Data-Driven Improvements in NGS Target Enrichment Performance

Yehudit Hasin-Brumshtein, Leonardo Arbiza, Kristin Butcher, Siyuan Chen, Hutson Chilton, Richard Gantt, Christina Thompson, Ramsey Zeitoun

1. Abstract

Target enrichment is used in a wide range of applications, from cancer exome sequencing to viral detection. Generation of high performance panels is an involved process that needs to take into account multiple factors, such as GC, sequence content, panel size and production variability. Here we describe the experiments and analysis undertaken to improve our design principles for high-efficiency target enrichment. Our design objective is quantitative optimization of key capture performance metrics. Towards that goal we assessed, both computationally and experimentally, multiple factors: **sequence complementarity, target context, thermodynamics and our production process**. We show that our design process results in high performance first-pass panels, and that for particularly difficult custom panels we are able to improve the performance through our Design-Build-Test-Learn (DBTL) cycle approach. Finally, we show how our panel design principles and data can be combined to address current and emerging target enrichment applications.

2. Understanding Sequencing Efficiency

An efficient target enrichment aims at maximizing coverage of most target bases, while minimizing overall sequencing required to achieve that coverage. In the commonly used Picard pipeline two metrics reflect efficiency of target enrichment: uniformity and off-target rate (Figure 2.1). Quantitative understanding of these metrics, and their interplay, is essential to evaluation of panel performance and thus to robust panel design.

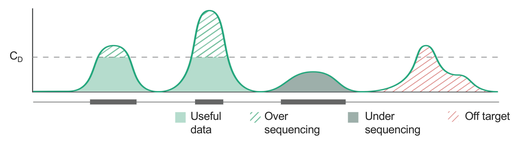

Figure 2.1 Sources of Sequencing Inefficiency. Illustration of sequencing uniformity and off-target. C_D indicates the desired coverage, required for meaningful interpretation. Lack of uniformity and off-target are main sources of "wasted" sequencing effort.

Here we examine how optimization of uniformity and off-target affects target enrichment efficiency, and show that in real-life applications minimizing uniformity will often have a larger impact than minimizing off-target. For example, Table 2.1 shows that to match a capture with fold-80=1.3 and 70% on-target, capture with fold-80=1.7 would need 92% on-target, while capture with fold-80=1.9 will need to exceed the theoretical maximum (>100%).

Panel 1 Fold-80 / On-Target	On-target rate required for panel 2 to match the performance of panel 1, given varying levels of uniformity			
	1.3	1.5	1.7	1.9
1.3 / 70%	70%	81%	**92%**	**102%**
1.5 / 70%	61%	**70%**	79%	89%
1.7 / 70%	54%	62%	**70%**	78%
1.9 / 70%	48%	55%	63%	**70%**

Table 2.1 Equilibrium Points for Off-target and Uniformity Quantitative Impact. The table shows interplay between Uniformity and On-target rate (defined as 1-PCT_OFF_BAIT). Values are required on-target rate to account for differences in uniformity, with panel 1 on-target rate = 70%. Performance equivalence was defined as 80% of bases reach predetermined coverage threshold.

3. First Pass Design Performance

A Design-Build-Test-Learn (**DBTL**) strategy was implemented towards developing a framework for generating reproducibly high-performing panels (Figure 3.1).

This iterative learning approach requires each step to be performed with reproducible results towards building on results of previous iterations. The reproducibility and expected performance of both the **build** and **test** steps of the DBTL system is shown. The reproducibility data is shown for a representative 800 kb panel consisting of roughly 7,400 probes. Replicates were synthesized 1 month apart.

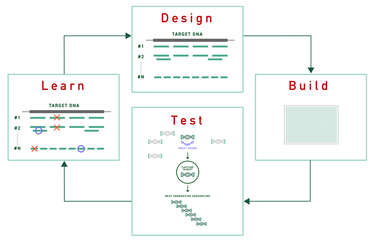

Figure 3.1 Design-Build-Test-Learn: Workflow of design-build-test-learn strategy that was used to generate process for designing a target enrichment system.

