Unique Dual Indexing Primers and Universal Blockers Enable Multiplexed Target Enrichment Applications

Kristin Butcher, Leonardo Arbiza, Siyuan Chen, Richard Gantt, Brenton Graham, Sabina Gude, Rebecca Nugent, Christina Thompson

TWIST BIOSCIENCE

1. Abstract

Target enrichment has been an enabling approach to sequencing specific regions of the genome and calling variants while reducing sequencing cost per sample. The ability to create customizable panels has been a key driver for increased utilization of target enrichment within various workflows and applications.

However, two important components, **(1)** unique dual-indexed (UDI) Y-adapters for library creation and **(2)** specific adapter blockers used to improve on-target rates during target enrichment, have seen few recent advancements and are fast becoming limiting factors in multiplexed target enrichment applications. Synthesis of UDI Y-adapters is challenging due to their relative long length (>65bp), 5'-phosphate chemical modification, and purity requirements that demand HPLC purification of two oligonucleotides for every desired index set. They also commonly generate significant amounts of undesired adapter dimer during ligation reactions. Some 'universal' adapter blockers are commercially available, but their application is limited to subsets of index lengths and artificially constrain experimental design.

To address these limitations, fundamental changes to **(1)** installation of UDI sequences during library creation and **(2)** adapter blockers were investigated and provided numerous benefits. Application of the Twist Universal Adapter System reduced undesired adapter dimer formation below the lower limit of detection under common reaction conditions, increased total gDNA library yield 2X with a 20% reduction of PCR amplification cycles, and enabled facile switching between index designs. In addition, facile synthesis allows for access to thousands of UDI pairs so the number of samples and amount of data generated during sequencing can be maximized. To complement this robust adapter system, Twist Universal Blockers eliminate any requirement for specific adapter blockers. This design functioned independently of index length, increased capture of on-target reads when compared to sequence-specific adapter blockers on a molar basis, was agnostic to target enrichment panel size, and remained functionality equivalent in singleplex and multiplex workflows.

Taken together, these advancements in the application of shared adapters for UDI sequence installation and universal adapter blockers eliminate numerous bottlenecks for target enrichment multiplexing applications.

2. Barcode Independent Universal Blockers

A critical metric that limits performance of multiplex hybrid capture is the percentage of bases on-target. To address this, commercial hybrid capture systems typically employ two types of blocker systems to improve this metric for human samples: **(1)** Cot DNA, a product enriched for repetitive human gDNA and **(2)** adapter blockers, designed to limit adapter cross-hybridization. While sequence specific adapter blockers provide acceptable performance for small barcode sets, they hinder adoption of expanded index sets, restrict changing barcode length between experiments, and prohibit use of unique dual indices (UDI) or unique molecular identifiers (UMI).

A universal blocker system was developed to address these bottlenecks by increasing capture of on-target reads independent of barcode sequence length or design for all TruSeq® compatible adapters (**Figure 2.1**). Functionality is also retained regardless of target panel size (**Figure 2.2**) and across singleplex and multiplex target enrichment workflows (see **Sections 3** & **4**). As a result, this universal adapter blocker technology drastically increases on- and near-bait capture, allows for index flexibility between experiments, simplifies design and execution of multiplex experiments, and directly reduces sequencing costs.

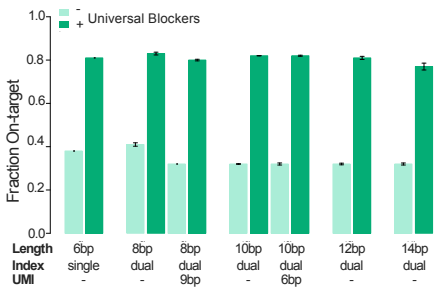

Figure 2.1: Performance is independent of barcode design. On-target performance of universal blockers (Twist Bioscience) across a variety of single and dual index TruSeq-compatible adapters. Individual libraries were generated from a single genomic source (NA12878; Coriell) and TruSeq-compatible adapters with barcode lengths ranging from 6 to 14 bp (UMI length not included). Hybrid capture was performed either in the absence or presence of universal blockers using an exome target enrichment panel (33.1 Mb; Twist Bioscience) using 500 ng of gDNA per individual library following the manufacturer's recommendations for 16-hour hybridization reactions. Cot DNA was present in all samples. Fraction of bases on-target is defined by the equation 1 - PCT_OFF_BAIT. Error bars denote one standard deviation or range of observations; N ≥ 2.

