Increasing NGS throughput with >3,000 Unique Dual Indices (UDIs)

Lydia Bonar, Tom Richardson, Maen Sarhan, Adriana Arneson, Richard Gantt, Palak Sheth, Rebecca Liao, Derek Murphy, and Esteban Toro

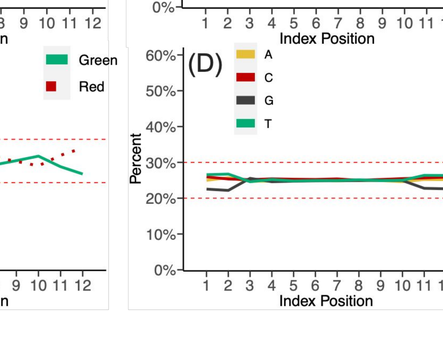
1. Abstract

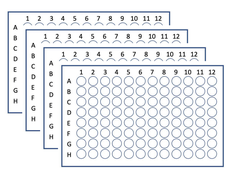
The number of commercially available Unique Dual Index (UDI) primer pairs remains limited, with few commercial offerings above 1,536 UDI pairs. In an era of ever-increasing sequencing capacity, this unnecessarily limits sample throughput and increases the cost per sample. To address these limitations, Twist Bioscience developed a new set of 3,072 UDIs that enables higher sample throughput in large experiments.

The design of large sets of barcodes can be difficult due to the exponential nature of barcode-barcode interactions. Twist's design approach includes (i) sophisticated index design, (ii) subsetting index pairs into bins that are base and color channel balanced, and (iii) empirical validation of each index pair across multiple sequencing platforms. This approach enabled the generation of a large and performant set of UDI primers described here.

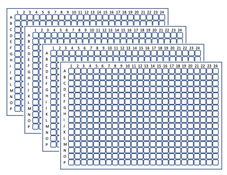
2. Index Design

Index design is a multi-step process, as shown in diagram 1 on the right. Barcode sequences are first filtered individually for desirable sequence characteristics such as GC content, absence of hairpins or homopolymers, etc. Then, barcode-barcode interactions are taken into consideration and pairs of barcodes are chosen, making sure that the pool of sequences works well as a set. Next, barcode pairs are assigned to plates taking into consideration the color balance of two- and four-color sequencing instruments. This enables any plate of barcodes to perform well as a standalone set in a sequencing experiment. Finally, all barcodes are tested empirically to filter out poor performers, and re-balanced by color into final plates (8 x 384-well plates and 32 x 96-well plates), to generate the final set of 3,072 barcode pairs. All barcodes are 12nt in length to allow for sufficient sequence space to meet all performance requirements.





32 x 96-well plates



8 x 384-well plates

3. Color balance

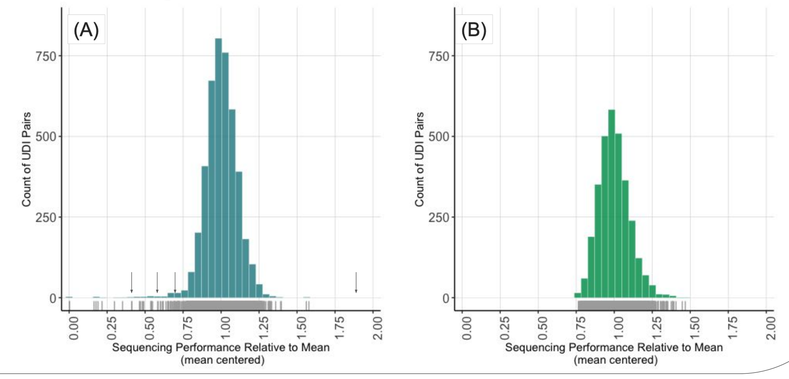
Our barcode generation algorithm is designed to balance color frequency in two- and four-color instruments. Figure 1 shows the color balance of two sets of barcodes, generated using different design algorithms. Between the two, only the second algorithm generates designs that are balanced in both types of instruments.

Figure 1 : Examples of two algorithms for color channel balance for 2- and 4-color sequencing chemistries. Algorithm 1 demonstrates results that provide good balance for 2-color sequencing chemistry (A), but unbalanced for 4-color sequencing chemistry (B). Algorithm 2 demonstrates an index set that is well balanced for 2-color (C) and 4-color sequencing chemistry (D).

4. UDI Primer Uniformity

Individual barcode sequences can sometimes yield abnormally high/low read counts consistently. This creates sequencing runs with poor uniformity, where some samples receive too many reads while others don't receive enough. To avoid this problem, the original set was screened empirically in order to select UDIs that show uniform read counts across the entire set of barcodes. The final UDI set is suitable for high multiplex applications, providing uniform sequencing performance, as shown in Figure 2.

