AMP 2017: #I34

Breaking the NGS Noise Barrier to Accurately Detect Variants below 1% Allele Frequency

Sukhinder Sandhu, Ashley Wood, Bita Lahann, Jordan RoseFigura, Justin Lenhart, Vanessa Kelchner, Laurie Kurihara, Vladimir Makarov, Timothy Harkins

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

Introduction

The use of circulating, cell-free DNA (cfDNA) for early detection and monitoring of disease is rapidly growing. This necessitates accurate variant detection at and below 1% allele frequencies due to a low population of tumor DNA within cfDNA samples. Reliable, low-frequency variant detection by next-generation sequencing (NGS) is challenging due to non-specific background noise from PCR and sequencing errors. We have employed molecular identifiers (MIDs) to uniquely label individual DNA molecules prior to amplification, facilitating the distinction of true variants from PCR and sequencing errors. MID incorporation also results in increased data retention by removing PCR duplicates while preserving fragmentation and sister strand duplicates.

Here, we have applied MIDs to both our Accel-NGS® 2S whole genome library prep followed by hybridization capture with various cancer panels, and to our Accel-Amplicon™ library prep that uses multiplex PCR ideal for small panels that are amenable to deep sequencing. To validate this technology, we performed low frequency spike-in experiments at 1% and 0.5% with genomic DNA, as well as various cfDNA samples. 2S libraries were prepared and hybridization capture was performed with an 800 kb pan-cancer panel. Libraries were sequenced to greater than 8000x coverage and a consensus sequence was generated with BMFtools. All known variants present at 1% and 0.5% allele frequencies were maintained in the resulting data. Further, true variants were preserved while PCR and sequencing errors were removed, demonstrating improved specificity using MIDs. We also validated variant calling below a 1% allele frequency using the Accel-Amplicon EGFR Pathway Panel with MIDs. After validation of our MID technology, libraries were prepared with cfDNA and tumor samples from individuals with ovarian, liver, stomach, and colon cancers. Libraries 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. This study highlights the ability of MID technology to enable variant detection at and below 1% allele frequencies, critical to track known variants and identify novel pathogenic mutations in cfDNA samples.

Improved Data Analysis with MIDs

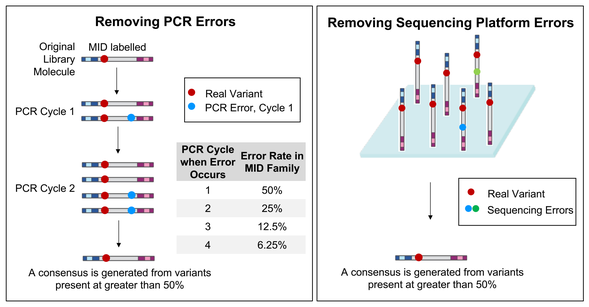


Figure 1. MIDs label individual molecules prior to exponential amplification by PCR facilitating the accurate identification and removal of PCR duplicates. Furthermore, molecules containing the same MID can be used to generate a consensus sequence that retains true variants but removes artificial mutations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

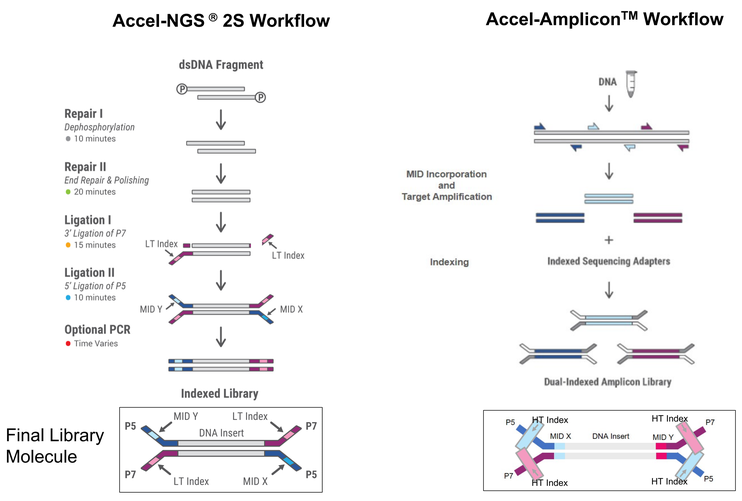


Figure 2. MIDs can be added to both our Accel-NGS 2S whole genome library prep used for hybridization capture and to our Accel-Amplicon library prep that uses multiplex PCR for targeting. Both library preps use the MID to label unique library molecules prior to amplification. Accel-NGS 2S libraries with 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. Each dsDNA substrate receives two independent MID adapters "X" and "Y". The Accel-Amplicon 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 in-line MID/barcode.

