AACR 2018: #2230

Low Frequency Variant Detection in Cell Free DNA by Applying Molecular Identifiers to Targeted NGS

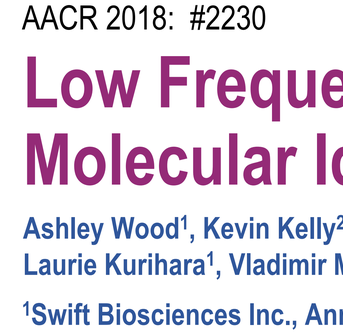
Ashley Wood[1], Kevin Kelly[2], Sukhinder Sandhu[1], Melissa Soucy, Mida Pezeshkian[1], Vanessa Kelchner[1], Jordan RoseFigura[1], Justin Lenhart[1], Honey Reddi[2], Laurie Kurihara[1], Vladimir Makarov[1]

[1]Swift Biosciences Inc., Ann Arbor, MI [2]The Jackson Laboratory, Farmington, CT

Introduction

The growing use of liquid biopsy for early detection and monitoring of disease necessitates accurate variant detection at <1% allele frequencies due to a low population of disease DNA within circulating, cell-free DNA (cfDNA). Reliable, low-frequency variant detection by next-generation sequencing (NGS) is challenging due to background noise from PCR and sequencing errors. We employed molecular identifiers (MIDs) to uniquely label individual DNA molecules prior to amplification, facilitating the distinction of true variants from PCR and sequencing errors. We incorporated MIDs in both our amplicon library prep that uses multiplex PCR for targeted NGS and our whole genome library prep followed by targeting with hybridization capture using an 800kb pan-cancer panel. We performed low frequency spike-in experiments at <1% allele frequencies. We prepared MID libraries with various amplicon panels including a 17 amplicon EGFR pathway panel and a 104 amplicon SNP panel. Deep sequencing to >30,000x was done to maximize MID family size (number of PCR duplicates) and optimize generation of a consensus sequence. This analysis identified all known variants present at 1%, 0.5%, and 0.25% allele frequencies. Next, the hybridization capture libraries were prepared with low-frequency spike-in samples, sequenced to >8000x, and all known variants at 1% and 0.5% allele frequencies were maintained in the consensus data. In both cases, the number of false positives was reduced, resulting in improved specificity. Further, EGFR amplicon libraries and hybridization capture libraries were prepared using cfDNA samples from lung, ovarian, liver, stomach, and colon cancers. Variant calling based on MID generated consensus sequences identified mutations in cfDNA samples as well as corresponding tumor and normal samples when available. This study highlights the ability of MID technology to enable low frequency variant detection, critical to track known variants and identify novel pathogenic mutations in cfDNA samples.

Improved Data Analysis with MIDs



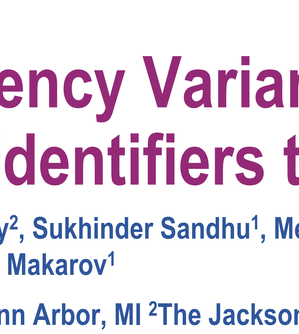


Figure 1. MIDs label individual molecules prior to exponential amplification by PCR facilitating the accurate identification and removal of PCR duplicates. Furthermore, molecules containing the same MID can be used to generate a consensus sequence that retains true variants but removes artificial mutations generated by polymerase errors during PCR amplification and sequencing. Here we depict how PCR duplicates from one MID family are used to create a consensus.