Build: An NGS quality control step is performed on every custom panel generated where probe representation is measured post-production. This ensures the process completed as expected and the probe content and representation reflects the intended design. Reproducibility between two panels based on NGS probe counting is high and can support DBTL (Figure 3.2).

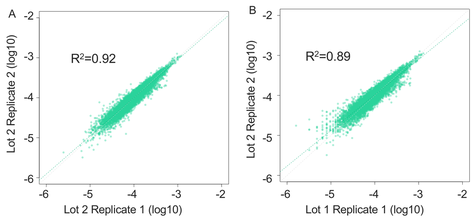

Figure 3.2 Lot to Lot Variability From Build: Each synthesis involves amplification step. A panel containing roughly 7,400 probes (800 kb) was re-synthesized ~1 month apart (Lot1 and Lot2), with two amplification replicates in each Lot (Replicate 1 and 2) **A**) Reproducibility of probe representation within same synthesis, different amplifications. **B**) Reproducibility of probe representation between syntheses.

Test: An NGS target enrichment probe to probe performance was done to ensure reproducible capture and testing of the built panel (Figure 3.3). The overall sequencing HS metrics also showed high concordance between lots.

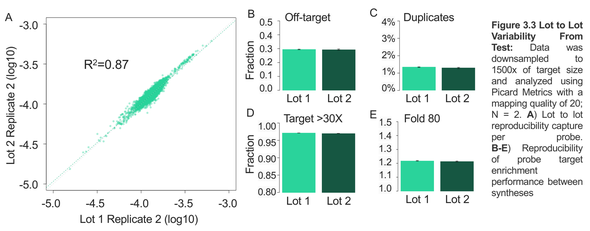

Figure 3.3 Lot to Lot Variability From Test: Data was downsampled to 1500x of target size and analyzed using Picard Metrics with a mapping quality of 20; N = 2. **A**) Lot to lot reproducibility capture per probe. **B-E**) Reproducibility of probe target enrichment performance between syntheses

Following the application of learning and design (see below) the results of this learning was used to design high-performance panels in a first attempt. Six panels ranging from 16 kb to 13 Mb were synthesized and shown to have high coverage metrics (30x coverage) which was made possible by a multivariate optimization of key metrics (Figure 3.4).

A

Panel Name	Size (Mb)	Probes	Genes
Mitochondrial DNA	0.02	139	37
Cancer Hotspot	0.04	384	50
Neurodegenerative	0.6	6,024	118
Cancer + Hotspot	0.8	7,446	127
Actionable Cancer	1.7	19,661	522
Pan-Cancer	3.2	31,002	578
Exploratory Cancer	13.3	135,937	5,442

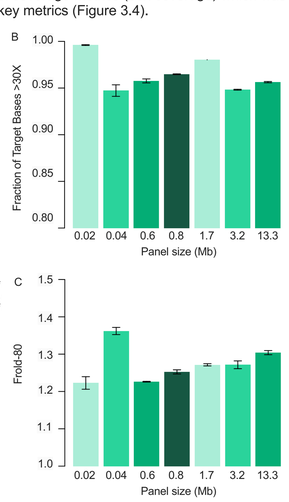

Figure 3.4 First Pass Capture Performance*: Information of capture performance of 6 different panels. **A**) Description of panels and size. **B**) Uniformity (Fold 80) **C**) 30x Coverage performance of each panel as defined by Picard HS metrics.

*Hybrid capture was performed using several target enrichment panels (Twist Bioscience) using 500 ng of gDNA (NA12878; Coriell) per single-plex pool following manufacturer's recommendations. Sequencing was performed with a NextSeq 500/550 High Output v2 kit to generate 2x76 paired end reads. Data was downsampled to 150x of target size and analyzed using Picard Metrics with a mapping quality of 20; N = 2.