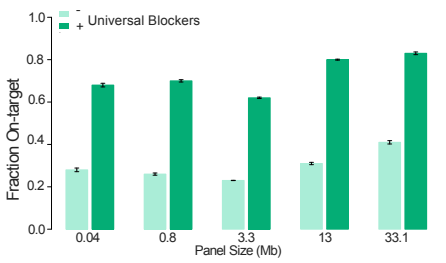

Figure 2.2: Performance is independent of panel size. On-target performance of Universal Blockers (Twist Bioscience) across a variety of panel sizes. Individual libraries were generated from a single genomic source (NA12878; Coriell) and 8 bp dual-indexed TruSeq-compatible adapters. Hybrid capture was performed either in the absence or presence of universal blockers with target enrichment panels of various sizes (0.04Mb to 33.1Mb; Twist Bioscience) and 500 ng of gDNA per individual library following the manufacturer's recommendations for 16-hour hybridization reactions. Cot DNA was present in all samples. Fraction of bases on-target is defined by the equation 1 - PCT_OFF_BAIT. Error bars denote range of observations; N = 2.

3. Improved Unique Dual Index Library Creation

With standard full length Y-adapters, NGS library preparation of challenging samples with 'T-A' overhang ligation often requires excessive PCR cycles, creating unwanted bias from amplification. In addition, the presence of adapter dimer at even low molar concentrations can overestimate the concentration of final libraries and lead to underloading when pooling on a mass basis for multiplexed hybrid capture applications. A 10bp Unique Dual Indexed (UDI) system with Universal Adapters was developed to alleviate these library preparation bottlenecks by increasing total recovered yield with a reduced number of PCR cycles (**Figure 3.1**) and eliminating adapter dimer formation under standard ligation conditions (gDNA:adapter ratio of 1:200; **Figure 3.2**).

Used in coordination with universal blockers, this adapter system not only improves numerous aspects of NGS library preparation, but lays the foundation for execution of high multiplex hybridization capture.

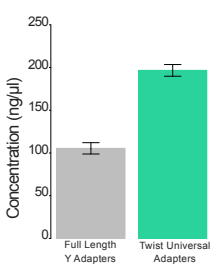

Figure 3.1: Improved Final Yield. Application of the 10bp UDI system for NGS library creation returns higher total yields when compared to standard full length Y-adapters with fewer PCR cycles. For all data, 50ng from a single genomic source (NA12878; Coriell) was processed with a commercially available NGS Library Preparation Kit (Twist Bioscience). Samples were generated with either commercially available standard TruSeq compatible Y-adapters and amplification primers with 10 PCR cycles (N = 12) or the 10bp UDI system with 8 PCR cycles (N = 12). gDNA:adapter ratios were held at a constant molar ratio of 1:200 for all reactions. Equal volumes of the final prepared libraries were evaluated with a Qubit dsDNA Broad Range Assay Kit (Invitrogen) for determination of final yield.

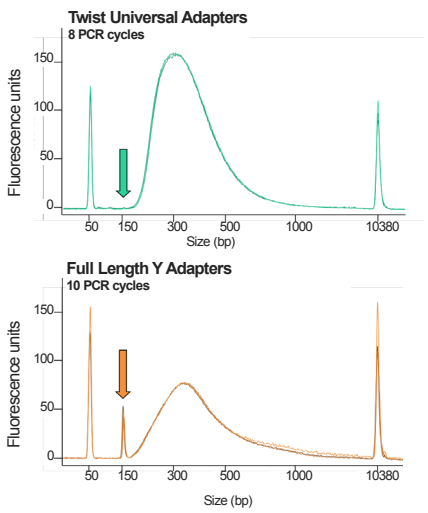

Figure 3.2: Elimination of Adapter Dimer Formation. Application of the 10bp UDI system eliminates formation of adapter dimer during gDNA library preparation. 50ng from a single genomic source (Coriell; NA12878) was processed with a commercially available NGS Library Preparation Kit. Samples were generated with either commercially available standard Full Length TruSeq compatible Y-adapters and amplification primers with 10 PCR cycles or the 10bp UDI system with 8 PCR cycles. gDNA:adapter ratios were held at a constant molar ratio of 1:200 for all reactions. Equal volumes of the final prepared libraries were evaluated with an BioAnalyzer 7500 Assay® (Agilent®). Overlays of representative electropherograms from both sample types presented (N = 3).

Materials & Methods

Unless otherwise noted, sequencing was performed with a NextSeq® 500/550 High Output v2 kit to generate 2x76 paired end reads. Data was down-sampled to 150x of target size and analyzed using Picard Metrics with a mapping quality of 20.

4. UDI Primer Performance

As the amount of data generated per sequencing run has continued to increase, the number of samples that can be analyzed per run has been artificially limited due to a lack of large (>1,000 members) and high performing unique dual index (UDI) sets. One limiting factor in the identification of such UDI sets has been the lack of cost effective workflows to generate thousands of NGS libraries with a representative genomic insert.

