CleanPlex® UMI: Detecting Rare Variants using a Targeted Amplicon Sequencing Approach with A Novel Double-Strand Molecular Barcoding Scheme

Lucie Lee, Yang Lily Liu, Li Jacey Zhang, Jeffery Liu, Lifeng Lin, Guoying Liu, Zhitong Liu | Paragon Genomics, Inc, 3521 Investment Blvd Suite 1, Hayward, CA 94545, USA

Abstract

Utilizing molecular barcodes, or unique molecular indices (UMI), with next generation sequencing (NGS) for the detection of somatic variants with ≤ 0.1% allele frequency has attracted increasing interest in the field of early cancer detection, cancer treatment monitoring, drug resistance screening, and liquid biopsy applications. We present CleanPlex® UMI - a patent-pending multiplex PCR-based ultra-sensitive molecular barcoding technology that significantly reduces false positive calls at low allele frequencies. The NGS library preparation workflow consists of a multiplex PCR step that uniquely barcodes target sequences, a resolving step that removes redundant barcodes, and a final PCR step that adds sequencing adapters to the library. The sequenced reads can be grouped by barcodes and traced back to the sense and anti-sense strand of the original DNA fragment. This combined barcoded library preparation and variant calling algorithm drastically reduces the number of false positive calls caused by PCR and sequencing errors. The described technology utilizes a three-step workflow and 3 hours to make molecular-barcoded NGS libraries, while demonstrating high sensitivity in detecting alleles with 0.1% frequencies.

Workflow

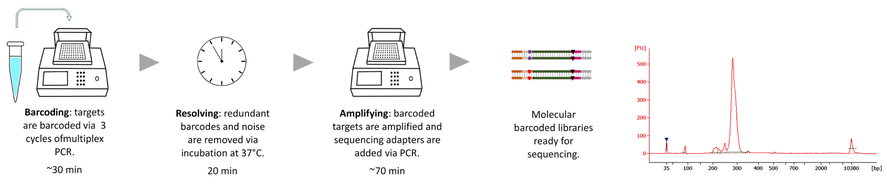

Figure 1: Left panel. The workflow includes adding molecular barcodes onto both sides of targets by 3 cycles of multiplex PCR, removing redundant barcodes and adding sequencing adapters and sample indexes via PCR amplification. **Right panel.** An example library generated by the CleanPlex® UMI technology.

Mechanism

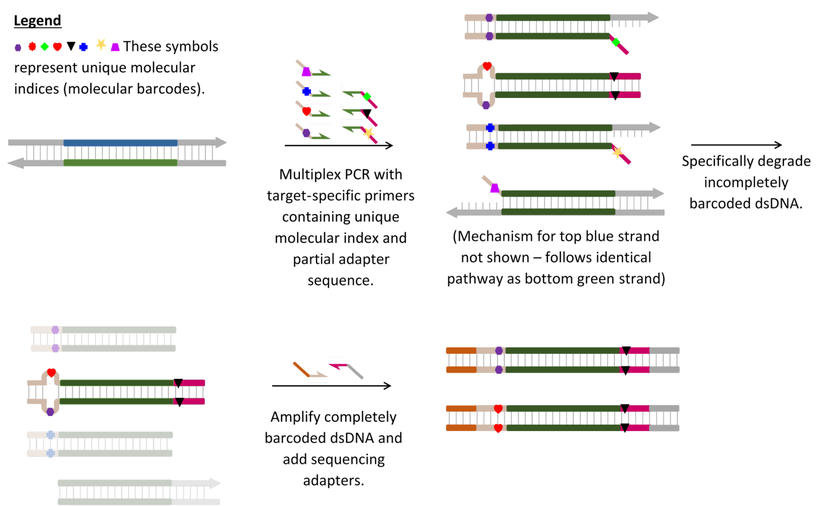

Figure 2: With a pool of primers containing unique molecular indices, 3 cycles of multiplex PCR yields two barcoded dsDNA families per ancestor DNA strand. Incompletely barcoded products are selectively degraded in the resolving step. Sequencing adapters are added to dsDNA families and amplified.

Data Analysis Scheme

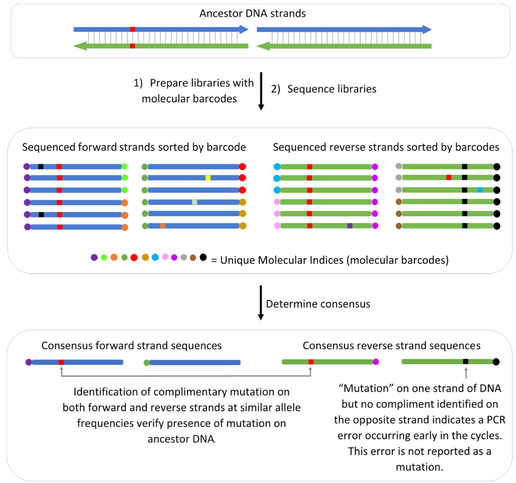

Figure 3: Diagram demonstrating how consensus sequences for barcode families are built, and how variant call concordance is determined using the consensus forward and reverse strands.

