AMP 2017: #ST120

Using Liquid Biopsies for Low Frequency Variant Detection

Andrew McUsic, Justin Lenhart, Ashley Wood, Sukhinder Sandhu, Cassie Schumacher, Laurie Kurihara, Vladimir Makarov, Tim Harkins

Swift Biosciences, 58 Parkland Plaza, Suite 100, Ann Arbor, MI 48103, Tel: 734.330.2568

Introduction

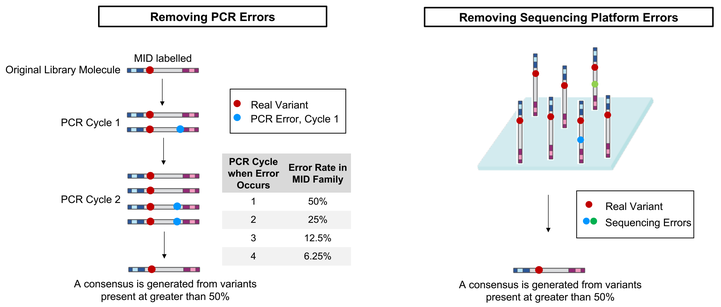
INTRODUCTION: Liquid biopsy assays enable non-invasive profiling of circulating, cell-free DNA (cfDNA) and circulating tumor cell DNA to assist in early-stage diagnosis of disease and monitoring treatment response. Since high sequencing depth is required to profile cfDNA variants, most liquid biopsy assays use targeting to cost-effectively achieve deep coverage of target loci for detection of pathogenic variants as low as 1% allelic fraction. An assay that produces uniform, comprehensive coverage from low DNA quantities is critical for obtaining the necessary sensitivity. We developed a liquid biopsy workflow for low frequency variant detection from a 10 mL blood draw combined with Accel-NGS® library preparation.

METHODS: Whole blood samples were collected in Streck Cell-Free DNA BCT® vials from patients with late stage cancer and cfDNA was extracted with the Promega Maxwell RSC. A total of 10 ng cfDNA was used to make an Accel-NGS 2S Hyb library followed by hybridization capture using IDT xGen® Pan Cancer probes. Molecular barcodes were incorporated to label each library molecule uniquely prior to PCR amplification. Sequencing was performed to a minimum of 8000X mean bait coverage. Variant calling was performed using Vardict and LoFreq.

RESULTS: Extraction yielded 8-32 ng cfDNA with a size peak of 170 bp and a mean qPCR integrity score of 0.22, characteristic of high quality cfDNA lacking cellular DNA. The Accel-NGS 2S Hyb Library Kit exhibited 90% conversion with cfDNA, providing highly complex libraries with uniform target coverage (>99% of bases covered >100X). Molecular barcodes enabled removal of PCR duplicates while preserving fragmentation and strand duplicates to maximize coverage. Barcoded molecules were grouped to generate consensus sequences after removal of false positives originating from PCR and sequencing errors, distinguishing signal from noise. Sensitive and precise detection of variants was achieved down to 0.5% allele frequency. We also validated variant calling below a 1% allele frequency using the Accel-Amplicon EGFR Pathway Panel with MIDs. Libraries were prepared with cfDNA and tumor samples from individuals with ovarian, liver, stomach, and colon cancers, were sequenced to a minimum of 13,000x coverage, and we determined data retention after de- duplication with and without the use of MIDs. We observed a significant increase in data retention that led to a 2- to 5-fold increase in coverage using MIDs. Variant calling identified pathogenic mutations in all cfDNA samples, including those present in a corresponding tumor sample when available.

CONCLUSIONS: We developed two liquid biopsy workflows for low frequency variant detection from clinically relevant quantities of cfDNA. This approach provides a powerful method for detecting, identifying, and monitoring disease.

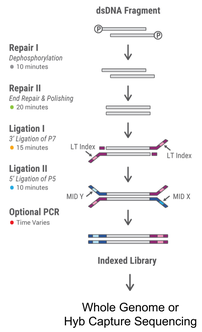
Improved Data Analysis with MIDs



PCR Cycle when Error Occurs	Error Rate in MID Family
1	50%
2	25%
3	12.5%
4	6.25%

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 PCR duplicates from one MID family to demonstrate that PCR and sequencing errors should not exist at greater than 50% and are therefore eliminated in the consensus sequence.

