GUIDE

Twist UMI Design Overview and Usage Guidelines

For Research Use Only (RUO). Not for use in diagnostic procedures.

The Twist UMI Adapter system consists of Twist UDI primers and Twist UMI adapters. This guideline details the Twist UMI adapter design and how to properly implement UMIs into a sequencing workflow.

INTRODUCTION

While next-generation sequencing (NGS) is the new standard for detecting genetic variants, it is not perfect and there are various sources of errors that can bias the results of any NGS workflow. The sequencing process itself and various upstream steps such as PCR amplification during library preparation have baseline error rates that cannot be avoided. In certain use cases, these processes introduce too much bias into the sequencing results and are particularly impactful when identifying low-frequency variants or when sequencing is performed on low-input samples. To correct for these biases and accurately detect variants in settings with low input or where deep sequencing is required, unique molecular identifiers (UMIs) can be incorporated into an NGS workflow.

The Twist UMI Adapter system consists of Twist UDI primers and Twist UMI adapters. Twist UMI adapters are 5-bp matched molecular indices compatible with 'T-A' overhang workflows. During library preparation, these UMI adapters tag unique DNA molecules which are subsequently PCR amplified with Twist UDI primers. This enables users to identify amplified DNA molecules and match them to their original template DNA, creating UMI families. By analyzing each family, data pipelines can correct amplification errors and generate more accurate reads based on the consensus of all DNA strands within a family (**Figure 1**).

This guideline details the Twist UMI adapter design and how to properly implement UMIs into a sequencing workflow.

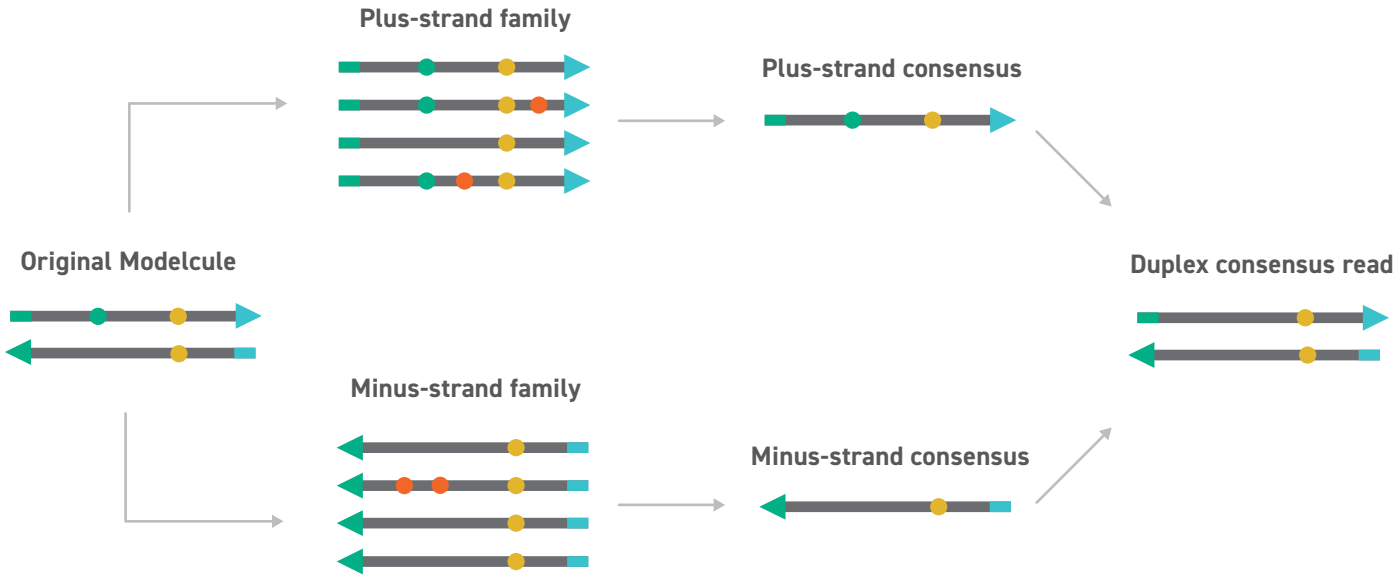

Figure 1. Overview of Obtaining Consensus Reads Using UMI Adapters.

