Development of MaxPlex Exome Capture, a High Throughput, Highly Multiplexed Target Enrichment of Human Exome

Steve Oh, Elaine Maggi, Michael Bocek, Leonardo Arbiza, Rebecca Nugent, Derek Murphy, Esteban Toro, Siyuan Chen

TWIST BIOSCIENCE

1. Abstract

Target enrichment through DNA hybridization-capture analysis reduces the costs and complexity of analysis by focusing the interrogation of a DNA library to the specific genomic regions of interest. The development of targeted whole exome capture panels has allowed for sequencing output and analysis to be directed at just 1% of the genome associated with protein coding regions. This reduced burden on sequencing throughput, in combination with sample indexing strategies, has allowed for multiple samples to be pooled on a single sequencing run. With the increase in capacity afforded by the newest generation of sequencing instruments, an increased number of samples can be combined on a single sequencing run, further reducing both costs and processing time. In 2019, to further streamline upstream sample processing, Twist Bioscience introduced modifications to our whole exome sequencing workflows that allowed for single-day multiplexed capture of up to 16 Core Exomes.

Since these innovations, Twist has developed an improved capture system that further expands the multiplexing capability of our target enrichment system. Using a newly optimized capture panel, Exome 2.0, we demonstrate this new capability through the simultaneous capture of **96 genomic libraries**. Here we present data generated on an Illumina NovaSeq 6000 sequencing platform which shows equivalent performance using the MaxPlex capture system when compared to our standard mid-plexed protocols. These results highlight the capability of the MaxPlex capture approach and further demonstrate the potential of a high-throughput whole exome sequencing solution.

2. Workflow

While the Fast Hybridization Target Enrichment protocol allows for exome capture to be performed in a single day, throughput is limited to 16 libraries. 96 and 192 exomes can be pooled and sequenced on the NovaSeq 6000 S2 and S4 flow cells, respectively. To make efficient use of the sequencing capacity of this instrument, increasing the number of hybrid capture reactions would typically be required. However, this would increase the chance for variability between samples as well as the allocation of resources.

Using the Fast Hybridization Target Enrichment protocol as a template, minimal changes were introduced to the overall protocol to create the MaxPlex Target Enrichment; a simplified method to increase throughput and to allow for a more seamless adoption for users already familiar with the midplex Fast Hybridization protocol. Further, the wide range of hybridization incubation times allows compatibility with either single-day or overnight workflows.

Of note, to accommodate the higher volumes resultant from pooling 96 libraries, the time needed to dry down the DNA and hybridization reagents would be longer than that required for midplex captures. However, time may vary depending on the DNA concentrations from the preceding library preparations.

HYBRIDIZATION TARGET ENRICHMENT WORKFLOW (AMPLIFIED INDEXED LIBRARIES)	FAST HYBRIDIZATION (1-16 PLEX) TIME	MAXPLEX (96 PLEX) TIME
STEP 1 **Prepare libraries for hybridization** Indexed library pool — STOPPING POINT	1 hour	2 to 4 hours (Sample dependent)
STEP 2 **Hybridize capture probes with pools** Hybridized targets in solution	3 to 4 hours	Flexible 3 to 16 hours
STEP 3 **Bind hybridized targets to streptavidin beads** Captured targets on beads	1.5 hour	1.5 hour
STEP 4 **Post-capture PCR amplify, purify, and perform QC** Enriched libraries — STOPPING POINT	1 hour	1 hour
STEP 5 **Sequence on an Illumina platform** Libraries ready for sequencing on Illumina platform	-	-

Table 2.1. Protocol Overview. Hybridization target enrichment workflows for MaxPlex (right column) and Fast Hybridization (middle column) are compared to highlight similarities in the two systems.

3. Considerations for Highly Multiplexed Capture

Increasing the number of libraries for a highly multiplexed hybrid capture creates several issues that affect both the quality of the results and the logistics of executing the protocol. Among these considerations, maintaining library complexity and minimizing off-target rates are two main challenges which need to be addressed.