MIDs Label Unique 2S Library Molecules



Uniform Incorporation of MIDs

Library	Unique MIDs*	Mean MID Copy #	MID Copy Uniformity	Total Index 2 (i5) Reads
PCR-Free	262,104	20.98	98%	5,498,922
9 PCR Cycles	262,036	18.50	99%	4,848,690

*Theoretical maximum 262,144

Figure 2. The Swift MID is a 9 base random sequence that uniquely labels individual library molecules prior to amplification. Accel-NGS 2S libraries with MIDs are constructed using Illumina®-compatible adapters with a strand-specific MID in the i5 index position and a sample index in the i7 position. Each dsDNA substrate receives two independent MID adapters "X" and "Y". To validate uniform MID incorporation, libraries were analyzed for number of unique MID sequences and MID copy uniformity. Swift's 2S MID libraries attain close to theoretical representation of MID combinations with high copy uniformity (percentage of MIDs that are within 20% of the average MID copy number). These validation metrics are important indicators of the technology's ability to uniquely identify library molecules.

gDNA and cfDNA Spike-in Experiments

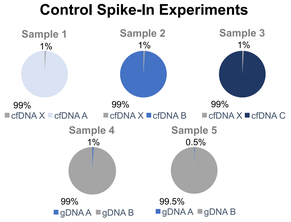


Figure 3. cfDNA was extracted from blood of four individuals with unique genetic backgrounds and gDNA samples from different genetic backgrounds were obtained (Coriell Institute). To determine the effect of MIDs on low frequency variant calling, sample spike-ins were performed at 1% or 0.5% frequency 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 800 kb target containing 127 genes, and sequenced on an Illumina HiSeq® to a minimum of 8000x coverage.

Results

Identification of Variants Down to 0.5%

cfDNA Spike-In Variants

Chr: Position	Sample 1 Exp.	Sample 1 Obs.	Sample 2 Exp.	Sample 2 Obs.	Sample 3 Exp.	Sample 3 Obs.
	ALLELE FREQUENCY					
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%**

gDNA Spike-In Variants

Chr: Position	Sample 4 Exp.	Sample 4 Obs.	Sample 5 Exp.	Sample 5 Obs.
	ALLELE FREQUENCY			
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%**
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. Data from 2S Hyb spike-in experiments 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. 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.

Increased Specificity with MIDs

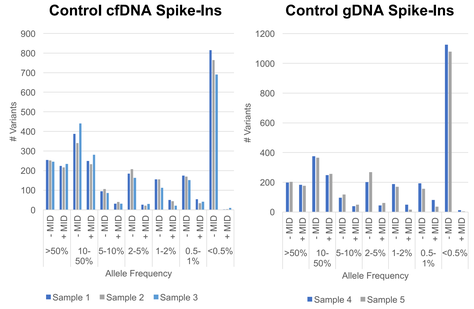


Figure 5. Total variants called at various allele frequencies with or without the use of MIDs are depicted from the spike-in experiments. MIDs only have a subtle effect on the number of variants called at high allele frequencies, but substantially reduce the number of low frequency variants called. This 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.

Variant Analysis from cfDNA Samples

cfDNA Library Preparation and Sequencing with MIDs

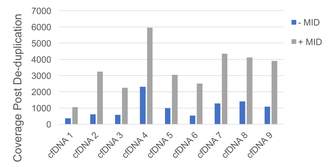
Sample	Cancer Type	Patient	Library Input (ng)	Read #	Raw Coverage	Duplication Rate	% On Target	#COSMIC Mutations (0.5-15% Allele Frequency)
cfDNA 1	Ovarian	A	20	81,387,833	14,135	94%	74%	21
cfDNA 2	Bile duct	B	20	100,905,882	22,452	78%	70%	14
cfDNA 3	Kidney	C	20	78,450,496	17,484	76%	71%	17
cfDNA 4	Stomach	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 6. 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 de-duplication. Deep sequencing maximized the number of PCR duplicates sequenced for each molecule used to generate a consensus sequence. 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. Of the variants found in the two tumor samples but not in the normal sample, a pathogenic TP53 mutation was identified at close to 100% in samples from both tumors and in the cfDNA sample at 11%.

