White Paper

WHY UMI?

Reasons to Consider a UMI-based Approach in Next-Generation Sequencing

Authors: Madison Tyler, Jessica Runyon

August 2021

A BRUKER COMPANY
CANOPY BIOSCIENCES®

HIGHLIGHTS

- Next-generation sequencing has revolutionized the life sciences and will likely continue to dominate biological research well into the future
- Unique molecular identifiers (UMIs) function as molecular barcodes that are added to nucleotide fragments prior to library preparation
- UMIs are used in NGS technologies to eliminate false-positive reads that are often introduced during library preparation and increase sensitivity of true variant detection
- RareSeq™ is a novel error-corrected NGS method that uses UMIs to detect ultra-rare allele variants in hematologic malignancies

INTRODUCTION

Next-generation sequencing (NGS) has revolutionized the life sciences by allowing researchers to study biological systems at an unprecedented level. Complex research questions that require a depth or breadth of genomic information beyond what traditional methods can offer, can now be addressed. With NGS, researchers can sequence whole genomes and target regions (Bewicke-Copley et al., 2019), analyze epigenetic factors (Sarda & Hannenhalli, 2014), and conduct gene expression analysis with RNA sequencing (Kukurba & Montgomery, 2015). Due to the ultra-high throughput, speed, and low costs, NGS technologies will likely continue to dominate biological research well into the future.

Library preparation is a critical first step in NGS and usually involves PCR amplification of the targets of interest. Although necessary, PCR amplification is a principal source of bias when generating sequencing data, as certain transcripts may be under or overrepresented in the final sequencing library. When the starting material is limited, contains long fragments, or has a high GC content, processing errors may be propagated (van Dijk et al., 2014). Such processing errors skew transcript abundance measurements, which poses a problem during read count analysis.

Furthermore, transcripts of biological impact and interest may be present in very low quantities. Although many genetic loci have been identified to date, a large proportion of genetic variation that predisposes an individual to disease have not yet been identified (Bomba et al., 2017). NGS technologies have rapidly evolved to meet the growing demands of scientists, who need methods that can confidently detect a single mutation in a population of normal genes. However, these methods are extremely specialized, using reagents that target selected disease-specific variants.

Unique molecular identifiers (UMIs) pragmatically solve these problems through a modification to conventional NGS. UMIs have emerged as valuable tools to reduce false-positive reads that are often introduced during the library preparation and sequencing steps and increase the sensitivity of variant allele detection.

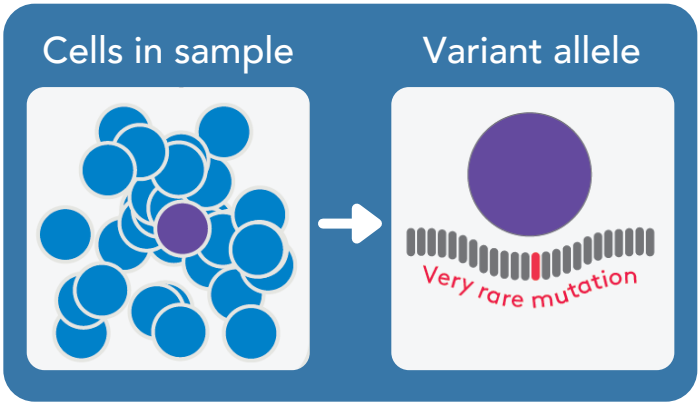

Figure 1. Left: NGS can be used to sequence entire cellular genomes or constrained to specific areas of interest small numbers of individual genes. Right: A key goal of a UMI-based approach in NGS is to identify very rare mutations and distinguish them from processing errors that arise during library preparation and sequencing.

For research use only. Not for use in diagnostic procedures.

WHAT ARE UMIS?

A unique molecular identifier (UMI) is a short molecular barcode used to uniquely tag DNA or RNA fragments. UMIs have a random sequence composition that is added prior to library preparation to ensure each fragment with a UMI is unique. After data is collected, bioinformatics software groups nucleotide fragments into "read families" according to UMI. In this way, biologically significant variant alleles present in the original sample can be distinguished from errors introduced during processing, as shown in Figure 2. This enables the software to provide an accurate quantitative readout of transcript abundance by eliminating false-positive reads.