POTENTIAL APPLICATIONS

LOW-FREQUENCY VARIANT CALLING

Sequencing to detect low-frequency variants (e.g., liquid biopsy using cfDNA) comes with a host of challenges. Among these challenges is the often low abundance of cfDNA in the plasma which limits the amount of input DNA that can undergo library preparation and sequencing. A low input amount with a low-frequency variant means only a few mutant molecules will be present within a given sample. As mentioned previously, it is challenging to develop a sensitive workflow without UMIs that can accurately call variants at this level of rarity as errors that arise from PCR amplification and sequencing cannot be distinguished from true variants. By analyzing UMI families, which consist of reads that stem from the same DNA molecule, deep sequencing can provide more accurate information regarding the original DNA molecules.

QUANTITATIVE ESTIMATES OF LIBRARY DIVERSITY

Beyond error correction, UMIs can assist in distinguishing between PCR and library duplicates to get accurate estimates of total molecular counts. For example, a major concern for quantitative measurements of RNA is the presence of PCR duplicates in the analysis. By adding UMIs to the original molecules prior to library preparation, duplicates can be identified and removed during data analysis as each read with the same UMI sequence as another is likely a duplicate. Analyses that count unique UMI reads can provide more accurate gene expression estimates. Similarly, UMIs can assist in getting accurate counts in applications such as bisulfite sequencing for methylation where the template is damaged or bottlenecked during library preparation. In bisulfite sequencing, DNA samples undergo bisulfite conversion, wherein unmethylated cytosines are converted to uracil (methylated cytosines remain the same). Thus, after PCR amplification, these unmethylated cytosines are identified as thymine in sequencing data. Consequently, the diversity of the sequencing data is reduced and there is an increased probability of identifying a unique read as a duplicate. In these cases, the use of UMI adapters allows for more accurate deduplication.

UMI DESIGN STRATEGY

Twist UMIs were designed as a set of 32x32 discrete pairs of adapter sequences 5 bp in length, forming a total of 1024 potential combinations of inline UMI sequences. To ensure that UMI family members can be easily distinguished even with sequencing errors, there is a minimum Hamming distance of 2 between any pair of UMI adapter sequences. This allows for error correction of any single base change.

The design of our 5-bp UMI sequences ensures the absence of homopolymers and includes sequences with moderate GC content. Longer UMI sequencing lengths will allow for further optimizations and potential pairs, but will also require more sequencing cycles and it was determined that the 5-bp length provided a strong balance between desirable UMI sequences and sequencing efficiency. A major concern with any analysis pipeline involving UMIs is the possibility of UMI collisions, which are events in which two reads will have identical sequences and UMI adapters despite originating from two separate molecules. With the design detailed above, the likelihood of UMI sequence errors leading to UMI collisions is low. The average family size in most UMI sequences experiments is generally small (often 5 or less). Because family sizes are generally smaller than the 1024 available pairs of duplex UMIs, the probability of a collision in any given family is quite low.

LABORATORY CONSIDERATIONS

INPUT AMOUNT

The DNA input mass into library preparation is a critical factor in both the sensitivity and the required sequencing depth for a low-frequency variant calling experiment. If a large quantity of DNA is used (>100 ng), it is likely that the experiment will require extremely high sequencing depths (>100,000x depth) to recover multiple supporting reads for a large proportion of consensus molecules. Conversely, low DNA input amounts (<10 ng) will impact the sensitivity of the assay to low-frequency variants. To see why, take a hypothetical example where we provide 3 ng of DNA input into library preparation and capture. Since a haploid human genome weighs approximately 3 pg, this amount of DNA translates to ~1000 copies of the genome. We have empirically found that cfDNA generally has a conversion efficiency of ~30% through library preparation and capture, which would leave ~300 sequenceable molecules at each nucleotide position. If the variant alleles are present in a low fraction (i.e., 0.1%), there would be on average less than one available molecule in the library supporting each variant. Because not every molecule in the library will have sufficient UMI family members to collapse into a consensus error-corrected read, it may be impossible to detect variants at the majority of the expected sites in the genome.

