illumina | ARCHER®

Technical Note

The Use of Molecular Barcodes in Anchored Multiplex PCR

Sample Index vs. Molecular Barcode

Sample barcoding, also called sample indexing, is a common approach to labeling samples for multiplex sequencing and analysis. All nucleic acids in a sample are labeled with the same sequence tag, and the resulting library is pooled with other libraries and sequenced in parallel in a single run. Then, during analysis, the sample-specific indexes enable the software to separate the multiplexed sequence data in sample-specific data sets *(Figure 1A)*.

Molecular barcoding differs from sample indexing, in that each molecule in a sample is labeled with a unique sequence prior to PCR amplification. With each nucleic acid in the starting material tagged with a unique molecular barcode (MBC), sequence analysis software can filter out duplicate reads and PCR errors to report unique reads *(Figure 1B and 2)*.

Anchored Multiplex PCR (AMP™)

AMP technology takes advantage of both tagging approaches for sample multiplexing and accurate mutation calling by ligating an adapter molecule to the starting cDNA or DNA fragments prior to PCR amplification. This adapter contains a sample-specific index of pre-defined sequence and a random 8-mer molecular barcode. After library preparation and sequencing, the Archer Analysis bioinformatics software sifts through the sequencing data and identifies each read's originating sample (via the sample-specific index) and the unique nucleic acid starting molecule (via the MBC).

Advantages of Molecular Barcodes

There are two major advantages to using molecular barcodes (MBCs) for targeted sequencing.

Molecular barcoding exceeds the limited number of unique reads that are identified by alignment-based de-duplication.

With highly complex libraries, more unique reads can be obtained with MBC-based de-duplication than is possible by removing PCR duplicates based on alignment coordinates. With alignment based de-duplica-

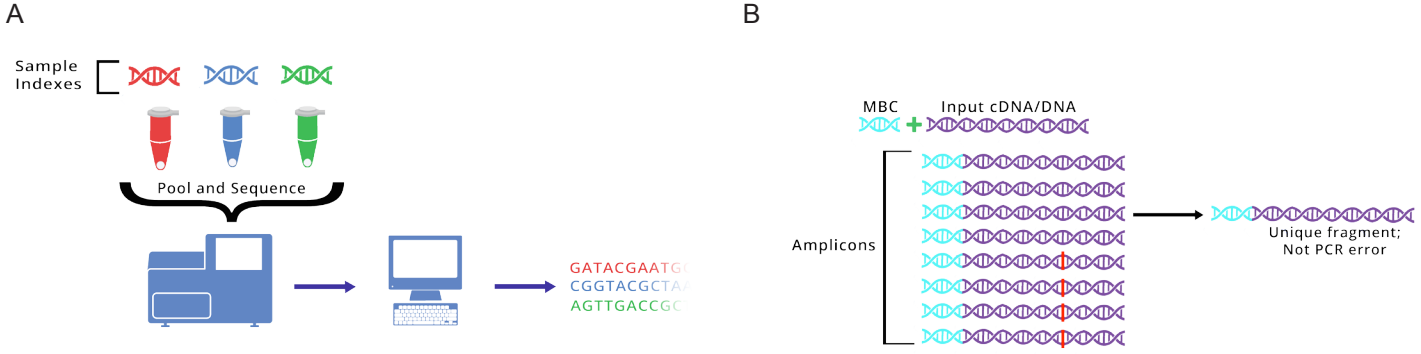

Figure 1. Sample indexes differ from molecular barcodes (MBCs). *A) Each sample is labeled with a unique sample index (red, blue, green). The sample libraries are then sequenced, and analysis software bins the data based on the index in each sample. B) Each cDNA or DNA molecule in a sample is labeled with a unique MBC. After sequencing, software generates consensus sequences based on MBCs, which helps remove PCR errors (red lines) or amplification bias.*