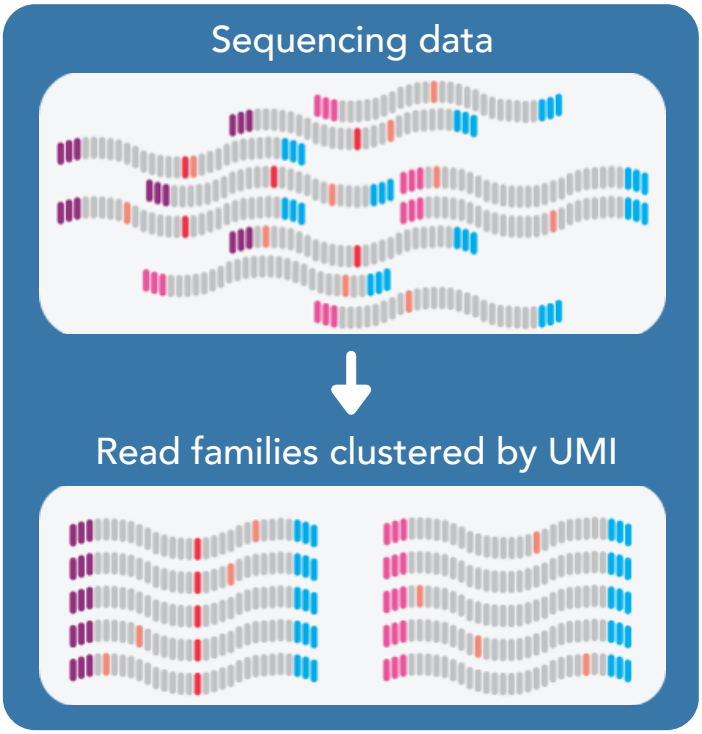

Figure 2. Top: Pool of sequencing data containing false-positive reads or processing errors (orange lines) and true variants (red lines). Bottom: Sequencing data clustered into read families based on UMIs (represented by purple, blue, and pink tags on the ends of nucleotide fragments).

There are two key reasons researchers employ a UMI-based approach in NGS research. First, UMIs enable deduplication analysis which leads to more accurate read abundance measurements. Second, UMIs increase read sensitivity and lower the limit of detection to identify rare and ultra-rare variants.

UMIS ENABLE DEDUPLICATION ANALYSIS FOR MORE ACCURATE READ ABUNDANCE MEASUREMENTS

A variety of sequencing applications use a read counting approach to estimate the abundance of a particular nucleotide fragment across biological conditions or different cell types. Yet, PCR amplification can lead to bias that can propagate to quantification measurements. To address this, UMIs enable deduplication analysis for more accurate read quantification results. In deduplication analysis, reads that align to the same position in the reference genome are marked as duplicates generated during PCR amplification and removed. Deduplication analysis is most appropriate for research requiring low depth of sequencing and is less appropriate as depth increases (Kukurba & Montgomery, 2015). Highly expressed transcripts are more likely to generate multiple reads that map to the exact same coordinates in the reference genome, which will be incorrectly marked as the same molecule during deduplication.

UMIS ENABLE RARE VARIANT DETECTION THROUGH ENHANCED READ SENSITIVITY

Detecting potentially malignant rare variants is critical since early indication of disease often correlates with patient outcome. Yet, detecting rare mutations with sequencing remains a significant challenge determined in large part by the limit of detection – the ability to resolve nucleotide fragments to distinguish a true variant from a processing error. The limit of detection of standard NGS is only 1:100, so a true variant will be mistaken for a processing error every one in one hundred times. Scientists studying clonal heterogeneity in search for early indicators of disease require methods with far greater sensitivity for confident results. A UMI-based approach greatly improves the limit of detection beyond what standard NGS offers, thereby increasing read sensitivity for the detection of rare variants. In fact, some UMI-based approaches such as RareSeq dramatically improve the limit of detection from 1:100 to 1:10,000, providing the sensitivity required to detect ultra-rare variant alleles (Crowgey et al., 2020).