PANEL SIZE AND DIVERSITY

Larger panels will require correspondingly larger numbers of sequenced reads to achieve similar depth (and therefore similar sensitivity and specificity to low-frequency variants) and will require more computational resources to analyze. Typically, the analysis and sequencing requirements become a significant burden for panels greater than a few hundred kilobases in capture space. Correspondingly, panel performance (particularly off-target) is an extremely important consideration for these experiments. Given the high depths of sequencing involved, even slight increases in off-target can correspond to large amounts of additional sequencing required to achieve the desired depth over the panel targets. It is important to evaluate the panel performance prior to read collapse and filtering since many off-target reads are likely to be non-specific (will occur essentially randomly throughout the genome). These off-target reads generally cannot form consensus UMI families and will be removed during the consensus filtering process and, thus, the post-consensus collapse capture metrics will give a misleadingly small estimate of the panel off-target. Therefore, assessments of a panel's off-target performance must be performed prior to implementation of a UMI adapter system.

VARIANT TYPES AND ALLELES

Another important consideration for designing experiments is the nature of the variants that are being detected. Short variants (like SNPs and short indels) tend to be easier to detect in targeted sequencing experiments for two main reasons. First, because probes tend to target the reference sequence of the target genome, short variants have fewer mismatches from the reference and therefore are more likely to be captured efficiently. Second, during bioinformatic processing short variants will be tolerated better by the genome aligner and are less likely to be clipped if they occur near the edges of the read. Therefore, the estimated allele fractions for short variants will often be more accurate than those for longer variants (particularly SVs). It's also important to consider the context in which a variant occurs. For example, short mononucleotide repeats are prone to single-base indel events during sequencing. Although UMI error correction can mitigate false positives in these contexts, the high rate of errors still makes it likely that the same error will simultaneously occur in multiple reads.

SEQUENCING DEPTH

Of the factors discussed, sequencing depth is probably the easiest of these factors to adjust as sequencer space is generally limited only by cost. Sequencing a sample too little runs the risk of not finding all the available diversity in the sample and not having enough independent observations of each molecule for effective error correction. Conversely, there is a point at which most of the available diversity of the sample has been sequenced and there are diminishing returns in obtaining more observations of each UMI family. A general rule of thumb for most experiments is to try to obtain a mean coverage over the targeted regions of the genome that is roughly the mass input into sequencing (in nanograms) times 1000. For an experiment with 30 ng of input mass, a depth of roughly 30,000x should provide the most "efficient" yield of consensus reads. However, for extremely low-frequency alleles, it may be beneficial to increase depth at the expense of sequencing costs.

PROCESSING UMI SEQUENCING DATA

There are various command-line toolkits available for processing and analyzing bioinformatic data. Among them, fgbio can work on read-level or variant-level data and includes tools specialized for processing UMI sequencing data. We recommend fgbio for applications that require error correction (e.g., low-frequency variant calling). Alternatively, UMI-tools is another toolkit with functions suited for quantitative assessments of diversity. This section generally describes how to run a bioinformatics pipeline (**Figure 2**) and considerations for specific command-line tools used to process or analyze UMI sequencing data.

Briefly, the recommended bioinformatics pipeline for processing UMI data involves converting the initial FastQ file into an unaligned BAM file in order to extract the UMI sequences and store them as tags in the BAM file. The unaligned BAM is converted back to a FastQ file format and the reads are then aligned to a reference genome. To allow for UMI grouping, the unaligned BAM is merged with the aligned BAM containing the UMI tags. At this point, HsMetrics can be collected to understand assay performance. The reads can then be grouped into UMI families that all likely stem from the same original molecule. Consensus reads can then be called using various filtering and collapsing methods. These consensus reads are realigned to the genome after collapse to make use of any quality improvements coming from the consensus call. For applications that use UMIs to count distinct reads (rather than error correction) a single member may be picked from each family (deduplication) instead of generating a consensus read.