To address this limitation, a high throughput NGS library generation method was developed. This method uses enzymatic fragmentation of a single human genomic source and automation to generate thousands of NGS libraries in a cost effective manner. The presence of human genomic material as insert is critical to ensure that any bias during NGS library creation is equally represented throughout the population.

Using this approach, 1,152 NGS libraries containing unique dual sequences were constructed and screened in an iterative fashion for even sequencing performance. As a result, a set of 384 UDI sequences were identified that provide sequencing performance relative to the mean of +/-25% either as a single large pool or as individual sets of 4 x 96 members. While this approach identified a set of UDIs suitable for high multiplex hybridization applications and investigations, the groundwork is also laid for identification of larger UDI sets in the near future.

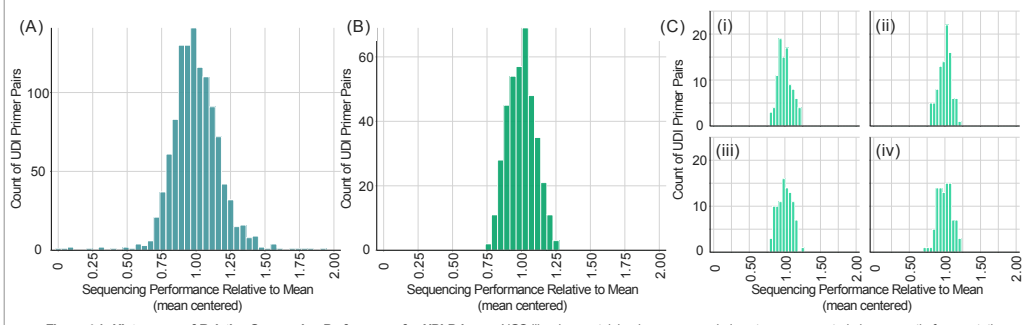

Figure 4.1: Histograms of Relative Sequencing Performance for UDI Primers. NGS libraries containing human genomic insert were generated via enzymatic fragmentation (Twist Enzymatic Fragmentation Kit). Individual libraries were pooled by mass and sequenced with a NextSeq 500/550 High Output v2 kit to generate 2 x 10bp index reads. The total count of individual pairs of index reads (1 mismatch allowed) was determined and the relative performance of each individual pair was calculated relative to the mean. **(A)** Initial screening set of 1,152 UDI Primer Pairs sequenced as a single pool, **(B)** 384 UDI Primers pairs sequenced as a single pool, **(C)(i-iv)** 4 individual pools of 96 UDI Primers pairs sequenced independently.

5. Robust Multiplexing Performance

To support the described blocker and adapter systems, target enrichment probes must also be designed to match the desired throughput of the system. Probes were designed to maximize the capture of unique molecules and minimize sequencing duplicates to delivery high multiplex performance. The adapter-blocker-probe system described was developed to maintain high-performance with up to a 16-plex capture. This is demonstrated on three panels of 800 kb, 3.3 Mb and a fixed Exome of 33.1 Mb. Consistent capture coverage at 30x is observed across all samples and multiplexing conditions (**Figure 4.1**).

While an increase in duplicate rate is expected as the number of samples in a multiplex increases, the magnitude of this increase is minimal. For an 800kb panel duplication rate increases from 1.8% to 2.7% between 1-plex and 16-plex captures, respectively, and similar observations were made with larger panels. The minimal impact to performance is confirmed with consistent 30x coverage.

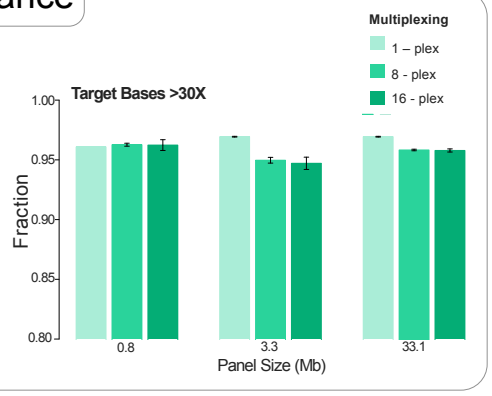

Figure 4.1: Multiplexing Performance. Multiplexing was performed for three panels at three degrees of multiplexing. Hybrid capture performed using target enrichment panels of various sizes (800kb cancer panel, 3.3Mb cancer panel, or 33.1 Mb exome panel; Twist Bioscience) using a total of 1500 ng library (NA12878; Coriell) per multiplexed pool following manufacturer's recommendations.

6. Fast Hybridization for Throughput Improvements

While the development of the blocker-adapter-probe system provides a robust foundation for multiplexing, throughput continues to be limited by the length of hybridization. However, the quality of a capture is related to how close it is to equilibration and is time dependent. The target enrichment system described thus far is a 16 hour hybridization and shorter times enable higher throughput with the undesired cost of performance in target coverage and other critical metrics. To radically increase the throughput of samples that can be processed, the conditions for hybridization must be optimized such that the system comes to equilibrium on a shorter timescale.