Methods

1. Input Considerations for Low Frequency Allele Detection

Input DNA (ng)	Genome Copy#	1% AF Copy#	0.5% AF Copy#	0.1% AF Copy#
1	330	3.3	1.6	< 1.0
5	1650	16.5	8.3	1.7
10	3300	33.0	16.5	3.3
20	6600	66.0	33.0	6.6
40	13200	132.0	66.0	13.2

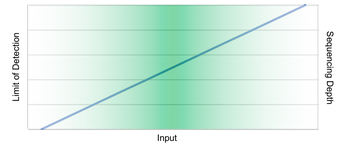
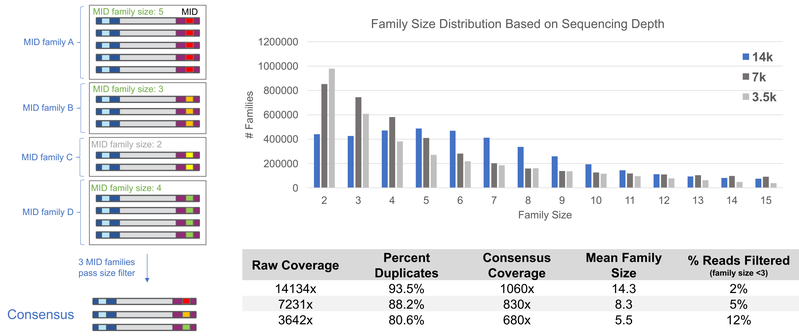


Figure 3. In order to ensure representation of DNA fragments carrying low frequency alleles within a sample, sufficient genome copy number is needed. Sample inputs with a predicted low frequency allele copy number of < 10 may not be present in the DNA sample due to Poisson distribution of DNA fragments in solution. However, higher input genome copy number is proportional to a higher library complexity and impacts the depth of sequencing required. Therefore, a balance between the limit of detection (LOD) and the required sequencing depth must be considered to determine the appropriate input for an experiment. Recommended inputs for specified LOD allele frequency are highlighted in green.

2. Sequencing Depth Considerations for Complexity



Raw Coverage	Percent Duplicates	Consensus Coverage	Mean Family Size	% Reads Filtered (family size <3)
14134x	93.5%	1060x	14.3	2%
7231x	88.2%	830x	8.3	5%
3642x	80.6%	680x	5.5	12%

Figure 4. Identification of low frequency variants requires that at least three PCR duplicates (Family Size ≥ 3)are sequenced per MID family to enable consensus sequence analysis since only alleles present in the majority of PCR duplicates are retained in the consensus. To achieve this for the majority of molecules, this requires sequencing DNA libraries to near saturation (i.e., every unique molecule from the library is sequenced more than once). Here we show how sequencing to 14000x, 7000x, and 3500x influences MID family size for a library prepared from 20ng cfDNA and enriched using the 800kb IDT xGen® Pan-Cancer Panel.

3. Bioinformatics Workflow

Make MID fastq

2S Hyb Analysis: MID/UMI is the 3rd read from PE sequencing, and can be obtained after running bcl2fastq (Illumina) from all sequencers.

Amplicon Analysis: MID/UMI are the 1st five bases of each Read1 & 2, and can be obtained by trimming those using Trimmomatic, Cutadapt, fastx-trimmer, etc.

