Enhanced sequencing with UMIs

C. Novella-Rausell
M. Grudniewska-Lawton

Unique molecular identifiers, in short UMIs, are here to stay. A technology introduced almost ten years ago (Kivioja et al., 2012) has changed how we perform and analyze next generation sequencing (NGS) experiments today.

Polymerase chain reaction: friend or foe?

In the majority of currently available sequencing methods, library preparation relies on nucleic acid fragmentation followed by size selection and template amplification via polymerase chain reaction (PCR) (Figure 1). The amplification step allows us to exponentially increase the amount of starting material, thereby increasing the library quantity and warranting that each molecule in our original sample is sequenced. This is particularly important when working with low-input samples. Despite the obvious advantage of amplification, this reaction introduces bias in applications that rely specifically on "counting" of amplified molecules, for instance in RNA sequencing or copy number variation studies (Wan *et al.*, 2017). The efficiency of PCR amplification is dependent on the sequence of the amplicon (Kozarewa *et al.*, 2009) which, in turn, determines the level of inhibition by self-annealing (Acinas *et al.*, 2005). In practice, if one sequence is amplified 20% more than another it will be 237 (1.20^{30}) times more abundant after 30 rounds of amplification.

By definition, the PCR will produce identical copies of the starting material. However, these copies do not represent the original abundance of the transcripts; we refer to them as duplicates. These are not "true copies" of our targets, and it is in our best interest to either remove or acknowledge them. PCR duplicates can be identified computationally by their mapping coordinates. However, it has been shown that computational removal of duplicates underestimates the abundance of transcripts, specifically those that are short or highly expressed (Fu Y. *et al.*, 2018). A gene can be mistakenly identified as lowly expressed because we partially removed its true copies while trying to eliminate PCR duplicates. Thus, removal of duplicates that relies solely on their coordinates may negatively affect downstream analyses.

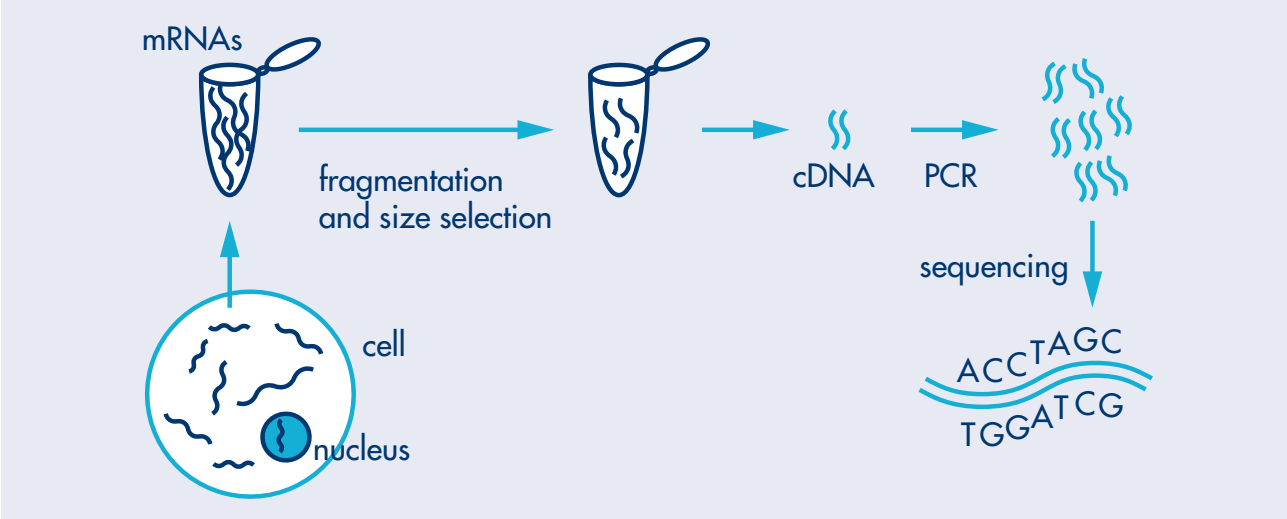

Figure 1. Library preparation and sequencing workflow.

Moreover, amplification with DNA polymerase is inherently flawed; base insertions, deletions, substitutions and rearrangements due to template switching can result in incorrect variant calls. This is troublesome in the diagnostics field, where some single nucleotide polymorphisms occur at an allele frequency as low as 0,1% (Filges *et al.*, 2019). These rare variants can be masked by the background noise produced by DNA polymerases, which have a larger error rate than the rarest variants (Fox *et al.*, 2014).

