ACMG 2017: #175

Leveraging Molecular IDs to Improve Coverage of Low Frequency Variants in Ovarian Cancer Using Next Generation Sequencing

Andrew McUsic[1], Ashley Wood[1], Matthew Dashkoff[2], Timothy Harkins[1], Sukhinder Sandhu[1], Olga Camacho-Vanegas[2], Peter Dottino[2], Laurie Kurihara[1], John Martignetti[2], Vladimir Makarov[1]

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

[2]Icahn School of Medicine at Mount Sinai, Departments of Genetics and Genomic Sciences and Obstetrics/Gynecology & Reproductive Sciences, 1425 Madison Avenue New York, NY 10029

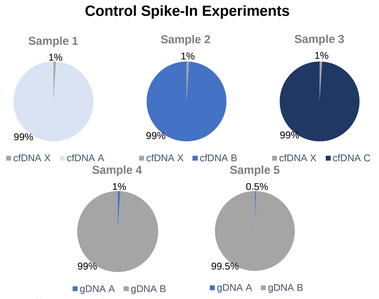
Abstract

Introduction: The detection of low frequency allele variants by next generation sequencing (NGS) holds great promise for elucidating the progression of cancer and other pathologies but also presents significant challenges for sensitivity and specificity. Below a threshold of 5% allelic fraction, abundance levels of mutations approach the assay limit of detection, and true positives are difficult to distinguish from false positives due to errors in both the amplification methods used to produce the NGS library as well as read errors inherent to the sequencing platform. Without a specific technology for uniquely labeling each library molecule, it is difficult to distinguish duplicate molecules produced by PCR from unique templates or otherwise understand the contribution of each clonal family of library molecules, especially for liquid biopsy samples where positional information is not a reliable method for distinguishing fragments due to common sites for enzymatic shearing of circulating, cell-free DNA (cfDNA). The use of Molecular IDs (MIDs) to label unique library molecules provides a method for detecting and removing PCR duplicates from the sequencing data while preserving unique information otherwise conventionally discarded during deduplication due to the presence of duplicates arising from fragmentation and complementary strands. Leveraging these MIDs allows the generation of consensus sequence from PCR duplicates to reduce PCR and sequencing errors and also increases confidence in the detection of low frequency mutations by quantifying the presence of mutations in multiple unique molecules.

Methods: Tumor samples were biopsied from a stage 3B, grade 3 ovarian carcinosarcoma (malignant mullerian mixed tumor from 75-year-old Caucasian female) and cfDNA was obtained from blood plasma collected at the time of surgery. The Accel-NGS® 2S Hyb DNA Library Kit from Swift Biosciences was combined with MID technology and a pan-cancer hybridization capture enrichment panel to prepare DNA isolated from the series of metastatic tumors for sequencing by Illumina® MiSeq® and HiSeq®. Data retention was defined as the increase in useful coverage obtained when considering MIDs versus removing duplicates without consideration of the MIDs. We determined data retention after de-duplication by comparing results from standard Picard tools without the use of MIDs to UMI-tools (Fulcrum Genomics) to utilize MIDs. Variant calling was performed using BMFtools, LoFreq, and VarDict.

Results: When sequencing the panel to an average depth of 200-300X by MiSeq, we observed an increase in data retention for both Covaris-sheared FFPE and cfDNA libraries that led to a 1.1-1.3 fold increase in coverage using UMI-tools over Picard de-duplication. This effect was most pronounced at lower DNA input quantities and with higher depth of sequencing. When sequencing the same libraries to an average depth of 1000-2000X by HiSeq, we observed a 1.7-2.9 fold increase in coverage when leveraging MIDs. These libraries were further evaluated for variant calling, and we identified variants with as low as 1% frequency present across all tumor samples and within cfDNA, as well as variants unique to each tumor sample.

Identification of Variants Down to 0.5%



gDNA Spike-In Variants

Chr: Position	ALLELE FREQUENCY Sample 4 Expected	Sample 4 Observed	Sample 5 Expected	Sample 5 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%**
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%**

cfDNA Spike-In Variants

Chr: Position	ALLELE FREQUENCY Sample 1 Expected	Sample 1 Observed	Sample 2 Expected	Sample 2 Observed	Sample 3 Expected	Sample 3 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%**

Figure 3. cfDNA was extracted from blood of four individuals with unique genetic backgrounds and Coriell gDNA samples from different genetic backgrounds were obtained. 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 the 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. A consensus sequence was generated for each MID family (BMFtools) 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.