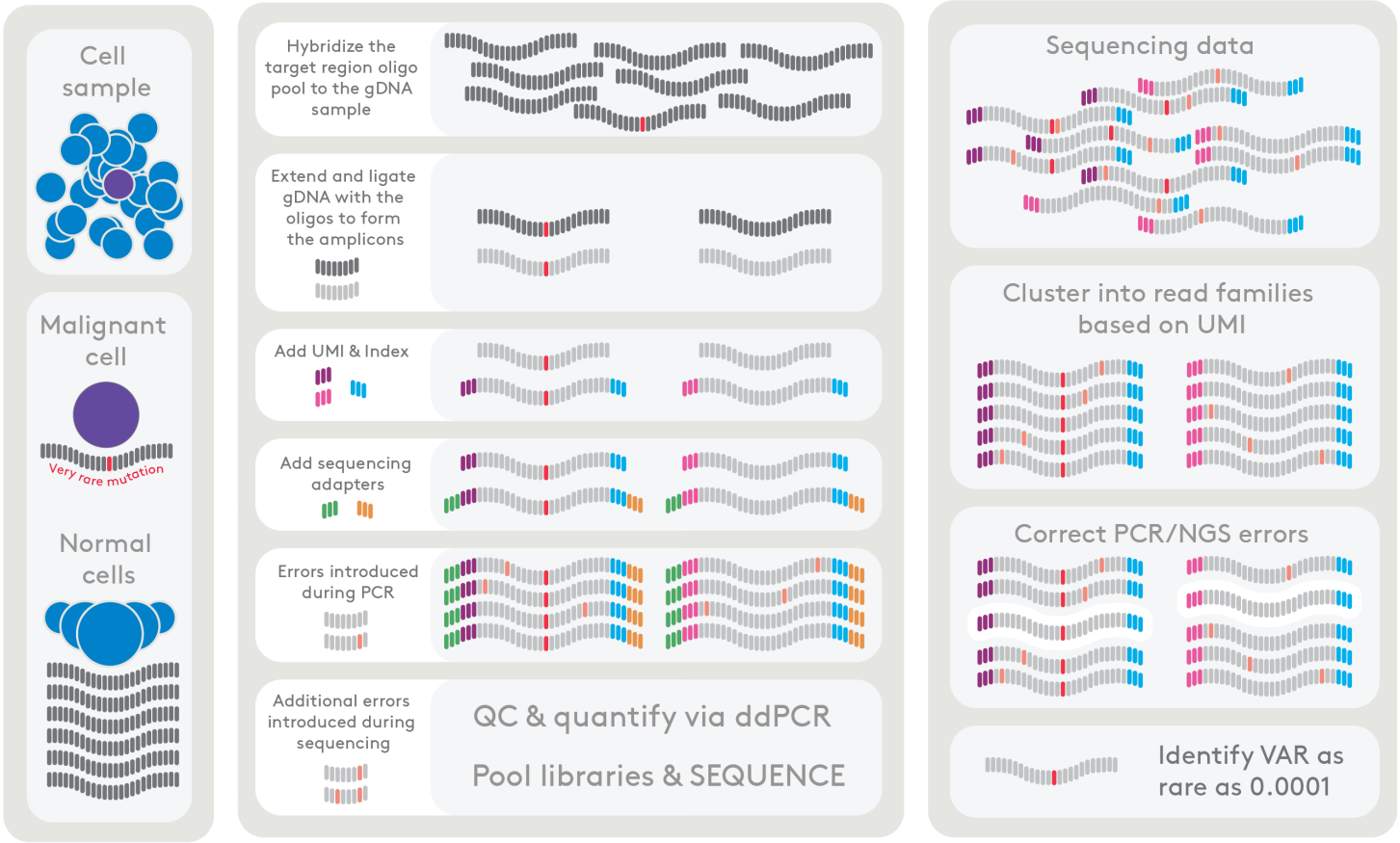

Figure 3. Left column: The RareSeq workflow begins with extraction of genomic DNA from cells in a sample and subsequent QC. Middle column: The gDNA is then used as the starting material for the library preparation and sequencing steps. Right column: Using bioinformatics software, the sequencing data is clustered into read families based on UMI to quantify read counts and identify true variants.