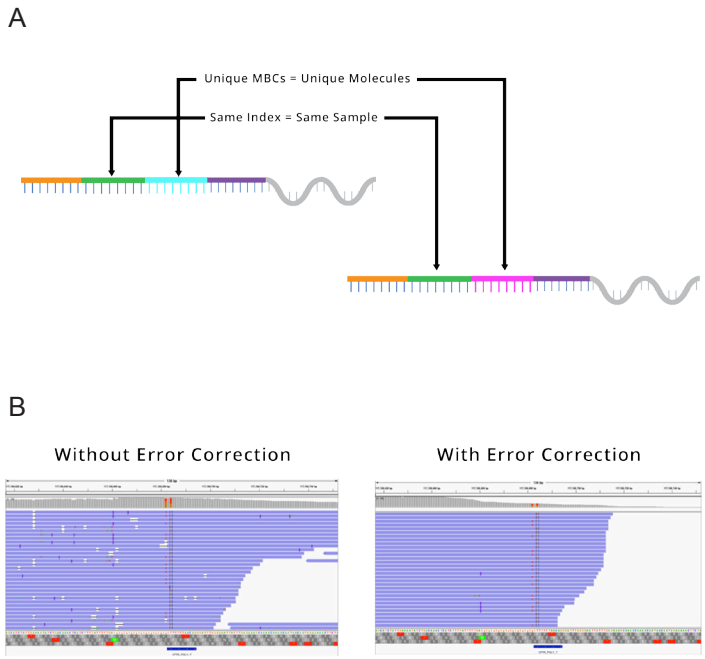

Figure 2. Archer MBCs and downstream error correction. *A) Archer MBC Adapters contain both a sample-specific index sequence and a random 8-mer MBC for both sample indexing and unique molecule identification, respectively. B) Screenshots showing sequence reads with and without MBC error correction.*

tion, the number of unique molecules counted for a given panel target is limited by both the fragment length and the constraints of the sequencing platform. In addition, reads with the same alignment coordinates may or may not be PCR duplicates. For example, a target primer pair meant to capture information at the most 5' end of a wild-type transcript will produce multiple reads that start at the transciption site. Alignment-based de-duplication of these reads would consider such reads PCR duplicates even if they were, in fact, unique transcripts.

Molecular barcodes bypass these limitations, as these tags are situated within the adapter such that the initial base calls of the read are that of the MBC, allowing differentiation of unique reads and binning for more accurate downstream analysis.

AMP reduces the time spent de-duplicating while providing high quality reads for downstream analysis.

In Archer Analysis, a hybrid key is generated for each read, consisting of the 8-mer molecular barcode and 10bp from the beginning of the R1 sequence, which expands the sequence space beyond the MBC to enhance binning specificity. Reads with the same key are binned together using an efficient clustering algorithm, where up to two mismatches between keys are tolerated. Reads with the same molecular barcode are binned together as PCR duplicates, so the read with the highest quality base calling is picked as the representative for the group. This method of de-duplication is faster than those based on alignment only, as it relies on simple string comparisons rather than more complex alignment algorithms.

Binning based on MBCs ensures that most real PCR duplicates containing low-level sequencing errors are binned together. After the MBC bins are created, optional error correction may be applied to bins of sufficient depth. For less populated bins, the read with the highest median base call qualities are picked as the representative for the bin for downstream processsing.

The result of deduplication is a population of high-quality, unique reads that are used to characterize the sample and calculate standard library metrics.

Molecular Barcode-enabled CNV calling

Uniquely tagging input fragments with molecular barcodes enables the counting of molecules post-sequencing. The MBC allows precise counting of the unique fragments that originate from the loci targeted by the panel. The number of unique MBCs associated with each target region is used to deduce copy number as shown below in Figure 3.

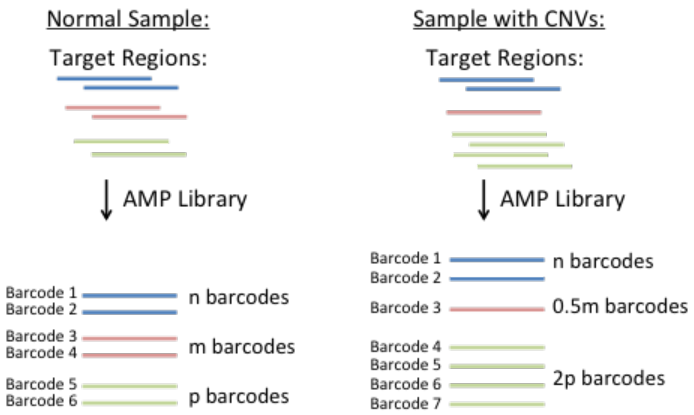

Figure 3. *Unique MBCs enable copy number counting*

Correlation to aCGH and qPCR

In a validation of molecular barcode-enabled CNV detection, a 25-gene panel was tested on a subset of NCI-60 cell lines by comparing our copy number measurements to those determined by both aCGH and qPCR. Results from both orthogonal methods strong-