Figure 2 : Read uniformity of UDI Primer sets. Histograms of the relative number of reads produced by each UDI in a pool of libraries (1 mismatch allowed during de-multiplexing) **(A)** Initial screening set of 4,352 UDI Primer Pairs sequenced as a single pool, **(B)** Final 3,072 UDI Primers pairs sequenced as a single pool. Note the presence of outliers with abnormally high/low read counts in the original pool (black arrows).



5. UDI Primer Fidelity

Individual barcodes may show high numbers of errors, which can lead to lost reads or barcode mis-assignment. Figure 3 shows that, for the original set, some UDIs had an elevated number of reads with errors in the barcode sequence (A). After screening those poor performers out, the final set shows a substantial reduction in errors (B). Our set of HT UDIs have high percentages of error-free barcode sequences on patterned (C) and unpatterned (D) sequencing flow cells.

All barcodes have a minimum hamming distance of 3, which makes barcode assignment unambiguous even for reads that have one error. Our set of UDIs typically yield >99% percent of reads assigned with 0 or 1 index errors.

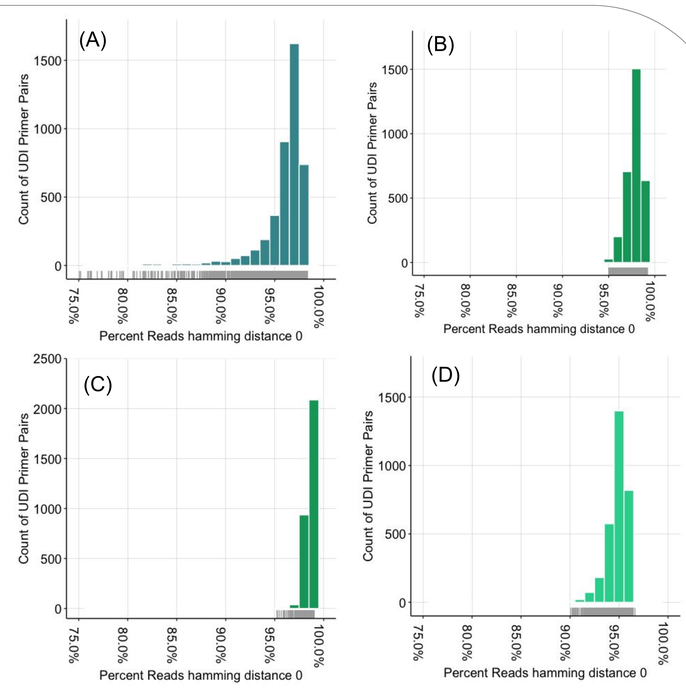


Figure 3 : Error rate of UDI sequences. The percent of reads with no mismatches was determined for UDI sets sequenced as a single pool. **(A)** Initial screening set of 4,352 UDIs sequenced on a NextSeq 2000 sequencer (P2 kit). **(B-D)** Final set of 3,072 UDIs sequenced on a NextSeq 2000 sequencer (P2 Kit) **(B)**, a NovaSeq 6000 sequencer (S4 kit) **(C)**, or a NextSeq 550 sequencer (high output v2 kit) **(D)**.

6. Quality Control

Manufacturing batches of UDIs require high purity standards in order to support sensitive applications. Figure 4 shows a representative plate where barcode purity has been measured, under conditions where confounding template contamination effects are controlled. As shown, the level of purity in the plate is exceptionally high.

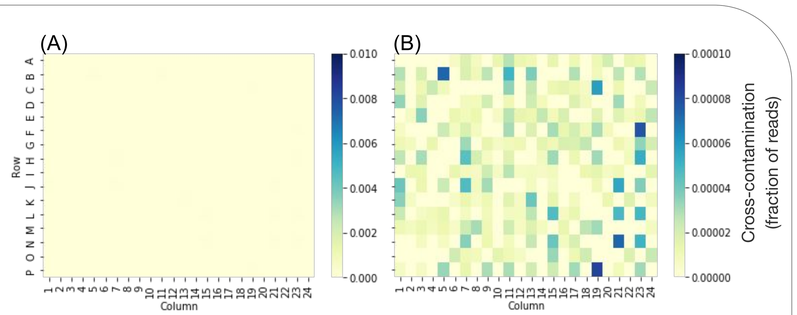


Figure 4 : Purity measurement of UDI sequences in a representative plate. Three plates of unique inserts were prepared using the same UDI plate. Template cross-contamination was controlled for by excluding any cross-contamination that was not consistent between templates in different plates. Cross-contamination was defined as the fraction of filtered reads that contain both UDIs and an incorrect insert (for all three inserts). Figures **4A** and **4B** show the same plate, with different color scales.

Materials and Methods

Unless otherwise noted, all NGS libraries contain human genomic insert generated via Twist Enzymatic Fragmentation Library Preparation kit v2. Individual libraries were pooled by mass and sequenced with a NextSeq 2000 P2 flow cell to generate 2 x 12bp index reads.

Disclosures: All authors are current or former employees of Twist Bioscience