Library complexity can be described as the number of unique molecules found in a given sample. In the context of an NGS library prep, this is a measure of how well a sequenceable product represents the source sample material. As such, decreased library complexity leads to the loss of information and decreased sensitivity of detecting variants within the sample. While diversity of the source genomic DNA defines the ultimate level of complexity, the amount of library input into hybrid capture also plays an important role (**Figure 3.1.A**). It follows that capture inefficiencies would have a negative impact and necessitate increased library DNA input to off-set the loss in sensitivity. The MaxPlex protocol recommends an input of ≥167 ng per indexed library to improve robustness across experimental variability while keeping total volumes low, thereby minimizing the total time required to complete the workflow. **Figure 3.1.B** shows an example in which hidden inefficiencies decrease unique molecules relative to a theoretical yield. While 83 ng input should contain ≅90% of the maximum level of unique molecules (**Figure 3.1.A**), we observe only ≅60%. However, increasing the hybrid capture input mass to 167 ng input returns >90%.

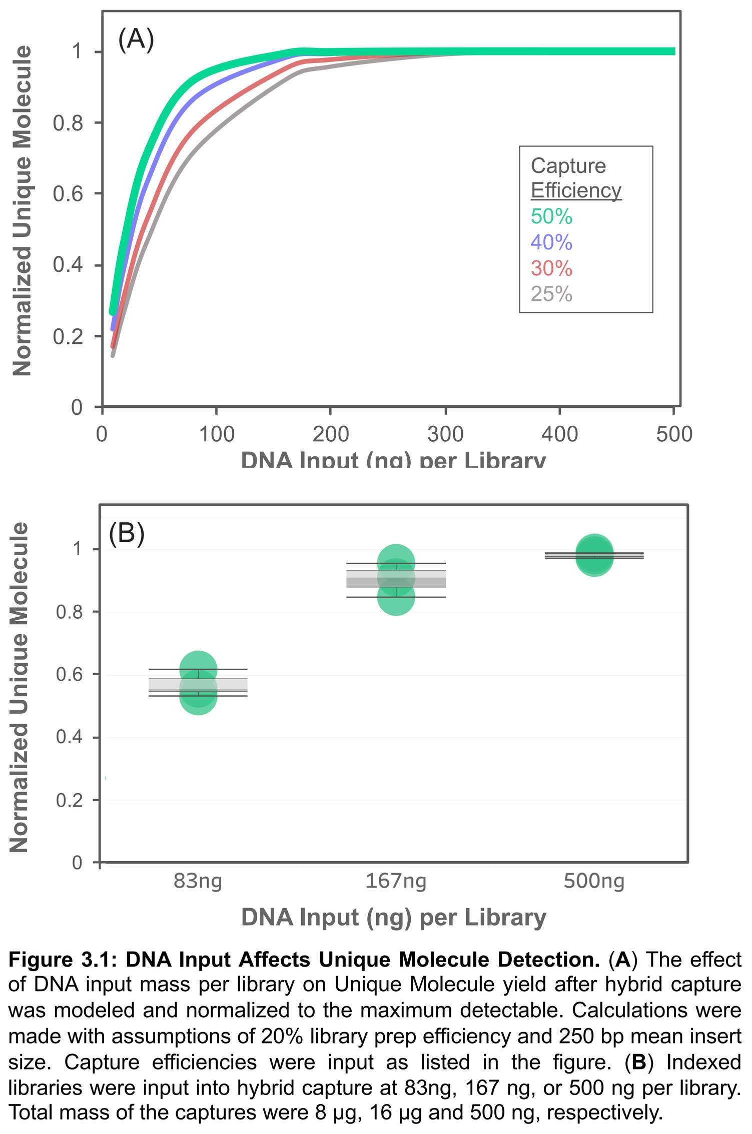


Figure 3.1: DNA Input Affects Unique Molecule Detection. (**A**) The effect of DNA input mass per library on Unique Molecule yield after hybrid capture was modeled and normalized to the maximum detectable. Calculations were made with assumptions of 20% library prep efficiency and 250 bp mean insert size. Capture efficiencies were input as listed in the figure. (**B**) Indexed libraries were input into hybrid capture at 83ng, 167 ng, or 500 ng per library. Total mass of the captures were 8 µg, 16 µg and 500 ng, respectively.