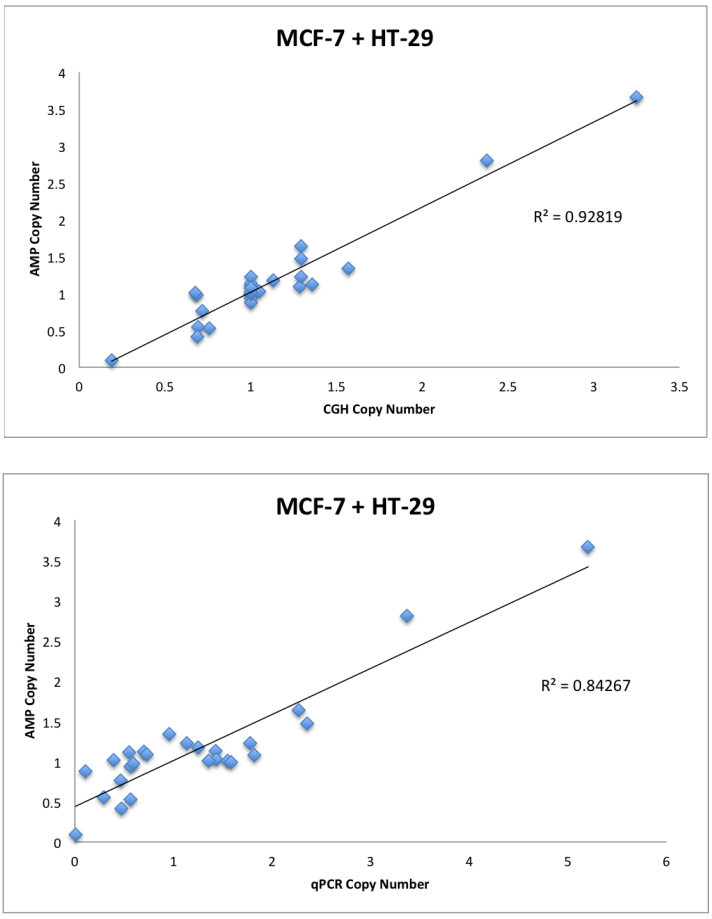

Figure 4. *Plots of CNV determined by aCGH (top) and qPCR (bottom) vs CNV determined by AMP of NCI-60 cell lines HT-29.*

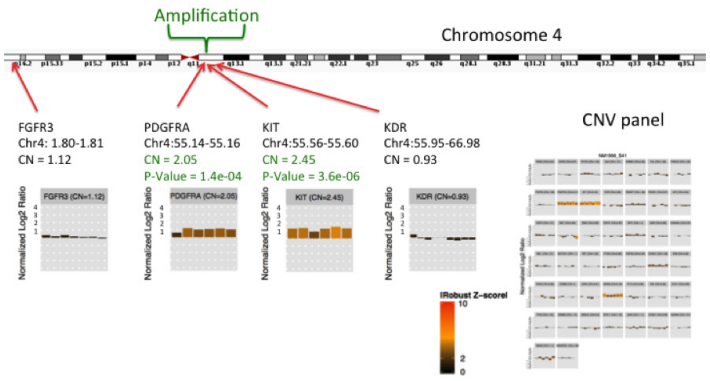

Figure 5. *Amplifications of PDGFA and KIT genes (CN = 2.05 and CN = 2.45 respectively) flanked by copy number neutral genes FGFR3 and KDR on chromosome 4 (CN = 1.12 and 0.93 respectively). 50 ng of DNA extracted from an FFPE sample was sequenced at 2M reads.*

ly correlated with data from our NGS-based method. Hundreds of samples were multiplexed on a single MiSeq® run and detected CNVs, both amplifications and deletions, of 2X magnitudes (and often lower) at extremely high confidence. Correlations between AMP CNV detection aCGH and qPCR are shown in Figure 4.

Sensitive CNV detection in VariantPlex Panels

Concurrent CNV and somatic variant detection is enabled in multi-gene Archer VariantPlex panels - like the Solid Tumor panel, which visually reports copy number on 43 relevant genes. Archer Analysis reports out copy number on 43 genes in the Archer VariantPlex Solid Tumor Panel v1. During an FFPE sample screen with this panel, a focal amplification of platelet-derived growth factor receptor alpha (PDGFRA) and c-KIT was detected at extremely high confidence (Figure 5).

Conclusion

We have demonstrated that molecular barcodes utilized in anchored multiplex PCR power the quantitative nature of Archer assays. Unique molecule tracking enables downstream deduplication, error correction, and sensitive and highly-multiplexed CNV detection.

A special thanks to Dr. Milhan Telatar for help in validating AMP CNV to aCGH and qPCR.

For more information, visit www.archerdx.com/fusionplex-assays/mbc-adapters

Limitations of use:

For Research Use Only. Not for use in diagnostic procedures.