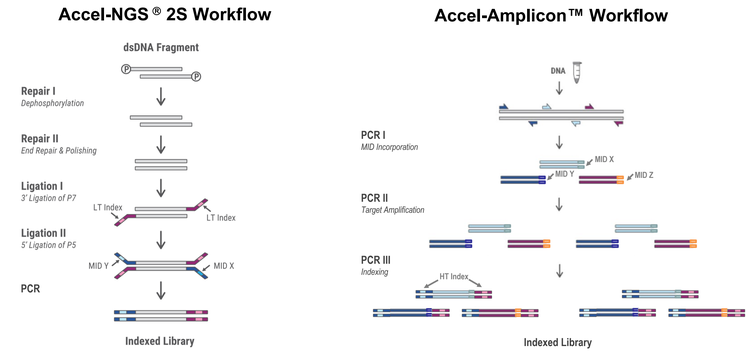
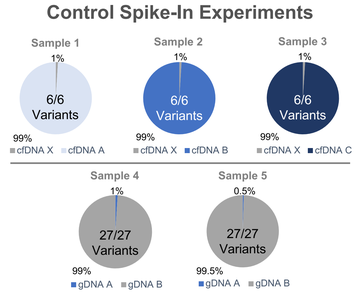
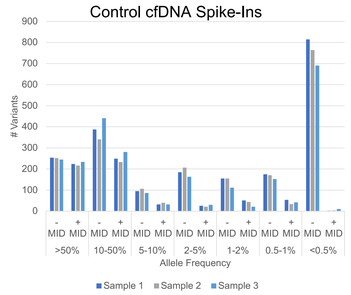


Figure 2. MIDs can be added to both our Accel-NGS 2S whole genome library prep used for hybridization capture and to our Accel-Amplicon library prep that uses multiplex PCR for targeting. Both library preps use the MID to label unique library molecules prior to amplification. Accel-NGS 2S libraries with MIDs are constructed using Illumina®-compatible adapters with a strand-specific 9 base random MID in the i5 index position and a sample index in the i7 position. The Accel-Amplicon workflow consists of a 2-cycle MID incorporation step, PCR amplification of targeted amplicon molecules, and an indexing PCR step that adds Illumina-compatible adapters with HT indexes. The final library molecules consist of a 10-base in-line random N sequence positioned at the start of Read 2.

Hybridization Capture with MIDs

Increased Specificity in Variant Calling with MIDs





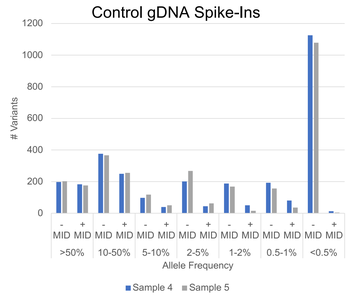


Figure 3. cfDNA was extracted from blood of four individuals with unique genetic backgrounds and gDNA samples from different genetic backgrounds were obtained (Coriell Institute). Sample spike-ins were performed at 1% or 0.5% frequency into 10 ng cfDNA or 100 ng gDNA. Libraries were prepared with Swift's Accel-NGS 2S Hyb Kit with MIDs, enriched with the IDT xGen® Pan-Cancer Panel that covers an 800kB target containing 127 genes, and sequenced on an Illumina HiSeq® to a minimum of 8000x coverage. Data were analyzed with BMF tools. A consensus sequence was generated for each MID family and data were analyzed for homozygous SNPs present in the spike-in sample only. (Left) 6/6 known variants were present in all three 1% cfDNA samples and 27/27 known variants were present in both 1% and 0.5% gDNA samples depicting the power of MIDs for low frequency variant calling. (Right) Total variants called at various allele frequencies with or without the use of MIDs are depicted. MIDs have a only a subtle effect on the number of variants called at high allele frequencies, but substantially reduce the number of low frequency variants called by eliminating sequencing and PCR errors.

Increased Data Retention with MIDs

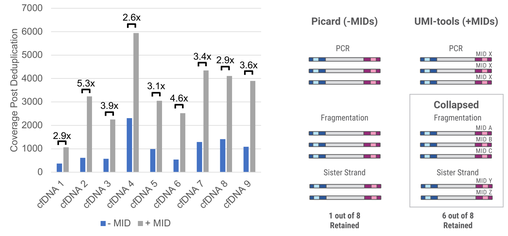


Figure 4. We evaluated the effect of MIDs on data retention after deduplication. Deduplication was performed with either standard Picard tools (- MID) or UMI-tools from Fulcrum Genomics (+ MID). MIDs allow for accurate identification and removal of PCR duplicates while maintaining sister strand duplicates and fragment duplicates. Deduplication using MIDs showed an increase in coverage for all samples analyzed.