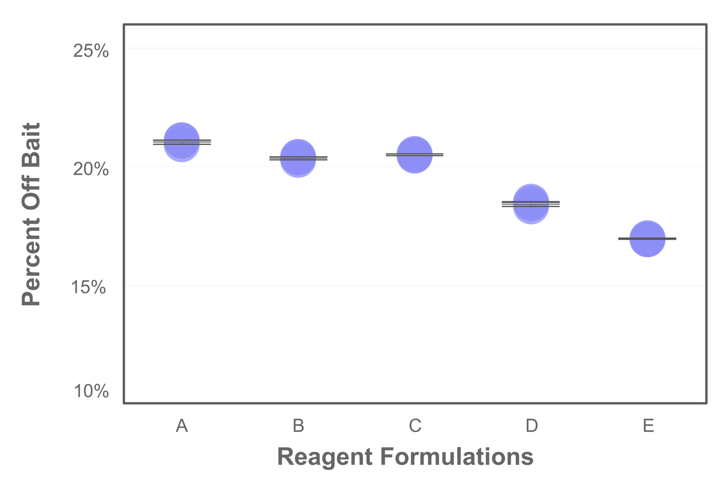


Figure 3.2: Improvements to Off-Target Rates. 96-plex libraries were captured to evaluate reagent reformulations during development. 8 µg of total library DNA was captured under conditions close to the final MaxPlex protocol. 800 kb panels were substituted to evaluate Off Bait rates only. Further refinement was later performed using MaxPlex Exome under final protocol conditions.

One of the fundamental strengths of hybrid capture is to focus the interrogation of a sample to targets of interest and to reduce the use resources on less informative regions. Therefore, off-target reads are undesirable as they increase the time and cost attributed to wasteful sequencing and analysis. By design, MaxPlex combines 96, separate libraries and can occupy a large sequencing space, exacerbating this issue.

For the MaxPlex system, reagents were reformulated to help combat off-target rates. Several iterations of the reagent formulations were tested during development (**Figure 3.2**) and demonstrated improvements of off-target performance.

4. Materials and Methods

Genomic libraries were prepared from purified DNA of NA12878, obtained from Coriell, using the Twist Library Preparation Enzymatic Fragmentation Kit 2.0 in 96-well plates. An equal mass of each of 32 libraries was pooled from 3 separate library prep plates, for a total of 96 libraries per pool, to account for day-to-day variability as well as potential artifacts arising from plate position. Hybrid capture was performed following the MaxPlex Target Enrichment Protocol with the Twist Exome 2.0 MaxPlex kit. To demonstrate flexibility, MaxPlex hybridization was incubated at either short (3 hours) or overnight (16 hours) times. Additionally, the 17 kb Twist Mitochondrial Panel was spiked in to demonstrate the capacity for additional panel customization. Captured libraries were sequenced on the NovaSeq® 6000 sequencer with S4 flowcells to generate 2x100 paired end reads. Manual XP loading was specifically avoided to minimize loading inconsistency. Data was down-sampled to 150x of target size and analyzed using Picard Metrics.

5. Maximizing Sequencing Efficiency with MaxPlex Exome

To compensate for inefficiencies and biases, target enrichment systems commonly require an excess of sequencing effort to retain sample information. In addition to developing MaxPlex-specific reagents, efforts optimizing the Twist Exome 2.0 panel for human (see Poster #530) were combined with the improvements for highly multiplexed hybrid capture. Here we demonstrate that the improvements achieved for midplex captures are translatable to the MaxPlex system

On Target rates of >94% were observed for MaxPlex Exome hybridized for 16 hours and 3 hours as well as MaxPlex Exome + mitochondrial spike-in and performed equivalently to Exome captured in an 8-plex (**Figure 5.1.A**). Fold-80, a measure of uniformity, is the amount of additional sequencing required to raise 80% of bases to the mean coverage. More evenly distributed and well-representative captures will have values closer to '1'. Developments in the MaxPlex system help maintain an even coverage uniformity as shown with Fold-80 values of 1.37-1.38 attained across the different MaxPlex conditions compared to an 8-plex capture (**Figure 5.1.B**).