2-Fold Increase in Data Retention with MIDs

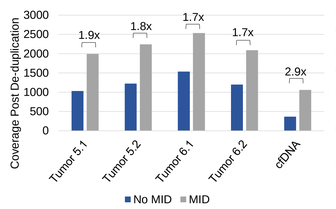


Figure 6. We evaluated the effect of MIDs on data retention after de-duplication from the ovarian cancer samples. Libraries were prepared from DNA isolated from fresh tissue or cfDNA. De-duplication was performed with either standard Picard tools (without MIDs) or UMI-tools from Fulcrum Genomics (with MIDs). MIDs allow for accurate identification and removal of PCR duplicates while maintaining complementary strand duplicates and fragment duplicates. De-duplication using MIDs showed an increase in coverage for all samples analyzed regardless of the sample type. This effect was maximized with lower DNA inputs and higher sequencing depths.

Loss of Heterozygosity in Ovarian Cancer Samples

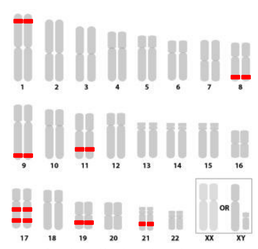
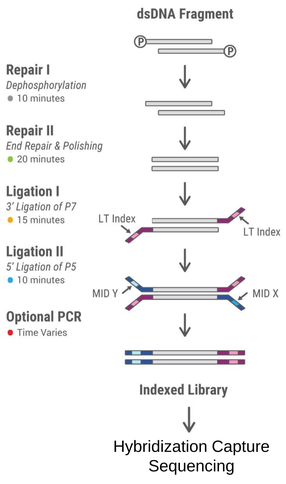


Image from https://chubaoyolu.org/

Chr 9: Position	Alt.	Gene	Cosmic ID	Allele Frequency Tumor 5.1	Tumor 5.2	Tumor 6.1	Tumor 6.2	cfDNA
139390822	G>A	NOTCH1	28763	**97.8%**	**97.7%**	47.7%	47.6%	50.0%
139397707	G>A	NOTCH1	1461158	**96.3%**	**98.3%**	53.1%	51.6%	48.6%
139407932	A>G	NOTCH1	3763788	**98.0%**	**97.9%**	50.7%	50.2%	52.5%
139410467	G>A	NOTCH1	-	**2.0%**	**1.8%**	45.8%	48.4%	49.9%
139418260	A>G	NOTCH1	1756050	**97.8%**	**96.0%**	50.3%	50.0%	49.8%

Figure 7. Tumor 5 (malignant) shows a substantial loss of heterozygosity (LOH) at multiple loci throughout the genome. Regions of LOH are marked in red. An example of this within the NOTCH1 gene is depicted where tumor 6 (benign) and the cfDNA show roughly 50% allele frequencies for multiple variants and samples from tumor 5 show 100% or 0% allele frequencies indicating a loss of one allele.

MIDs Label Unique 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 1. The Swift MID is a 9-base random N 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 double-stranded DNA 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.

Increased Specificity with MIDs

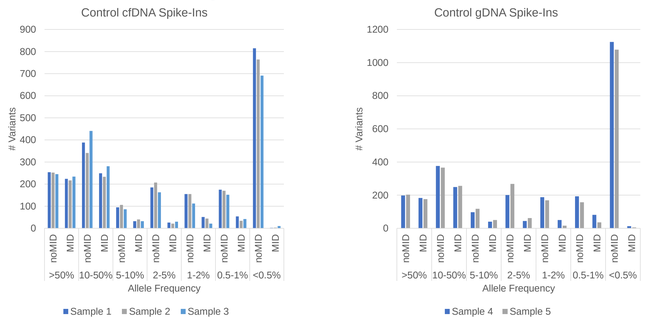
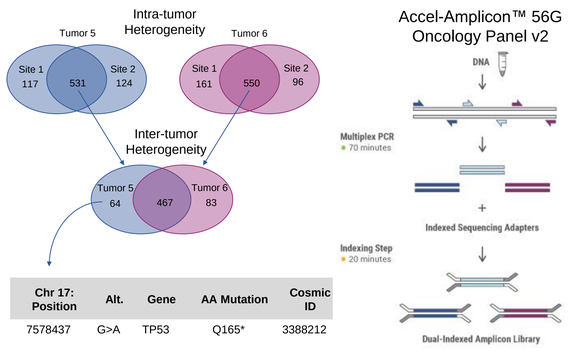


Figure 4. Total variants called at various allele frequencies with or without the use of MIDs are depicted from the spike-in experiments. MIDs only have a subtle effect on the number of variants called at high allele frequencies, but substantially reduce the number of low frequency variants called. This 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.