Fgbio (Fulcrum Genomics)

- Align with bwa
- Append MID/UMI from the MID fastq file generated above: *fgbio AnnotateBamWumi*
- Group reads by MID/UMI, includes error correction model: *fgbio GroupReadsByUmi*
- Generate consensus sequence: *fgbio GroupReadsByUmi*

BMFtools (ARUP Laboratories)

- Combine reads which share MID tag at fastq stage (secondary for Accel 2S, & Amplicon): *bmftools collapse secondary*
- Align collapsed reads with; mark tags to rescue reads after alignment, sort *bmftools mark | bmftools sort*
- Rescue reads post-alignment into pre-align MID groups using positional information *bmftools rescue*

Variant calling with Lofreq, Vardict, or variant caller of choice

Fgbio Output
- FilterBam (fewer options)
- Compatible with MID/UMI agnostic tools

BMFtools Output
- Multiple summary metrics: family stats, coverage, filter bam options, variant caller (stack) and Vet tools
- Compatible with MID/UMI agnostic tools

Results

1. Increased Data Retention with MIDs

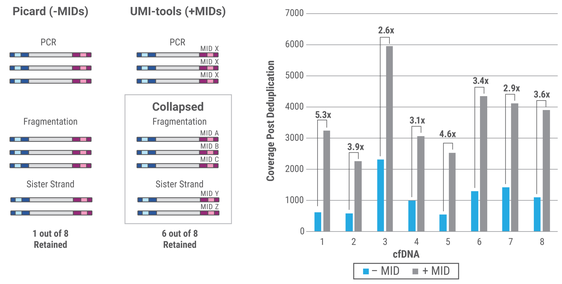


Figure 6. We evaluated the effect of MIDs on data retention after de-duplication. 20ng cfDNA from variety of cancer types was used. De-duplication was performed with either standard Picard tools (- MID) or UMI-tools from Fulcrum Genomics (+ MID). MIDs allow for accurate identification and removal of PCR duplicates while maintaining sister strand duplicates and fragment duplicates. De-duplication using MIDs showed an increase in coverage for all samples analyzed.

2. Identification of Variants Down to 0.5%

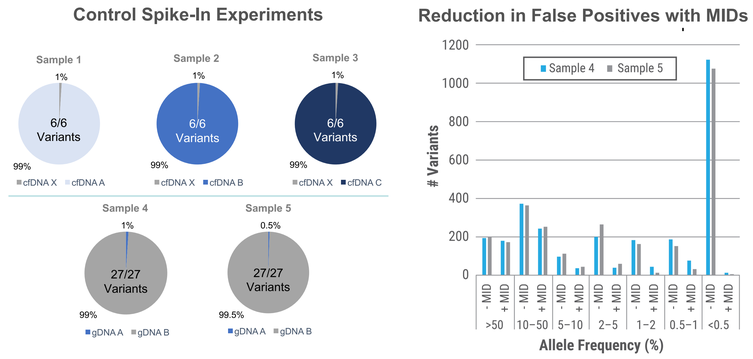


Figure 7. (Left) cfDNA was extracted from blood of four individuals with unique genetic backgrounds and gDNA samples from different genetic backgrounds were obtained (Coriell Institute). Sample spike-ins were performed at 1% or 0.5% frequency into 10 ng cfDNA or 100 ng gDNA. Libraries were prepared with Swift's Accel-NGS 2S Hyb Kit with MIDs, enriched with the IDT xGen Pan-Cancer Panel and sequenced on an Illumina HiSeq® to a minimum of 8000x coverage. Data from 2S Hyb spike-in experiments were analyzed with BMF tools 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. (Right) Total variants called at various allele frequencies with or without the use of MIDs are depicted from the gDNA 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.

3. EGFR MID Panel Performance

Variant Calling with the EGFR-MID Amplicon Panel

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

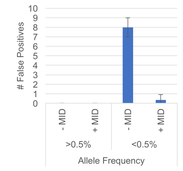


Figure 8. The EGFR MID panel contains 17 amplicons, 107-149bp in size, and shows >95% coverage uniformity (>0.2x mean) and >95% on target reads. Horizon Diagnostics Quantitative Multiplex DNA Standard (HD701) was spiked into Coriell DNA (NA12878) at 50% to obtain expected variants at allele frequencies 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. 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

- We've validated the input DNA quantity and depth of sequencing required to maximize data retention and the LOD when performing MID analysis. We recommend using the minimal input quantity for the desired LOD in order to minimize sequencing cost.
- We have successfully applied Fgbio and BMFtools for MID analysis. Both produce error free consensus bam, compatible with most MID agnostic tools. BMFtools provides more exploratory options and a plethora of summary metrics and filtering options.
- Incorporating MIDs in NGS library preparation increases data retention during de-duplication and 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