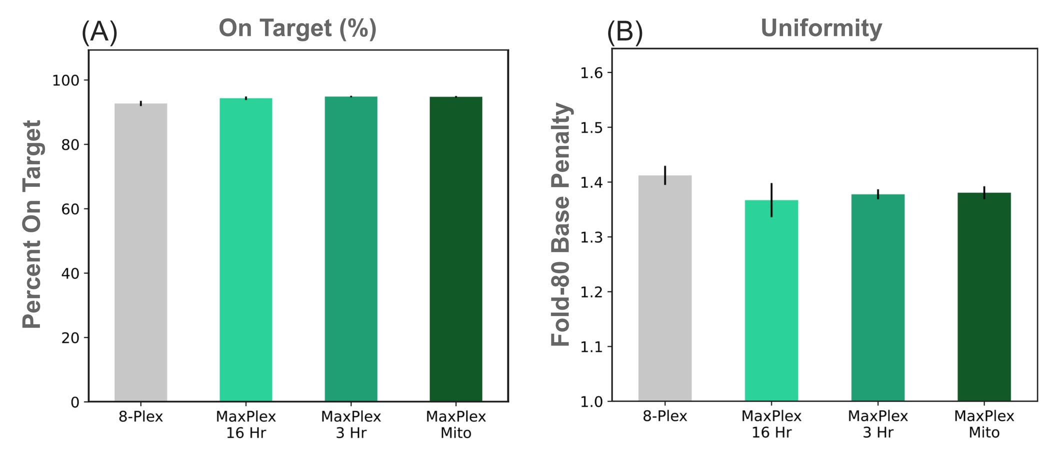


Figure 5.1: On Target Rates and Coverage Uniformity. 187.5 ng of each indexed library was combined for the 8-plex captures while 166.7 ng per library was combined for the MaxPlex captures under the various conditions described the Materials and Methods section. (**A**) Mean values for On Target and standard deviations were calculated for 8 libraries in the 8-plex captures and 96 libraries in the MaxPlex captures. (**B**) Similarly, means and standard deviations were calculated for Fold-80 values using Picard HsMetrics for the MaxPlex conditions described.

Fold-80 provides a high-level, quantitative value of coverage uniformity, it does not fully describe the underlying distribution of reads. For this, probes represented in the sequencing data were organized by GC content and plotted against coverage. A typical profile for midplex is displayed against representative libraries from each of the MaxPlex conditions denoted (**Figure 5.2**). With 3-hour hybridization times, MaxPlex and MaxPlex + mito show a slight bias toward low GC content. This can be rebalanced by increasing hybridization time to 16 hours which provides a layer of tunability for the system. Even with the difference in distributions observed, a high level of uniformity is achieved across all MaxPlex conditions.

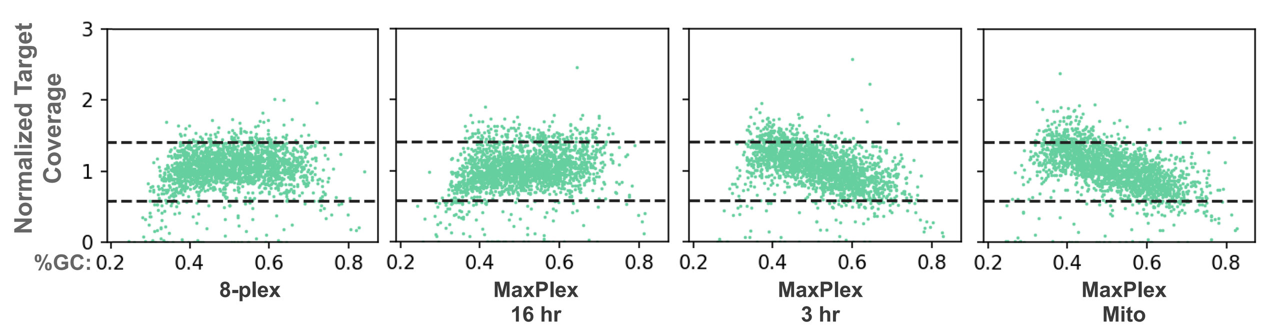


Figure 5.2: Coverage Distribution. Single, representative libraries were chosen for a typical midplex capture and for each of the MaxPlex captures described in the Materials and Methods section. For each sample, baits were binned according to %GC content (X-axis). Relative coverage was normalized to the mean for each bin and plotted on the Y-axis.

