Improved Data Analysis with MIDs

MIDs Label Unique Library Molecules

Identification of Variants Down to 0.5%

Increased Specificity with MIDs

Improved Data Analysis with MIDs

Concentration of Variants A©

Variant Analysis from a Single Individual

Loss of Heterozygosity in Ovarian Cancer Samples

Tumor Heterogeneity in Ovarian Cancer Samples

Results

Figure 1. The Swift MD is a 9 base random N sequence that uniquely labels individual library molecules prior to amplification. Acast-NGS 25 libraries with MIDs are constructed using Illumina® compatible adapters with a strand-specific MID in the 5’ index position and a sample index in the index 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 25 MID libraries attain close to 100% representation of MID combinations with high copy uniformity (percentages 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.

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.

Figure 3. cfDNA was extracted from blood of four individuals with unique genetic backgrounds and cross-spiked at a frequency of 1%. In addition, control gDNA samples from different genetic backgrounds were cross-spiked at a frequency of 0.5% and 1%. Libraries were prepared with the Swift 25 Hpy kit with MIDs, enriched with the IGT iCancer Pan-Cancer Panel that covers 408 target cancer genes, and sequenced on an Illumina HiSeq® to a minimum 2500x coverage. A consensus sequence was generated for each MID family (MIDN-MIDN) showing high sequence uniformity. Samples were then analyzed for homozygous SNPs present in the spike-in sample only. 618 and 27/27 homozygous SNPs were found in cfDNA and gDNA, respectively, demonstrating the power of MIDs for low frequency variant calling.

Figure 4. Total variants called at various allele frequencies with and 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.

Figure 5. Samples were analyzed from a 75-year-old female with stage 3B grade 3 ovarian carcinoma. Samples from two different sites on two tumors were taken during surgery. In addition, plasma cfDNA collected at the time of the tumor biopsy was analyzed. Libraries were prepared using an Acast-NGS 25 Hpy kit with MIDs and enriched for oncology-related genes and co-located with the iCancer Pan-Cancer Panel. Sequencing was performed on an Illumina HiSeq® to greater than 14,000x depth sequencing. Maximum pairwise analysis maximized the number of PCR duplicates for each molecule used to generate a consensus sequence.

Conclusion

- Labeling 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.

The Use of Molecular Identifiers (MIDs) for Improved Low Frequency Mutation Detection in Ovarian Cancer

Candia Brown1, Ashley Wood1, Sukhinder Sandhu1, Matthew Dashkoff2, Timothy Harkins1, Olga Camacho-Vanegas2, Peter Dottino2, Laurie Kurihara1, John Martingelli2, Vladimir Makarov1

Swift Biosciences, 58 Parkland Plaza, Suite 100, Ann Arbor, MI 48103; Tel: 734.330.2586

1Swift Biosciences, 58 Parkland Plaza, Suite 100, Ann Arbor, MI 48103, Tel: 734.330.2586
2Oak 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

Improving the sensitivity of next generation sequencing (NGS) technologies to detect low frequency variants involves removal of PCR duplicates to achieve accurate coverage. By uniquely labeling every DNA molecule within an NGS library, more confident detection of low frequency mutations and other genetic variants is possible. Molecular ID (MID) labeling of unique library molecules enables detection and removal of PCR duplicates while preserving fragmentation and complementary strand duplicates. Prior to removal, the MID can also be used to generate a consensus sequence from PCR duplicates, reducing sequencing and PCR-related errors. Here MID technology was combined with Swift Biosciences’ Acast-NGS® 25 Hpy DNA Library Kit to evaluate the effect on data analysis. Libraries were prepared with DNA from a series of tumors from a stage 3B ovarian carcinoma patient and circulating cfDNA from blood plasma collected at the time of surgery. We determined data retention after de-duplication using either standard Picard tools without using MIDs or UMI-tools (Fulcrum Genomics) that utilize MIDs. We observed an increase in data retention for both Concatenated cfDNA and cfDNA-libraries that lead to a 1.7 to 2.9 fold increase in coverage using UMI-tools over Picard de-duplication under the conditions tested and found this effect was maximized by lower DNA input quantities and higher depth of sequencing. These libraries were further evaluated for variant calling, using oncology-related genes only. We identified low frequency variants (as low as 0.5%) present across all tumor samples and within cfDNA, as well as variants unique to each tumor sample.

Improving the sensitivity of next generation sequencing (NGS) technologies to detect low frequency variants involves removal of PCR duplicates to achieve accurate coverage. By uniquely labeling every DNA molecule within an NGS library, more confident detection of low frequency mutations and other genetic variants is possible. Molecular ID (MID) labeling of unique library molecules enables detection and removal of PCR duplicates while preserving fragmentation and complementary strand duplicates. Prior to removal, the MID can also be used to generate a consensus sequence from PCR duplicates, reducing sequencing and PCR-related errors. Here MID technology was combined with Swift Biosciences’ Acast-NGS® 25 Hpy DNA Library Kit to evaluate the effect on data analysis. Libraries were prepared with DNA from a series of tumors from a stage 3B ovarian carcinoma patient and circulating cfDNA from blood plasma collected at the time of surgery. We determined data retention after de-duplication using either standard Picard tools without using MIDs or UMI-tools (Fulcrum Genomics) that utilize MIDs. We observed an increase in data retention for both Concatenated cfDNA and cfDNA-libraries that lead to a 1.7 to 2.9 fold increase in coverage using UMI-tools over Picard de-duplication under the conditions tested and found this effect was maximized by lower DNA input quantities and higher depth of sequencing. These libraries were further evaluated for variant calling, using oncology-related genes only. We identified low frequency variants (as low as 0.5%) present across all tumor samples and within cfDNA, as well as variants unique to each tumor sample.