenced on an Illumina HiSeq to a minimum of 8000x coverage. A SSCS was generated for each MID family (BMFtools) and data were analyzed for detection of previously identified homozygous SNPs derived from the spike-in DNA samples.

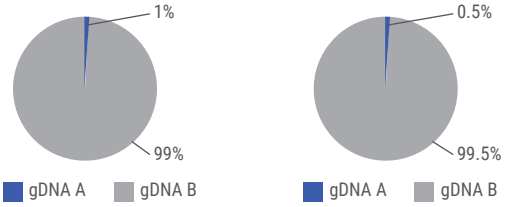
Results: Across all three samples, a total of 27/27 expected variants were consistently detected in both 0.5% and 1% gDNA samples.

A.



	Allele Frequency					
	Sample 1		Sample 2		Sample 3	
Chr: Position	Expected (%)	Observed (%)	Expected (%)	Observed (%)	Expected (%)	Observed (%)
2: 212244718	1.0	1.05	1.0	0.87	1.0	0.77
12: 25361074	1.0	1.15	1.0	1.16	1.0	1.01
12: 25361142	1.0	1.40	1.0	0.97	1.0	0.66
12: 25361646	1.0	1.39	1.0	1.40	1.0	0.59
12: 40688695	1.0	0.71	1.0	0.97	1.0	0.55
12: 115108136	1.0	0.90	1.0	1.96	1.0	0.70

B.



	Allele Frequency			
	Sample 4		Sample 5	
Chr: Position	Expected (%)	Observed (%)	Expected (%)	Observed (%)
2: 212243011	1.0	1.1	0.5	0.6
2: 212244761	1.0	0.9	0.5	0.3
2: 212245090	1.0	0.5	0.5	0.4
2: 212245489	1.0	1.3	0.5	0.7
3: 176738798	1.0	1.1	0.5	0.6
3: 176739663	1.0	1.2	0.5	0.7
3: 176741730	1.0	0.7	0.5	0.4
5: 38475507	1.0	1.4	0.5	0.7

	Allele Frequency			
	Sample 4		Sample 5	
Chr: Position	Expected (%)	Observed (%)	Expected (%)	Observed (%)
5: 38481335	1.0	0.8	0.5	0.6
7: 106547469	1.0	1.0	0.5	0.8
7: 106547921	1.0	0.5	0.5	0.6
7: 26224668	1.0	0.9	0.5	0.8
7: 55238464	1.0	0.9	0.5	0.8
7: 55238874	1.0	0.8	0.5	0.8
10: 8116598	1.0	1.1	0.5	0.5
11: 32409625	1.0	1.4	0.5	0.7
11: 32410002	1.0	1.1	0.5	0.5
11: 32410337	1.0	1.1	0.5	0.6
11: 32452240	1.0	1.1	0.5	0.4
17: 70121339	1.0	0.9	0.5	0.9
17: 70122108	1.0	1.1	0.5	0.8
19: 1224934	1.0	0.6	0.5	0.6
19: 1225054	1.0	0.8	0.5	0.7
19: 1228191	1.0	1.1	0.5	0.7
20: 31024274	1.0	1.0	0.5	0.8
20: 31025535	1.0	1.1	0.5	1.2
X: 41093413	1.0	0.8	0.5	0.5

Figure 4: *MID libraries including MID index sequences were prepared with Accel-NGS 2S Hyb Kits. 6 (A. table) and 27 (B. table) homozygous SNPs were detected for cfDNA and gDNA spike-ins, respectively.*

The Power of MIDs for Increased Specificity in Variant Calling

MIDs have a subtle effect on the number of variants called at high allele frequencies, but substantially reduce the number of low frequency variants called. By removing sequencing and PCR errors, the Swift MID technology eliminates noise and highly enriches for true variants; thereby, improving variant calling specificity.