6. MaxPlex Exome Capture Efficiency

With a highly multiplexed hybrid capture, the complexity and the bulk of the combined libraries increases. This can potentially impact how useful the resulting capture data will be for downstream analysis. The quality metrics included here show performance of the MaxPlex capture system in this context of data quality. To further define our evaluation of unique molecules, we distinguish between unique molecules, *per se*, as found in the original library prep versus HS Library Size, which is an estimate of unique molecules in the target enriched sample. The latter is important to gauge how much of the original complexity was captured and, subsequently, to decide whether additional sequencing of the enriched library will yield more information. MaxPlex captures retain a high level of library complexity ranging from 87% to ≅100% of that found in midplex captures (**Figure 6.1.A**).

With increased capture inefficiency, enrichment of targeted regions is at risk. To confirm that the ratio of exome over genomic background is consistent with that in midplex captures, averages for fold enrichment are measured against the midplex control and show equivalent, if not improved, performance ranging from 42-44-fold enrichment vs 41-fold for control (**Figure 6.1.B**)

As many of the quality metrics are dependent on reads that align to a reference, the amount of targets which do not get coverage are not directly evaluated and may be susceptible to suboptimal capture conditions. To further dissect performance of the MaxPlex system, rates of non-coverage are compared to midplex control. Overall, all MaxPlex conditions tested show consistent rates of targets with zero coverage at ≅1% (**Figure 6.1.C**).

Importantly, many downstream analysis require a predictable rate of coverage across all targets. While Fold-80 speaks to the uniformity of the population, it does not explicitly reveal a quantitative measurement of coverage. To help predict the output of coverage for a given sequencing input, the percent of target bases represented at various coverage outputs are presented in **Figure 6.1.D**. Here we observe that >95% of target bases are covered at 30X when sequencing reads are downsampled to 150X of the bait space. Further, at 50X coverage, >79% of target bases are represented. This reaffirms the high level of coverage consistency in the MaxPlex captures.