4. Improving Capture Uniformity for Custom Use Cases

Probes can be designed and balanced based on iterative datasets to reproducibly generate uniform capture panels on a first try (above). However, assumptions imbedded in our balancing algorithms may not faithfully represent experimental conditions of custom or non-standard panels. In order to address this, balancing can be custom addressed by applying the same theoretical strategies used to initially design highly uniform capture. This allows efficient sequencing by a wide-range of applications with lower technical risk.

This rebalancing approach is demonstrated on two panel types – a custom panel and spike-in panel of two sizes.

A standalone custom panel containing roughly 13,000 probes was rebalanced. This panel initially showed a slight GC bias that was remedied with resynthesis of a rebalanced panel. Based on HS metrics we see this effect is most prominent on the coverage of all target regions (30x coverage) and AT dropout biases (Figure 4.1).

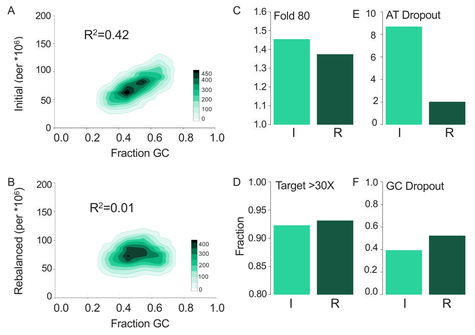

Figure 4.1 Standalone Capture Panel Rebalancing. A,B) Basecalls overlapping a probe calculated using BEDTools versus probe GC% **C-F)** Key Picard performance metrics of **Initial** (I) and **Rebalanced** (R) panels, calculated using **minimum** mapping quality of 20.*

A spike-in custom panel (added content to the exome) containing roughly 250,000 probes was also rebalanced. This panel initially showed a major GC bias that was remedied with resynthesis of a rebalanced panel. Based on HS metrics we see this effect is most prominent on the coverage of all target regions (20x coverage) and uniformity (Figure 4.2).

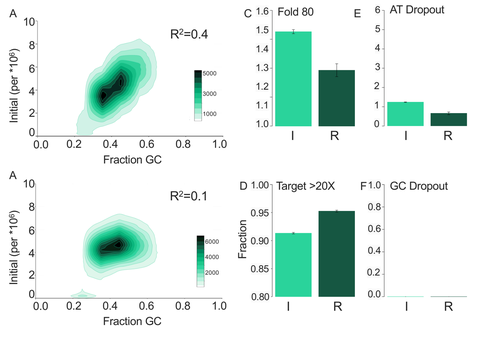

Figure 4.2 Spike-In Panel Rebalancing. A,B) Base calls overlapping a probe calculated using BEDTools versus probe GC% **C-F**) Key Picard performance metrics of **Initial** (I) and **Rebalanced** (R) panels, calculated using **minimum** mapping quality of 20. *

5. Using 30,000 Probes to Understand How Mismatches Affect Capture

Hybridization capture tolerates some mismatches and a probe will capture a range of sequences that are sufficiently similar to the perfect target - with varying efficiency. Depending on the application, this may be a desired property, or undesired hindrance. In either case quantitative understanding of the effect of mismatches on capture is important for optimizing probe design, and large scale data are lacking. To examine the effects of number and distribution of mismatches on capture efficiency we designed and synthesized two panels, Control and Variant. The Control panel contained probes selected from the Twist human exome panel that perfectly match the human genome reference. The Variant panel contained the same probes but with 1-50 mismatches distributed at random, or as one continuous stretch. (**Figure 5.1**)

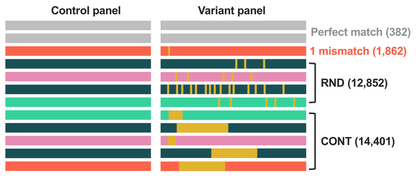

Figure 5.1 Experimental Design of Control and Variant Panels Each panel (Variant and Control) contained 28,794 probes. In the control panel, the probes were designed to be complementary to their targets. In the variant panel 1-50 mismatches (yellow) were introduced either randomly along the probe (RND) or all together in a single continuous stretch (CONT). Also, 382 control probes without mismatches were added to both panels for normalization (in grey), thus the Control and Variant pools contained a total of 29,176 probes.