Variant Analysis from cfDNA Samples

cfDNA Library Preparation and Sequencing with MIDs

Sample	Cancer Type	Patient	Library Input (ng)	Read #	Raw Coverage	Duplication Rate	% On Target	#COSMIC Mutations (0.5-15% Allele Frequency)
cfDNA 1	Ovarian	A	20	81,387,833	14,135	94%	74%	21
cfDNA 2	Bile duct	B	20	100,905,882	22,452	78%	70%	14
cfDNA 3	Kidney	C	20	78,450,496	17,484	76%	71%	17
cfDNA 4	Stomach	D	20	77,176,513	17,101	67%	71%	9
cfDNA 5	Colon	E	20	69,214,598	15,778	74%	72%	8
cfDNA 6	Colon	F	20	111,063,000	24,975	79%	71%	32
cfDNA 7	Unknown	G	20	100,453,057	23,139	75%	73%	10
cfDNA 8	Bile duct	H	20	76,763,375	17,694	72%	72%	3
cfDNA 9	Colon	I	20	100,766,501	23,210	77%	73%	5

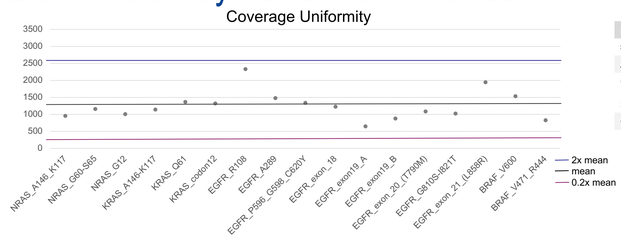
Tumor/cfDNA Variant Validation

Chr: Position	Reference	Alternate	Gene	Cosmic ID	Allele Frequency			
					Normal	Tumor site 1	Tumor site 2	cfDNA 1
17:7578437	G	A	TP53	COSM3388212	0.0%	95.6%	97.7%	11.2%

Figure 5. cfDNA libraries were prepared using the Accel-NGS 2S Hyb kit with MIDs and enriched for oncology-related genes and hotspots with the IDT xGen Pan-Cancer Panel (cfDNA 1) or the Agilent ClearSeq® Comprehensive Cancer Panel (cfDNA 2-9). Sequencing was performed on an Illumina HiSeq to greater than 14,000x prior to deduplication. Low frequency, COSMIC mutations were identified in all 9 cfDNA samples. For patient A (a 75-year-old female with stage 3B, grade 3 ovarian carcinosarcoma), additional samples were available. A normal sample and biopsies from two different sites on one of the patient's tumors were taken during recurrent surgery performed 9 months after the primary surgery and a pathogenic TP53 mutation was identified in both tumors and in the cfDNA sample but not in the normal sample.

Multiplexed PCR with MIDs

EGFR Pathway Panel Performance



Feature	Specification
# Amplicons	17
Amplicon Size	107-149bp
Coverage Uniformity at > 0.2x Mean	>95%
On Target	>95%