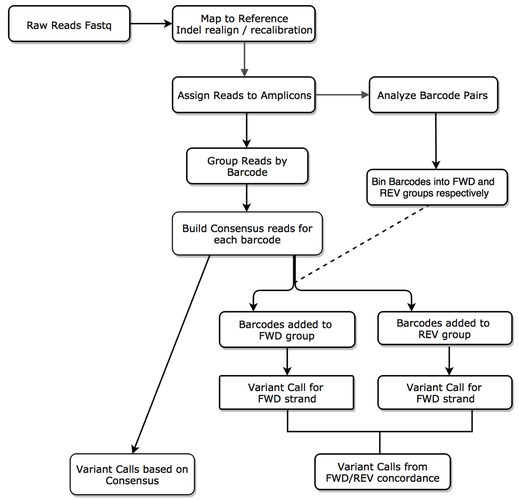

Figure 4: Flow chat demonstrating the bioinformatic algorithm for single stand variant calling based only on consensus, and double stranded workflow based on forward and reverse concordance..

Panel Design

We designed two development panels with 40 and 53 amplicons targeting some known somatic mutations related to lung cancer. Amplicon length ranges from 70-100bp and can be used with cfDNA or gDNA samples. Each forward and reverse PCR primers contains 12 or 16 random bases as molecular barcodes for dual molecular barcoding. After removing redundant barcodes in the resolving step, the libraries were amplified with primers containing sequencing adapters and sample indexes on both sides. Sequencing depth was based on the DNA input used in making the library, with 7500 and 8000 reads per amplicon were used for every nanogram of cfDNA and gDNA, respectively.

Variant Call

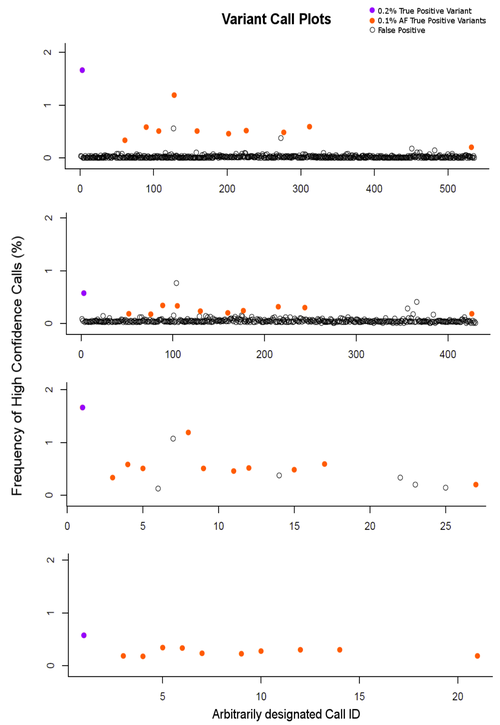

Figure 5: 40 ng of genomic DNA (gDNA) mix of 0.2% NA12878 in NA18507 was used for the plots above. A total of 11 reference mutations are expected: 10 mutations (orange) at 0.1% allele frequency (AF) and 1 mutation (magenta) at 0.2% AF. The four plots show a progression of single stranded to double stranded variant calling, and from unfiltered to filtered results. The concordance to expected allele frequency percentage improves with double stranded variant calling.

HGEs Recovered

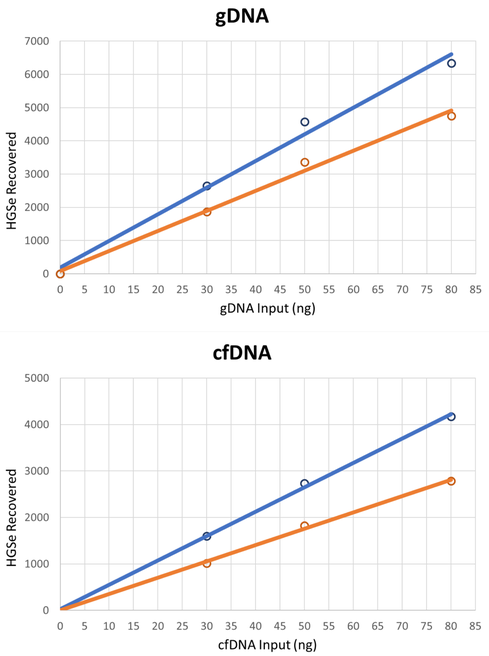

Figure 6: The number of HGEs (haploid genomes) recovered from variant calling by single and double stranded consensus for both gDNA and cfDNA. It required 35ng of gDNA (0.2% NA12878 in NA18507) or 55ng of cfDNA (Horizon HD780) to recover 3000 copies of HGEs (average 3 positives at 0.1% AF) by single stranded consensus; 50ng of gDNA or 80ng of cfDNA to recover 3000 copies of HGEs (average 3 positives at 0.1% AF) by double stranded consensus.

Conclusions

- CleanPlex® UMI is a multiplex PCR-based technology for molecular barcoding, with redundant barcodes removal and double stranded consensus in variant calling.
- Double strand consensus allows accurate calling of low AF% and dramatic separation of the background from the true variants, which enables removing of noise by filtering.
- Library generation involves a simple 3-hour workflow that incorporates CleanPlex® background removing technology, generating a clean library that requires less sequencing depth.
- Custom panel design utilizing this technology is available.

Paragon Genomics, Inc. | www.paragongenomics.com | contact@paragongenomics.com | 1-510-363-9918