UMIs: tag and conquer

The introduction of UMIs was driven by both the need to solve the aforementioned amplification bias and to increase sequencing and quantification accuracy. UMIs are short random strings of nucleotides that uniquely barcode each molecule prior to amplification, allowing us to confidently account for PCR duplicates and to detect low-frequency alleles. Each of these barcoded molecules will represent a true DNA or cDNA copy (Figure 2).

The way in which the oligomers will be integrated in the library preparation protocol will depend on the application at hand. For instance, in several single cell gene expression technologies, the UMI is already present in the capture beads whereas in Illumina-based RNA-seq applications the UMI will be included in the indexing step (Figure 2). Nevertheless, all implementations will require UMIs to be added prior to amplification steps.

Design

One important consideration when using UMIs is the barcode length which determines the barcode library complexity. The more complex the barcode library is, the less likely the same UMI would barcode two biologically different molecules, which, in turn, decreases the false positive rates. However, one should not simply choose the longest UMI; longer oligomers are more prone to inherent sequencing errors, the effects of which are still largely unknown. Moreover, including longer UMIs reduces the effective sequencing length. Ideally, UMI length should be optimized using experimental benchmarks. In GenomeScan, we have adapted seven-nucleotide UMIs.

Once the barcodes are included in the sequencing library, the analysis within the standard processing pipeline is straightforward. In short paired-end sequencers, UMI can either be attached to one of two reads or as a stand-alone readout FASTQ file. In order to attach the stand-alone UMI read to the header of one of the paired-end reads, one can use one of many available tools e.g. UMI-tools (Smith *et al.*, 2017).

Applications

The major advantage of using UMIs is the ability to differentiate true templates from PCR duplicates. Furthermore, relying on UMIs, scientists have been able to accurately identify rare, low-frequency mutations (Kennedy *et al.*, 2014), sequence ultra-low amounts of genetic material from single cells (Macosko *et al.*, 2015) and accurately quantify

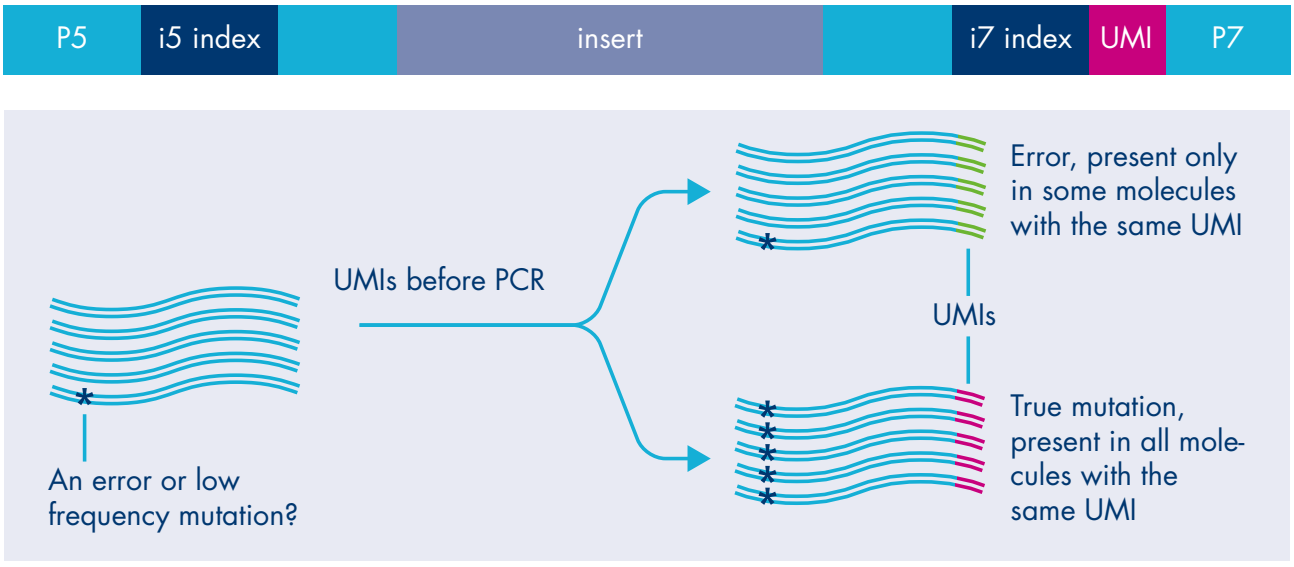

Figure 2. Dual-indexed library with UMI (top) and use case scenario (bottom).

reads (Smith *et al.*, 2017). Notably, the repertoire of UMI applications is continuously expanding, with a few examples highlighted below.