In order to address this, numerous step of the target enrichment buffer system were either optimized or re-engineered into a 'Fast Hybridization' system that arrives at equilibrium on a greatly reduced timescale and provides excellent sequencing performance in 4 hours or less (**Figure 5.1**). Capture performance was evaluated at 4 hours, showing expected uniformity, off target, and overall coverage are equivalent with panels designed for 16 hour capture (**Figure 5.2**).

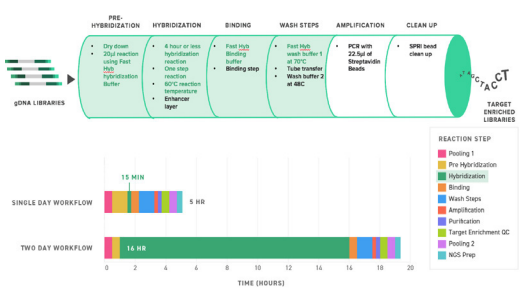

Figure 5.1: Fast Hybridization Workflow. (A) General workflow present in the 'Fast Hybridization' system (Twist Bioscience). (B) Timing of different steps are shown here and demonstrate the reduction of overall time to hybridization and sequencing. The vast majority of time savings come from the reduction of hybridization time from 16 to a maximum of 4 hours.

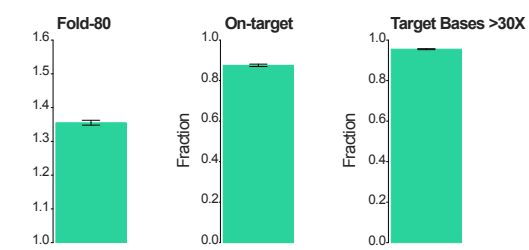

Figure 5.2: Fast Hybridization Performance. Hybrid capture was performed using an exome target enrichment panel (33.1 Mb; Twist Bioscience) using 187.5 ng of library (NA12878; Coriell) and the 'Fast Hybridization' protocol at 4 hours. N = 2. a) Uniformity (Fold 80) b) On Target, the fraction of bases on-target is defined by the equation 1 - PCT_OFF_BAIT. c) 30x Coverage.

In addition, this system generates high-coverage capture results across a range of times with minimal performance degradation (**Figure 5.3**). Finally, this system is compatible for high-multiplexing applications using the adapter, blocker, and probe systems described above and enables a significant improvement for target enrichment throughput in regards to both samples and time (**Figure 5.4**).

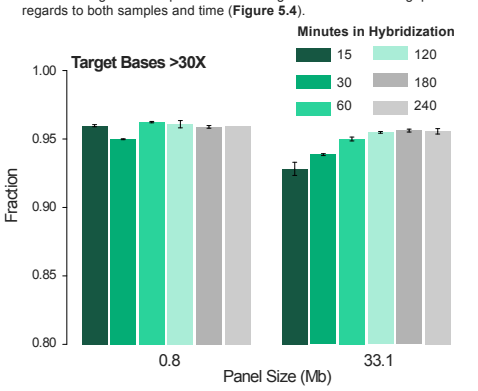

Figure 5.3: Fast Hybridization Time Course. Hybrid capture was performed using an exome target enrichment panel (33.1 Mb; Twist Bioscience) using 187.5 ng of library (NA12878; Coriell) and the 'Fast Hybridization' protocol at variable hybridization times. N = 2. Coverage at 30x is plotted. a) Exome panel (33.1 Mb), b) 800kb panel.

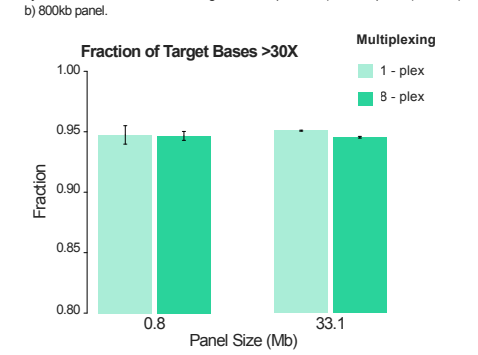

Figure 5.4: Fast Hybridization Multiplexing. Hybrid capture was performed using an exome target enrichment panel (33.1 Mb; Twist Bioscience) using 187.5 ng of library (NA12878; Coriell) per library per multiplexed pool (8-libraries, 1500 ng total) and the 'FastHyb' protocol at 4 hours. N = 2. Coverage at 30x is plotted. a) Exome panel (33.1 Mb), b) Cancer panel (800 kb).