RARESEQ IS DESIGNED TO DETECT ULTRA-RARE VARIANTS THAT MAY LEAD TO MALIGNANCY

Clonal heterogeneity has emerged as a critical topic in disease research, since cells harboring driver mutations may predispose individuals to malignancy. Detecting these ultra-rare variants is critical or monitoring minimal residual disease (MRD) in patients with acute myeloid leukemia, where early indication correlates with patient outcome (Creutzig et al., 2014). Researchers at the Washington University Medical School, St. Louis noted that standard NGS lacks the quantitative sensitivity for monitoring MRD due to an inherently high error rate. To address this problem, the researchers developed RareSeq, a targeted error-corrected sequencing approach that employs UMIs to dramatically increase sensitivity by one hundred times greater than standard NGS (Crowgey et al., 2020).

With this technology, Crowgey and colleagues (2020) were able to detect previously undetectable clonal leukemic mutations with a high degree of sensitivity, enabling more accurate MRD tracking. Mutations in the same genes implicated in MRD are thought to also be responsible for at least some cases of clonal hematopoiesis of indeterminant potential (CHIP). Researchers harnessed the power of RareSeq and detected hematopoietic clones in 95% of individuals studied, a number far greater than is detectable with standard NGS (Young et al., 2016).

Taken together, the data in these studies demonstrates the utility of RareSeq to identify ultra-rare mutations that may lead to hematologic malignancy. Canopy Biosciences offers RareSeq error-corrected sequencing as a service for scientists conducting research aimed to detect ultra-rare AML-associated mutations.

For research use only. Not for use in diagnostic procedures.

SUMMARY

As the need for more accurate and sensitive sequencing methods has increased, scientists have turned to UMI-based approaches for genomic research. In this article, we summarize the key applications of UMIs – namely their ability to provide more accurate read count quantification and enhance the limit of detection to confidently detect rare variants. Although a number of examples exist, we describe one UMI-based targeted sequencing approach called RareSeq, designed to detect ultra-rare variant alleles that predispose individuals to hematologic malignancies. Researchers have already demonstrated the utility of RareSeq to identify ultra-rare mutations that may be implicated in MRD and CHIP (Crowgey et al., 2020; Young et al., 2016).

In summary, UMIs have evolved as a key solution in NGS research to address previously unanswerable questions.

REFERENCES

Bewicke-Copley, F., Arjun Kumar, E., Palladino, G., Korfi, K., & Wang, J. (2019). Applications and analysis of targeted genomic sequencing in cancer studies. Computational and Structural Biotechnology Journal, 17, 1348–1359.

Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. Genome Biology, 18(1), 77.

Creutzig, U., Zimmermann, M., Dworzak, M. N., Gibson, B., Tamminga, R., Abrahamsson, J., ... Kaspers, G. L. (2014). The prognostic significance of early treatment response in pediatric relapsed acute myeloid leukemia: Results of the international study Relapsed AML 2001/01. Haematologica, 99(9), 1472–1478.

Crowgey, E. L., Mahajan, N., Wong, W. H., Gopalakrishnapillai, A., Barwe, S. P., Kolb, E. A., & Druley, T. E. (2020). Error-corrected sequencing strategies enable comprehensive detection of leukemic mutations relevant for diagnosis and minimal residual disease monitoring. BMC Medical Genomics, 13(1), 32.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. Cold Spring Harbor Protocols, 2015(11).

Sarda, S., & Hannenhalli, S. (2014). Next-Generation Sequencing and Epigenomics Research: A Hammer in Search of Nails. Genomics & Informatics, 12(1), 2.

van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. Experimental Cell Research, 322(1), 12–20.

Young, A. L., Challen, G. A., Birmann, B. M., & Druley, T. E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nature Communications, 7(1), 12484.

Contact us at **hello.canopy@bruker.com**

For research use only. Not for use in diagnostic procedures.