True duplicates and short transcripts

During nucleic acid fragmentation, short genes are more likely to yield identical reads than longer genes. Fragmentation is a random process and the probability of generating the exact same fragment twice depends on the sample space, which is the length of the sequence to be fragmented. Prior to the introduction of UMIs, one would mark a genuine copy of the transcript as a PCR duplicate due to the aforementioned bias. With UMIs, we can now confidently distinguish between two sequences that are identical and appreciate that they might come from different capture events of the same molecule (Fu Y. *et al.*, 2018).

Read error correction

Perhaps a less obvious application of UMIs is read error correction. When we identify two reads as true copies of a sequence, we can model and correct technical errors introduced by the sequencer. This is particularly important in applications where the identification of single nucleotide polymorphisms, or SNPs in short, is key (Peng Q.*et al.*, 2015).

Duplex sequencing

Recently, a new sequencing protocol has been developed in order to detect rare mutations, while leveraging the use of UMIs in paired-end sequencing (Kennedy *et al.*, 2014). In duplex sequencing, a double-stranded UMI is ligated on both ends of the read prior to amplification. PCR of each strand of the duplex results in two distinct PCR products, originating from the same strand of DNA. This allows the construction of single-strand consensus sequences (SSCS) for each of the two PCR products, accounting for the sequencing errors in each unique molecule. The chances of the same first-round PCR artifacts being present in both of the SSCS are negligible. Therefore, a true mutation will only be identified when it is present in both of the SSCS: in a duplex consensus sequence (DCS).

UMIs at GenomeScan

At GenomeScan we stand for the quality of our products and services. That is why we provide our customers with the latest advancements in sequencing technologies. For whole exome sequencing (WES) and targeted gene-panels, the GenomeScan expert team foresees major advances in oncology and pathology studies reporting on low frequency variants. For clinical geneticists, UMIs represent a means of optimizing sensitivity and specificity in WES-based diagnostics and in the development of robust CNV data analysis pipelines. We implemented the use of UMIs in all our DNA and RNA sequencing workflows.

You can benefit from UMIs too! Contact our representatives and start designing your next sequencing experiment with us.

References

Filges, S., Yamada, E., Ståhlberg, A., & Godfrey, T. E. (2019). Impact of Polymerase Fidelity on Background Error Rates in Next-Generation Sequencing with Unique Molecular Identifiers/Barcodes. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-39762-6

Fox, E. J., & Reid-Bayliss, K. S. (2014). Accuracy of Next Generation Sequencing Platforms. Journal of Next Generation Sequencing & Applications, 01(01). https://doi.org/10.4172/2469-9853.1000106

Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D., & Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. BMC Genomics, 19(1). https://doi.org/10.1186/s12864-018-4933-1

Kennedy, S. R., Schmitt, M. W., Fox, E. J., Kohrn, B. F., Salk, J. J., Ahn, E. H., Prindle, M. J., Kuong, K. J., Shen, J.-C., Risques, R.-A., & Loeb, L. A. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. Nature Protocols, 9(11), 2586–2606. https://doi.org/10.1038/nprot.2014.170

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature methods, 6(4), 291–295. https://doi.org/10.1038/nmeth.1311

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. Scientific Reports, 6(1). https://doi.org/10.1038/srep25533

Peng, Q., Vijaya Satya, R., Lewis, M., Randad, P., & Wang, Y. (2015). Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. BMC Genomics, 16(1). https://doi.org/10.1186/s12864-015-1806-8

Potapov, V., & Ong, J. L. (2017). Examining Sources of Error in PCR by Single-Molecule Sequencing. PLOS ONE, 12(1), e0169774. https://doi.org/10.1371/journal.pone.0169774

Should I remove PCR duplicates from my RNA-seq data? | DNA Technologies Core. (n.d.). DNA Technologies & Expression Analysis Core Laboratory. UCDAVIS Genome Center. Retrieved March 11, 2021, from https://dnatech.genomecenter.ucdavis.edu/faqs/should-i-remove-pcr-duplicates-from-my-rna-seq-data/

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Research, 27(3), 491–499. https://doi.org/10.1101/gr.209601.116

Vallabh Minikel, E. (2012, December 11). How PCR duplicates arise in next-generation sequencing. CureFFI. https://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R., & Rosenfeld, N. (2017). Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nature Reviews Cancer, 17(4), 223–238. https://doi.org/10.1038/nrc.2017.7