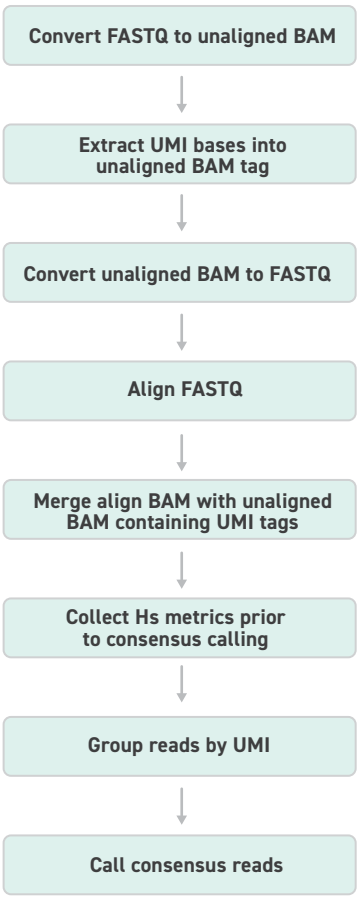

Figure 2. Summary of a general UMI bioinformatics pipeline

Some considerations need to be taken into account when collapsing consensus reads. Generally, reads are filtered according to the total number of supporting family members. In duplex-sequencing experiments these filtering criteria often specify the number of reads required from each original strand - for example, in fgbio, the filtering criteria consist of three numbers. The first number is a cutoff for the total number of required reads, the second number is a cutoff for reads required from the strand (plus or minus) with greater representation, and the third number is the cutoff for reads required from the less represented strand. For example, a filter of 3-2-1 would reject a UMI family with one read from each strand, since it would neither have the total number of required reads (3) nor the number of reads required from the more-represented strand (2). On the other hand, for the dedup command in UMI-tools, no filtering parameters are necessary since one representative from each family is always kept.

Requiring more stringent filtering will decrease the overall error rate in the collapsed reads, as it is increasingly unlikely that the same sequencing error will occur in all of the represented family members. At the same time, more stringent filters will remove a larger fraction of the original read data, as there will be fewer reads that fall into UMI families with sufficient evidence to pass the filter. In practice, we've found that requiring at least one read from each original strand (i.e. a 2-1-1 filter) is a good tradeoff between error rate minimization and retaining the sample's read diversity. Requiring additional reads from each strand has a more marginal effect on accuracy than requiring at least one read from each original strand, as the two different strands provide more independent observations of the genotype at target sites than reads deriving from the same original strand.

In cases where duplicate counting (rather than error correction) is the primary experimental goal, it is important to avoid filtering reads out because of a lack of evidence. In general, it is most appropriate to use a 1-1-0 filter for these experiments to ensure that the duplex collapse process does not discard reads because of low base or mapping quality (a 1-1-0 filter only requires a total of 1 read in a family so no reads overall are discarded). In RNA-seq applications duplex collapse is not appropriate (as the original molecules are not duplexes), while in methylation applications the differences in conversion readout between the plus (C>T) and minus (A>G) strands may lead to excessive numbers of masked bases in the duplex output. For this reason, we recommend the use of single consensus in these applications with utilities like the dedup functionality of UMI-tools.

CONCLUSION

Twist UMI Adapters can be used in various applications including calling low-frequency mutations and quantitative assessments of diversity. However, there are various considerations that one must take into account when using UMI adapters including input mass, panel characteristics, variant types, sequencing depth, and analysis tools. The above considerations and recommendations will assist in effectively designing efficient experiments that require the use of UMI adapters.

For additional details and other specific recommendations for running a DNA sequencing data pipeline please refer to our Processing Sequencing Data Utilizing Twist Unique Molecular Identifier (UMI) Adapter System guideline found here: https://www.twistbioscience.com/resources/guideguideline/processing-sequencing-data-utilizing-twist-unique-molecular-identifier-umi

HAVE FURTHER QUESTIONS?

For additional support please contact Twist Bioscience's support team at **customersupport@twistbioscience.com.**