Panel Metrics with cfDNA Samples

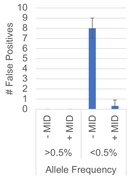
Sample	Cancer Type	%Bases On Target	% Coverage Uniformity
cfDNA 1	Lung	95.4	100.0
cfDNA 2	Lung	99.6	93.8
cfDNA 3	Lung	93.2	100.0
cfDNA 4	Lung	96.2	100.0
cfDNA 5	Lung	95.8	100.0
cfDNA 6	Lung	96.9	100.0
cfDNA 7	Breast	97.7	100.0
cfDNA 8	Prostate	95.7	100.0
cfDNA 9	Lung	94.1	100.0
cfDNA 10	Breast	92.6	100.0
cfDNA 11	Prostate	92.1	100.0
cfDNA 12	Colorectal	94.7	100.0
cfDNA 13	Mesothelioma	93.9	100.0
Ctrl 1 (SeraCare)	N/A	98.6	100.0
Ctrl 2 (SeraCare)	N/A	97.8	100.0
Ctrl 3 (HD701)	N/A	95.4	100.0

Figure 6. (Top) 10 ng of control gDNA (Coriell NA12878) was used to test the Accel-Amplicon EGFR Pathway Panel with MIDs. This 17 amplicon panel shows coverage at greater than 0.2x the mean for all amplicons before and after deduplication using MIDs (fgbio, Fulcrum Genomics). (Bottom) Panel metrics from an EGFR MID prototype panel are shown for libraries prepared with cfDNA samples from patients with various cancer types by The Jackson Laboratory. Input quantity varied for cfDNA samples based on material available.

Identification of Variants down to 0.25% from 10ng

Variant Calling with the EGFR Pathway Amplicon Panel with MIDs

Chr	Position	Gene	Variant	Reference	Alternate	Expected Allele Freq.	Observed Allele Freq.
7	55241707	EGFR	G719S	G	A	12.25%	15.50%
12	25398281	KRAS	G13D	C	T	7.50%	6.10%
1	115256530	NRAS	Q61K	G	T	6.25%	4.20%
7	140453136	BRAF	V600E	A	T	5.25%	5.27%
12	25398284	KRAS	G12D	C	T	3.00%	2.26%
7	55259515	EGFR	L858R	T	G	1.50%	1.22%
7	55242464	EGFR	ΔE746-A750	AGGAATTAAGAGAAGC	A	1.00%	1.11%
7	55249071	EGFR	T790M	C	T	0.50%	0.52%



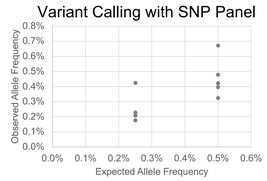


Figure 7. (Top) Horizon Diagnostics Quantitative Multiplex DNA Standard (HD701) was spiked into Coriell DNA (NA12878) at 50% to obtain expected variants at allele frequencies from 12.25-0.50%. Libraries were prepared using 10 ng of input DNA and the EGFR MID panel. Sequencing was performed on an Illumina MiniSeq® to greater than 100,000x prior to deduplication. PCR duplicates were defined based on MID analysis with fgbio (Fulcrum Genomics). All expected variants were consistently detected in the consensus sequence and the use of MIDs removed false positives at low allele frequencies. (Upper Right) The graph depicts the average number of false positives (n=3) called by LoFreq with and without the use of MIDs. (Lower Right) We also developed a 104 amplicon SNP panel to validate low frequency variant calling. With this panel and Coriell DNA spike-in samples, we were able to detect 5/5 variants present at 0.5% and 5/5 variants present at 0.25%.

Conclusion

- Labeling unique library molecules with MIDs prior to amplification allows for the removal of sequencing and PCR induced errors during data analysis.
- Both Accel-NGS 2S and Accel-Amplicon with MIDs are compatible with cfDNA.
- The use of MIDs for de-duplication results in increased data retention through the accurate distinction of PCR duplicates from fragmentation and complementary strand duplicates.
- Inclusion of MIDs in NGS library preparation requiring PCR improves variant calling by increasing specificity at low allele frequencies. Here we are able to detect known variants at less than 1% allele frequencies using hybridization capture and amplicon approaches for targeted NGS.

Swift Biosciences, Inc.
58 Parkland Plaza, Suite 100 • Ann Arbor, MI 48103 wwwswiftbiosci.com

www.swiftbiosci.com