Tumor Heterogeneity in Ovarian Cancer Samples

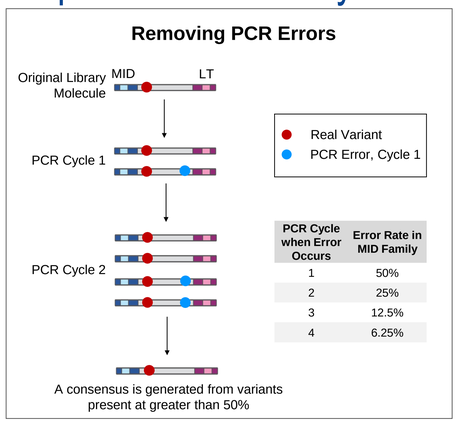


Chr 17: Position	Alt.	Gene	AA Mutation	Cosmic ID
7578437	G>A	TP53	Q165*	3388212

Sample	Library Prep Method	Allele Frequency
Normal	56G	-
Tumor 5.1	Hyb	95.6%
Tumor 5.2	Hyb	97.7%
Tumor 6.1	Hyb	-
Tumor 6.2	Hyb	11.2%
Tumor 1.1	56G	50.3%
Tumor 1.2	56G	78.4%
Tumor 1.3	56G	45.2%
Tumor 3.1	56G	65.2%
Tumor 3.2	56G	59.9%
Tumor 3.3	56G	50.5%

Figure 8. MID families were used to generate a consensus sequence (BMFtools) and variants down to 0.5% were called with LoFreq. The overlap between the two samples from tumor 5 and tumor 6 was determined and a comparison was then made between variants present in both samples from tumor 5 (malignant) and both samples of tumor 6 (benign). Of the variants unique to tumor 5, a pathogenic TP53 mutation (COSM3388212) was identified at close to 100% in samples from tumor 5 and in the cfDNA sample at 11%. Multiple samples were also acquired from two primary tumors from the same individual (tumor 1, left ovary, and tumor 3, right ovary) and analyzed using Swift Biosciences' Accel-Amplicon 56G Oncology Panel v2 that includes full exon coverage of TP53. Variant analysis revealed this TP53 mutation present at > 50% in the primary tumor samples, but not in a normal sample.

Improved Data Analysis with MIDs



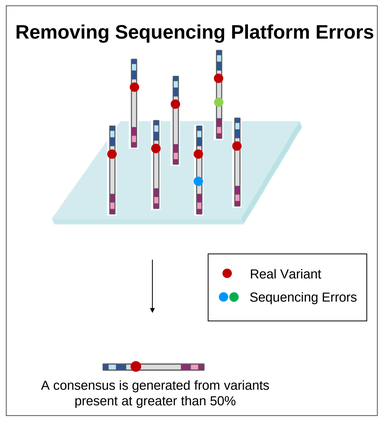
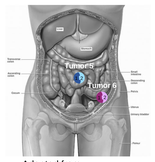


Figure 2. MIDs label individual molecules prior to PCR, facilitating the accurate identification and removal of PCR duplicates during data analysis. Furthermore, molecules containing the same MID can be used to generate a consensus sequence that retains true variants but removes artificial mutations 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.

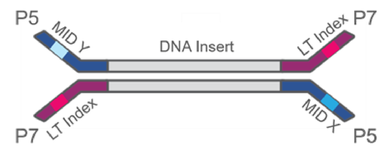
Variant Analysis from a Single Individual

Sample	Description	Final Pathology	Input	Reads	Average Bait Coverage (Pre-deduplication)	Duplication Rate (Without MID/With MID)	On Target
Tumor 5.1	Small bowel implant (site 1)	malignant	50ng	93,536,722	15,634x	89.7% / 84.1%	74.2%
Tumor 5.2	Small bowel implant (site 2)	malignant	50ng	94,802,633	15,962x	89.1% / 83.4%	74.4%
Tumor 6.1	Left pelvic peritoneum (site 1)	benign	50ng	89,990,124	15,096x	86.9% / 81.0%	74.6%
Tumor 6.2	Left pelvic peritoneum (site 2)	benign	50ng	85,759,502	14,387x	88.3% / 82.7%	74.0%
cfDNA	Plasma cfDNA	N/A	20ng	81,387,833	14,135x	93.5% / 88.6%	74.4%



Adapted from http://human-anatomy101.com

Figure 5. Samples were analyzed from a 75-year-old female with stage 3B, grade 3 ovarian carcinosarcoma. Samples from two different sites on two tumors were taken during recurrent surgery performed 9 months after the primary surgery. In addition, plasma cfDNA collected at the time of the recurrent surgery was analyzed. Libraries were prepared using an Accel-NGS 2S Hyb kit with MIDs and enriched for oncology-related genes and hotspots with the IDT xGen Pan-Cancer Panel. Sequencing was performed on an Illumina HiSeq to greater than 14,000x. Deep sequencing maximized the number of PCR duplicates for each molecule used to generate a consensus sequence.



Conclusions

- Labelling unique library molecules with MIDs prior to amplification allows for the removal of sequencing and PCR induced errors during data analysis.
- MIDs improve low frequency variant calling by increasing sensitivity and specificity and are shown here to facilitate detection of known variants present at 1% and 0.5% allele frequencies.
- 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 www.swiftbiosci.com

www.swiftbiosci.com