Our data suggests that distribution of mismatches is of paramount importance for capture efficiency: for example probes with 50 mismatches arranged in one continuous stretch capture as well as probes with 10-15 mismatches distributed randomly, while probes with 50 mismatches distributed randomly were completely ineffective. (**Figure 5.2**).

We also show that other factors such as GC, length of perfect match and hybridization temperature (**Figure 5.3**) modulate capture efficiency in the presence of mismatches.

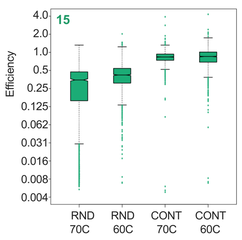

Figure 5.3 Effect of Temperature on Capture Efficiency in the Presence of Mismatches* Effect of decreasing hybridization temperature on capture efficiency of probes with 15 mismatches distributed either randomly (RND) or in a continuous stretch (CONT) across the probe.

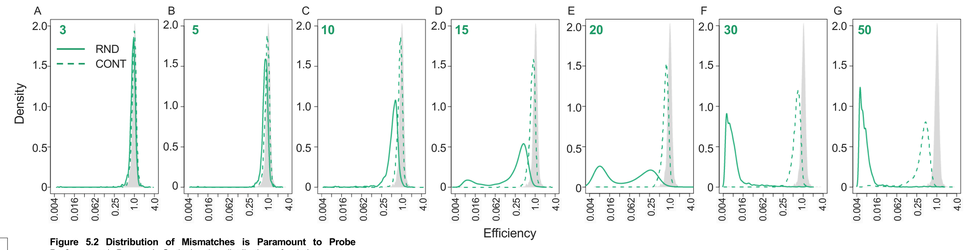

Figure 5.2 Distribution of Mismatches is Paramount to Probe Performance* Panels A–G depict the distribution of relative capture efficiency for probes with a single mismatch (gray) and probes with multiple mismatches (green lines; the number of mismatches is indicated in the left top corner). Solid line depicts the distribution for probes with randomly distributed mismatches (RND), and the dotted line indicates the distribution for probes with continuous mismatches (CONT).

This data can be used to address applications, like the capture of mixed metagenomic samples and the capture of bi-sulfide treated gDNA towards predicting pitfalls that require augmented probe design (**Figure 5.4**).

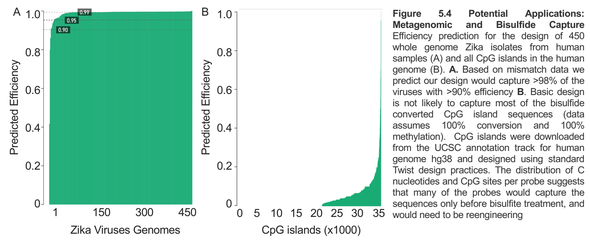

Figure 5.4 Potential Applications: Metagenomic and Bisulfide Capture Efficiency prediction for the design of 450 whole genome Zika isolates from human samples (A) and all CpG islands in the human genome (B). **A.** Based on mismatch data we predict our design would capture >98% of the viruses with >90% efficiency **B**. Basic design is not likely to capture most of the bisulfide converted CpG island sequences (data assumes 100% conversion and 100% methylation). CpG islands were downloaded from the UCSC annotation track for human genome hg38 and designed using standard Twist design practices. The distribution of C nucleotides and CpG sites per probe suggests that many of the probes would capture the sequences only before bisulfite treatment, and would need to be reengineering

6. Fine-Tuning Probe Specificity for Downstream Applications

Parameters that can affect target enrichment are sequence properties and genomic structure, where repeats, low complexity sequences, and other features can lead to uneven capture and off target.

