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 nonspecific background noise from PCR and sequencing errors. We have employed molecular identifiers (MIDs) to uniquely label individual DNA molecules prior to amplification, facilitating the detection of low variants from PCR and sequencing errors. MID incorporation also results in increased data retention by removing PCR duplicates while preserving fragmentation and strand orientation.

Here we have applied MIDs to our previously published whole genome library prep followed by hybridization capture with capture panel probes and then the Accel-AMPION™ library prep assay that uses multiple MID PCR for small panel that are essentially 100% hybridization efficient with any MID group and sparsely hybridized with any MID group with a+2% error, with various MID libraries. As well as various MID/UMI samples. 24 libraries were hybridization and preparation was performed with 800 ng cfDNA as a control at 10 ng cfDNA as a control. Libraries were sequenced to greater than 1000X coverage and a consensus was generated with bmftools. All known variants present at 1% and 0.1% allele frequencies were maintained in the resulting data. Further, low variants were preserved while PCR and sequencing errors were removed, demonstrating modified specificity using MIDs. We also evaluated variant calling below 1% allele frequency using the Accel-AMPION® EGFRI Pathway Panel with MIDs. After validation of our MID families were prepared with MIDs and tumor sample libraries contain with ovarian, liver, stomach, and colon libraries. Libraries were sequenced to a minimum of 15,000X coverage, and we determined rare variants through de-duplication with and without the use of MIDs. We observed a significant increase in data retention that is on the order of a factor of five to ten times higher than the same data set with or without MIDs present in a corresponding tumor sample when analyzed. This study highlights the ability of MID technology to enable variant detection at and below 1% allele frequencies, to track known variants and identify novel pathogenic mutations in cfDNA samples.

Methods

1. Input Considerations for Low Frequency Allele Detection

- **Input DNA**: ng
- **Genome Coverage**: 15X
- **1% AF**: 
- **0.1% AF**:

![Figure 1](https://www.swiftbiosci.com/image)

The purpose of the study is to understand the performance of MID incorporated amplicon sequencing on the Accel-AMPION® EGFRI Pathway Panel with MIDs.

2. Sequencing Depth Considerations for Complexity

- **Raw Coverage**:
- **Percent Duplicates**:
- **Consensus Coverage**:
- **Mean Family Size**:
- **% Reads Filtered**:

![Figure 2](https://www.swiftbiosci.com/image)

- **Consensus**:
- **Coverage**:
- **Size**:
- **Filtering**:

The purpose of the study is to understand the performance of MID incorporated amplicon sequencing on the Accel-AMPION® EGFRI Pathway Panel with MIDs.

3. Bioinformatics Workflow

- **MID family A**:
- **MID family B**:
- **MID family C**:
- **MID family D**:

![Figure 3](https://www.swiftbiosci.com/image)

**Figure 3.** MIDs can be added to both our Accel-NGS 25 workflow whole genome library prep for hybridization capture and our Accel-AMPION® library prep assay that uses multiplex MID PCR for targeting. Both library prep use the MID to label unique library molecules in a multiplex fashion. The Accel-NGS 25 workflow performs hybridization with a 2:2:1 MID Incorporation step. PCR samples are sequenced at the start of both series and 10 ng cfDNA. Each original DNA molecule receives an unique MID tag, yielding 10-12 reads in the MID sequencing.

Results

1. Increased Data Retention with MIDs

![Figure 4](https://www.swiftbiosci.com/image)

- **Call rate**:
- **Variants detected**:
- **False positive rate**:

**Figure 4.** We evaluated the effect of MIDs on increasing data retention by removing PCR duplicates. 20 ng cfDNA from a variety of cancer types was used. De-duplication was performed with several standard Prime tools (-MID) or MID tools from Fidicin Genomics (-MID) with the MID deletion and removal of PCR duplicates. De-duplication using MID tools resulted in greater coverage for all samples analyzed.

2. Identification of Variants Down to 0.5%

![Figure 5](https://www.swiftbiosci.com/image)

**Figure 5.** (Left) cfDNA was extracted from blood of four individuals with unique genetic backgrounds and given samples with minimal genetic differences were obtained. (Center) Twelve different tumor amplicons were performed in order to sequence the seven cancer types to assess the MID incorporated amplicon sequencing. The MID incorporated amplicon sequencing was performed on a T790M (T790M) and 2000s influences MID family size for a library prepared from 20 ng cfDNA and enriched using the 50000-1 G4 fastCap Panel-Cancer Panel.

3. EGFR MID Panel Performance

![Figure 6](https://www.swiftbiosci.com/image)

**Figure 6.** The EGFR MID panel consists of 17 amplicons, 15-17 ng cfDNA in size, and shows >98% coverage uniformly (±2% mean) and >98% on target reads. In-house Diagnostic Quantitative Multiplex DNA Test (DHQTM) was validated using cfDNA DNA Quantitation. The MID incorporated amplicon sequencing was performed on a T790M (T790M) to a minimum of 1000X coverage. Data from 20 ng cfDNA experiments were analyzed with MID tools and data were analyzed for-homologous DNA present in the sample only. All known variants were present in all MID samples and 2707 known variants were present in both 1% and 0.5% cfDNA samples. (Right) Total variants called at various allele frequencies with and without the use of MIDs are depicted from the MID split-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 variant calls are highly enriched for true variants and the variant's representative fragments.

Conclusion

- **We validated the input DNA quantity and depth of sequencing required to maximize data retention and the LOQ when performing analysis. We recommend using the minimal input quantity for the desired LOQ in order to maintain sequencing efficiency.**

- **We have successfully applied Fgbio and BMTools for MID analysis. Both provide error-free consensus bam, compatible with most MID agnostic tools. BMTools provides more exploitative 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 specifically at low allele frequencies. Here we are able to detect known variants at less than 1% allele frequencies using hybridization capture and population wide targeted sequencing.**