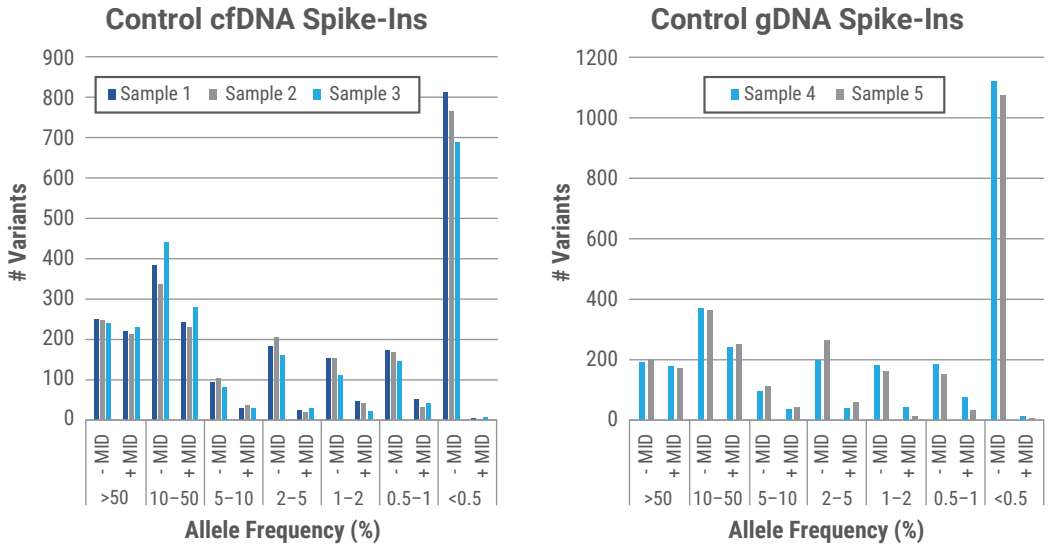


Figure 5: *Total variants called at various allele frequencies with or without the use of MIDs are depicted from the spike-in experiments. MIDs only have a subtle effect on the number of variants called at high allele frequencies, but substantially reduce the number of low frequency variants called. This is the result of removing sequencing and PCR errors such that variants called are highly enriched for true variants and the removed variants represent noise. In this way MIDs lead to increased specificity in low frequency variant calling.*

Bioinformatic Analysis of Low Frequency Alleles

After accurate demultiplexing of the sequencing run, there will be 4 (R1, I1, I2, and R2) or 3 (R1, R2, and R3) FASTQ read files per sample where the 9 bp MID sequence should be in the I2 file in the former, and R2 in the latter set of FASTQ files. For MID runs, the following command provides the bcl2FASTQ files:

bcl2fastq -R <Illumina_RunFolder> --sample-sheet <SampleSheet_P5-MID.csv> -o <output_folder_name> --minimum-trimmed-read-length 0 --mask-short-adapter-reads 0 --use-bases-mask Y*,I*,Y*,Y*

The goals of MID inclusion in data analysis include both accurate deduplication, which increases data retention, and removal of PCR and sequencing errors to enable low frequency allele detection.

There are multiple publicly available tools that can be used for data analysis. Below is the brief overview of the two sets of tools that have been utilized to analyze Swift's MID data.

Instructions for Bioinformatic Analysis Using BMFtools (https://github.com/ARUP-NGS/BMFtools)

1. Collapse FASTQ reads using MID using "secondary" option:
   - Bmftools collapse secondary -s 5 -p 12 -o tmp_prefix -f out_prefix -i Collapse Fastq reads using MID using "secondary" option
2. Alignment (including numerous processes, hence I/O piping, see manual online):
   - bwa mem -CYT0 -t<threads> $REF R1_collapsed.fq R2_collapsed.fq | bmftools mark | bmftools sort -T tmpfile –o pre_rsq.bam – **(Align)**
   - Bmftools rsq –sf tmp.fq –I0 pre_rsq.bam - | samtools sort –O bam –T tmp-sort –o tmp.bam – **(Rescue)**
   - bwa mem –pCYT0 $REF tmp.fq | bmftools mark –I0 | samtools sort –I0 –O bam –T tmp - | samtools merge -fh <tmp.bam> <final.bam> <tmp.bam> – **(Align Rescued Reads)**
3. Get MID usage metrics, family size, and target depth:
   - Bmftools famstats <final.bam>
   - Bmftools depth –sb <bed_file> <final.bam>
4. Variant calling using Lofreq has been validated using internal spike-in datasets.