Increased Data Retention with MIDs



Type of Read Duplicate	Read Removed - MID	Read Removed + MID
PCR duplicates	✓	✓
Sister strand duplicates	✓	✗
Fragment duplicates	✓	✗

Figure 7. We evaluated the effect of MIDs on data retention after de-duplication. De-duplication 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. De-duplication using MIDs showed an increase in coverage for all samples analyzed.

MIDs Label Unique Amplicon Library Molecules

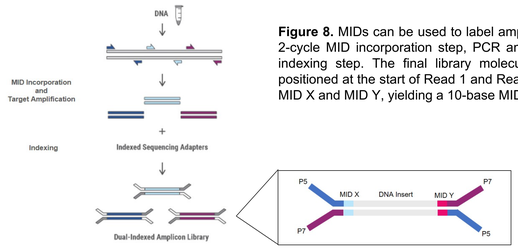
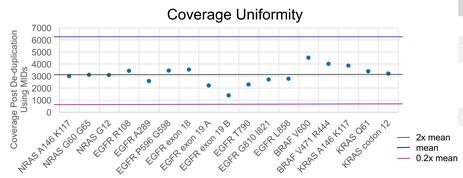


Figure 8. MIDs can be used to label amplicon library molecules. This workflow consists of a 2-cycle MID incorporation step, PCR amplification of targeted amplicon molecules, and an indexing step. The final library molecule consists of two 5-base random N sequences positioned at the start of Read 1 and Read 2. Each original DNA molecule receives an unique MID X and MID Y, yielding a 10-base MID/barcode.

EGFR MID Panel Performance



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

Figure 9. 10 ng of control gDNA was used to test the Accel-Amplicon EGFR MID Pathway Panel. This 17 amplicon panel shows coverage at greater than 0.2x the mean for all amplicons before and after de-duplication using MIDs (fgbio, Fulcrum Genomics).

Variant Calling with MIDs at Low Frequencies

Variant Calling with the EGFR-MID Amplicon Panel

Chr	Position	Gene	Variant	Reference	Alternate	Total Coverage	Variant Coverage	Expected Allele Freq.	Observed Allele Freq.
7	55241707	EGFR	G719S	G	A	3626	562	12.25%	15.50%
12	25398281	KRAS	G13D	C	T	3540	216	7.50%	6.10%
1	115256530	NRAS	Q61K	G	T	3428	144	6.25%	4.20%
7	140453136	BRAF	V600E	A	T	4972	262	5.25%	5.27%
12	25398284	KRAS	G12D	C	T	3540	80	3.00%	2.26%
7	55259515	EGFR	L858R	T	G	2790	34	1.50%	1.22%
7	55242464	EGFR	△E746-A750	AGGAATTAAGAGAAGC	A	3952	44	1.00%	1.11%
7	55249071	EGFR	T790M	C	T	2329	12	0.50%	0.52%

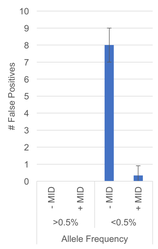


Figure 10. Horizon Diagnostics Quantitative Multiplex DNA Standard (HD701) was spiked into Coriell DNA (NA12878) at 50% to obtain expected variants at allele frequencies from 12.25-0.50%. Libraries were prepared using 10 ng of input DNA and the EGFR MID panel. Sequencing was performed on an Illumina MiniSeq® to greater than 100,000x prior to de-duplication. 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. The graph to the right depicts the average number of false positives (n=3) called by LoFreq with and without the use of MIDs.

Conclusion

- Labeling unique library molecules with MIDs prior to amplification allows for the removal of sequencing and PCR induced errors during data analysis.
- 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.
- 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.

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

www.swiftbiosci.com