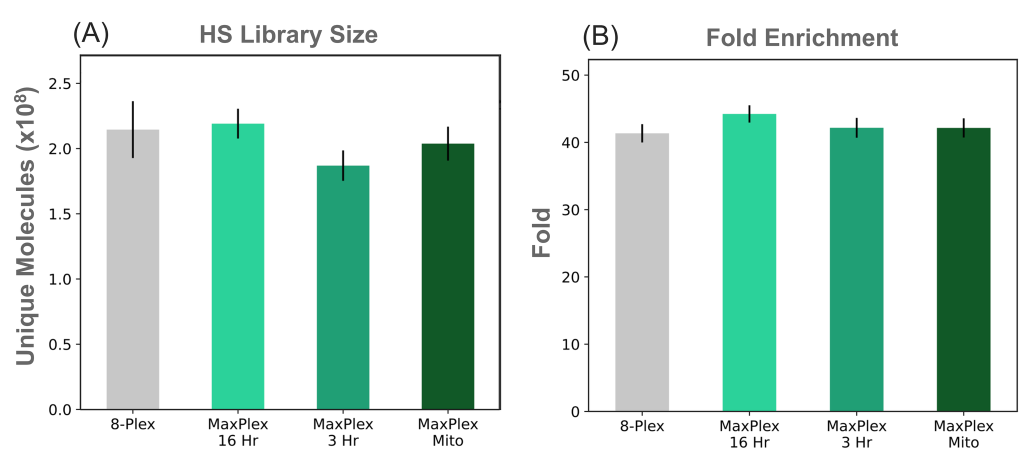
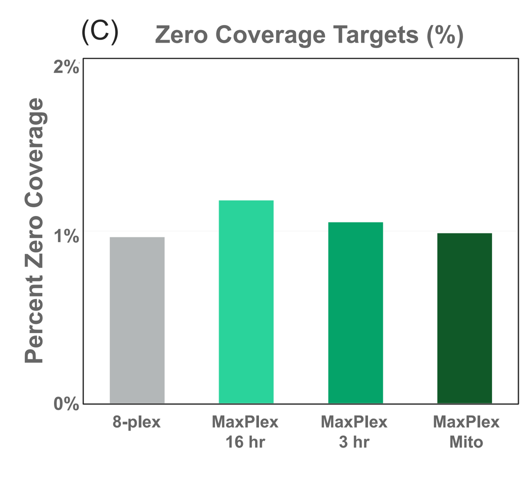
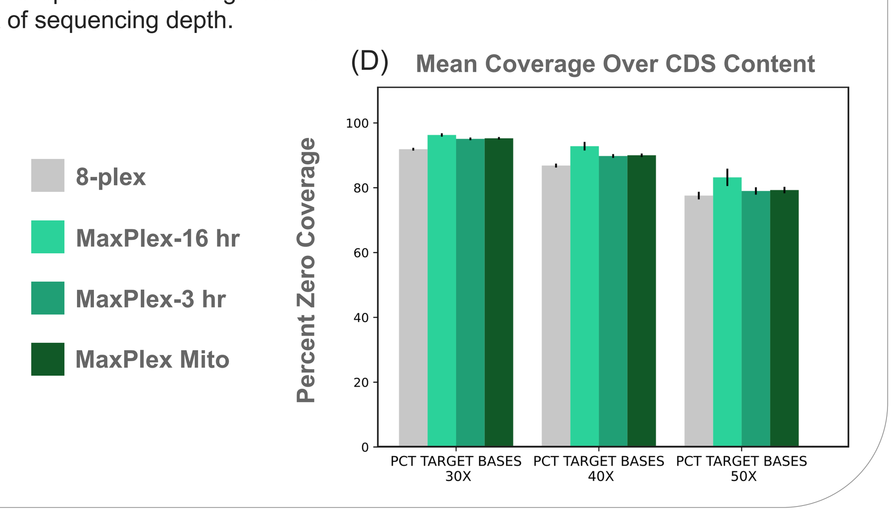


Figure 6.1: Capture Efficiency. Sequencing data for the capture conditions described in the Materials and Methods section were analyzer using Picard Hs Metrics. Mean values are marked in the figures above their respective samples and are averages of 8 libraries, for the 8-plex sample, and of 96 libraries, for the MaxPlex samples. (**A**) Hs Library Size is an estimate of the library complexity available in the target enriched samples. (**B**) Fold enrichment measures the level amplification of the Exome over the background genome. (**C**) Percent Zero Target Coverage is useful in understanding the level of targets that lack coverage. (**D**) Target Coverage reveals how many target bases are at a specified coverage for a given amount of sequencing depth.





7. Sequencing Considerations for MaxPlex Exome

Parameter	Metric
Exome 2.0 Panel Size	≅36.5 Mb
Sequencing Depth	200X
Libraries	96
Minimum Total per MaxPlex Exome	≅700 Gb

Sequencing of Exomes enriched using the MaxPlex system requires loading on to the Illumina NovaSeq 6000 platform or higher.

- qPCR QC: Because patterned flow cells are sensitive to underloading, it is important to QC the library by qPCR to guard against capture variability and to account for the lower number of PCR cycles in the MaxPlex protocol
- Sequencing Depth: Targeting 200X coverage provides some robustness against inter library variability. However, a higher depth may be recommended, if possible, to account for read quality and duplicate rates related to variability in sequencer loading and performance.
- Flow Cell Loading Workflow: Automated flow cell loading is recommended due to the sensitivity of the NovaSeq to artifactual duplicates.
- Service Provider: If relying on a service provider for NovaSeq sequencing, results may vary. In addition to optical duplicate rate, PF quality, and yield, some differences in GC bias have been observed.