Instructions for Bioinformatic Analysis Using Fulcrum Genomics (https://github.com/fulcrumgenomics/fgbio)

1. Align FASTQ reads (BWA)
2. Connect MID (I2 FASTQ) to bam alignments: fgbio, AnnotateBamWithUmis
3. Optional: Determine gain in data by using Picard MarkDuplicate with option BARCODE_TAG=RX. Output should show lower rate of "% Duplication" than without using MIDs.
4. Continue with following steps to get MID base consensus sequence using fgbio & Picard Tools:

   a. RevertSam: java -jar $picard RevertSam I=$OUT.midbased.markdup.bam O=$OUT.sanitised.bam SANITIZE=true REMOVE_DUPLICATE_INFORMATION=false REMOVE_ALIGNMENT_INFORMATION=false

   b. Fgbio SetMateInformation

   c. Fgbio GroupReadsByUmi

   d. Fgbio CallMolecularConsensusReads

5. Consensus Bam does not contain any mapping information, hence use BamToFASTQ to generate consensus Fastq reads which can then be aligned to reference followed by variant calling.

Glossary of Terms

Fragmentation Duplicate: Some DNA fragmentation is non-random, where fragmentation duplicates can be produced and sequenced (this is more frequent in cfDNA where nucleosomal positioning is non-random). Fragmentation duplicates share a common paired-end (PE) map position but have different MIDs. This is how they are distinguished from PCR duplicates that share a common MID.

MID Family: A group of PCR duplicates defined by common PE map position and MID sequence.

MID Family Size: The number of PCR duplicates within an MID family. A minimum of three are required in order to generate a SSCS.

MID Sequence: A 9-base random N sequence that is uniquely attached to each library molecule and is read during Index 2 sequence.

Number of MID Families: The number of unique molecules tagged with individual MID sequences, represents the copy number of unique strands that have been sequenced from a library.

PCR Duplicate: Progeny of exponential PCR amplification that share a common PE map position and MID sequence.

PE Map Position: PE map position is when a PE read is aligned to particular coordinates of a reference genome. This information is used for deduplication by standard Picard tools, which is why it cannot distinguish PCR duplicates from sister-strand or fragmentation duplicates.

Sequencing to Saturation: When a sample is sequenced to a depth where further sequencing does not produce additional unique library molecules; also characterized by most library molecules being sequenced multiple times to produce MID families of sufficient size for SSCS determination.

Sister-strand Duplicate: Each strand of a dsDNA input molecule are independently tagged with MIDs, amplified and sequenced. They share a common PE map position due to the Repair II polishing reaction during library preparation, but are identified as separate molecules by having different MIDs. This is how they are distinguished from PCR duplicates that share a common MID.

SSCS: A single-strand consensus sequence that eliminates PCR and sequencing errors, where a minimum MID family size of three is required (i.e., where 2/3 or 3/3 majority variants are retained).

References

1. Kennedy et al., *Nature Protocol*, 2586-2606 (2014).
2. Newman et al., *Nature Biotechnology*, 547-555 (2016).
3. Schmitt et al., *Proc Natl Acad Sci*, Vol. 109 No. 36 (2012).

Swift Biosciences, Inc.
674 S. Wagner Road, Suite 100 • Ann Arbor, MI 48103 • 734.330.2568 • www.swiftbiosci.com