Twist's TE system has been engineered to achieve extremely high uniformity, and off target capture can in principle be addressed by varying levels of stringency with which problematic sequences are avoided.

However, developing a good way of controlling stringency is not straightforward. A common strategy is to provide the option to filter designs by a variety of elements such as those identified by repeatmasker and different programs for detecting intragenomic homology. The problem with this heuristic approach is the uncertainty of its effects, often leading to the removal of probes that do not affect off target while missing others that will contribute significantly to the level of promiscuous capture (**see Mismatch study in Section 5**).

In addition to eliminating the guesswork out of including specific regions among targets and allowing fined grained stringency control, enabling informed design decisions, Twist methods have a variety of design optimization applications including minimal filtering and 1st pass optimizations in a variety of TE applications from genotyping to CNV detection.

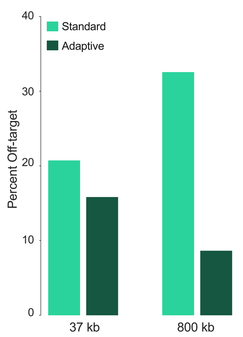

Figure 6.1 Experimentally Driven Adaptive Designs. One very powerful approach to eliminate the guess work in tuning filtering stringency is by the use of adaptive designs, where experimental results from a 1st pass design, are used to determine sequences that should be removed with great precision (**Fig 6.1**). Examples of improvements after a single pass adaptive design for moderate and aggressive off target reduction in panels with challenging target regions (respectively 37Kb and 800Kb, 3 probes and ~4% of probes removed).

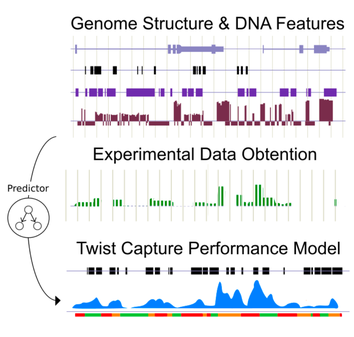

Figure 6.2 Modeling Adaptive Designs 1st pass designs and other design optimization tasks would strongly benefit if running experiments could be avoided. To do this we have combined a variety of DNA sequence and genomic structural features with experimental data to build a predictive model of capture performance (**Fig. 6.2**).

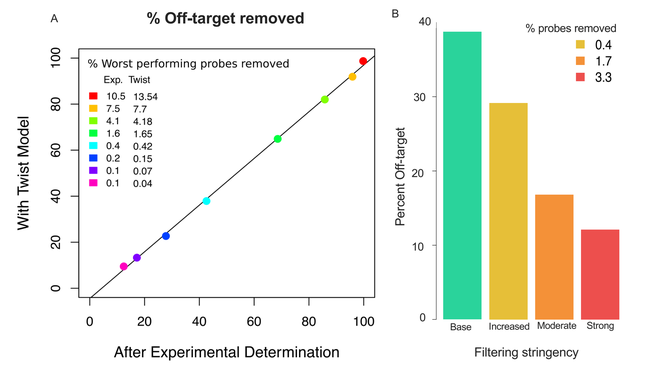

Figure 6.3 Adaptive Designs Without Experimental Data. Twist's capture model eliminates the guesswork associated with bait filtering allowing for adaptive-like designs without running experiments. **A**) The graph shows the level off target predicted by our model compared to that measured by experimentation (axes) and the fraction out of the total number of baits required in each case to achieve it. **B**) The graph shows results analogous to those in **Fig. 6.1**, for a custom design against a particularly hard set of target regions, various levels of stringency, and the effectiveness of bait removal based on our model.*

Learn more about the new Twist Exome and Custom Target Enrichment solutions at twistbioscience.com/products/ngs