The Effects of Mismatches on DNA Capture by Hybridization

Yehudit Hasin-Brumshtein, Leonardo Arbiza, Kristin Butcher, Hutson Chilton, Ramsey Zeitoun

INTRODUCTION

Targeted next-generation sequencing (NGS) focuses the sequencing effort on specific region by capturing and enriching those regions with complementary biotinylated probes. Though many factors can influence the efficiency of such capture, a primary consideration is how well the DNA sequence of the probe matches the target (sequence complementarity), as this affects both the efficiency and selectivity of capture.

Many target enrichment applications aim to isolate sequence variants, which can range in complexity from single nucleotide polymorphisms (SNPs) to the full range of sequences found in all the organisms in an ecological sample (metagenomics). For capture hybridization and targeted sequencing, probes are designed around a reference genome that is assumed to represent the targeted sequence in a sample. The performance of genomic analysis of human samples, for example, is often evaluated with the human NA12878 genome, a human reference sample with an established set of high-confidence genomic variants which was sequenced to an excess of 500x with different technologies and protocols (Zook et al. 2014, 2018; Eberle et al. 2017). The NA12878 genome contains regions with up to 22 variants per 120 nt, but most regions of this size harbor only 0–1 variants (Havrilla et al. 2019).

From the high-confidence variant calls for the NA12878 genome, we estimate that ~8% of the probes in the Twist exome designed to hg38 would have one mismatch with respect to the actual sequence, and ~1% would have multiple mismatches (**Figure 1A**). A quantitative understanding of the effects of those mismatches would be useful for optimizing probe design. For example, probes can be constrained to contain no more than a few mismatches, or specific genomic regions could be identified that might require probes designed for sensitive and selective capture of a variant. This concept can also be extended to predict the effectiveness of capture in mixed samples, such as in metagenomics applications where viral and bacterial genomes are expected to evolve continuously. For probes to be of use in these applications, some tolerance to sequence mismatch is expected.

To understand the factors that might influence probe design for these types of applications, we examined the effects of sequence complementarity (the number and distribution of mismatches) on capture efficiency in target enrichment. Though the effects of mismatches on capture efficiency have been described previously (Ke et al. 1993, Gotoh et al. 1995, Piao et al. 2008, Naiser et al. 2008), these studies focused on one or a few short (20–30 nt) sequences and examined only one or two mismatches in isolation. Target enrichment, however, usually involves thousands of longer probes (~120 nt). To examine the effects of mismatches on capture efficiency in a setting relevant to target enrichment, we synthesized and performed target enrichment with two large probe panels (~10% of the exome): one with 120-nt probes that matched the reference and sample, and one with 120-nt probes that had mismatches introduced by design (**Figure 1**). Our findings show that the number and spacing of mismatches have significant effects on capture efficiency and that these effects can be modulated by probe GC content and hybridization temperature.

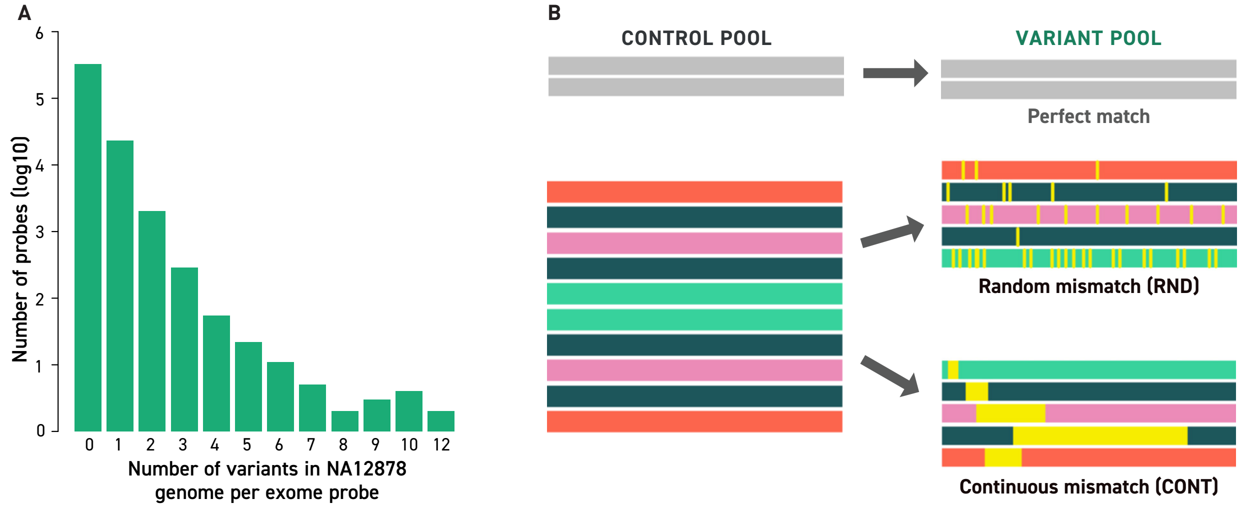

Figure 1. Probe design. A. Using the NA12878 high-confidence variant set (see text for details) and Twist exome probes, we found the NA12878 genome contained 0–12 mismatches per 120-nt exome probe (hg38). **B.** Each probe panel (Variant and Control) contained 28,794 probes. The Control probes were designed to be complementary to their targets. In Variant probes, 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; thus the Control and Variant panels contained a total of 29,176 probes.

STUDY DESIGN

We designed and synthesized two panels, each containing 28,794 120-nt probes covering 3.4 Mb of the human exome: the Control panel contained probes randomly selected from the Twist human exome panel, and the Variant panel contained the same probes but with varying numbers (1–50) and distributions (random, or as one continuous stretch) of mismatches (**Table 1** and **Figure 1B**). For normalization, 382 probes with no mismatches were also added to each panel for a total of 29,176 probes.

We evaluated the performance of both panels in a target enrichment experiment using NA12878 genomic DNA (gDNA) with Twist's standard 16 hour protocol that includes a 70°C hybridization temperature.

RESULTS

Multiple mismatches affect capture efficiency, but single mismatches do not

Single nucleotide polymorphisms (SNPs) are natural variations that occur at a single position in a sequence. The most common source of mismatch between probes and targets, they are virtually unavoidable when designing probes for capture from a population. Multiple mismatches between probe and target, on the other hand, occur in variant-rich regions, in mixed samples, or when multiple genomic regions differ by only a few nucleotides. Whether a particular application aims to maximize or minimize the capture of imperfectly matching targets, it is useful to quantify the bias introduced to capture by single and multiple mismatches.

Figure 2 shows the results of a comparison of the capture efficiency of a population of probes with and without single mismatches. The results demonstrate that a single mismatch does

# MISMATCHES	# PROBES		
	CONTROL	RANDOM	CONTINUOUS
0	382	–	–
1	–	1862	
3	–	1772	1971
5	–	1800	2090
10	–	1777	2033
15	–	1826	2109
20	–	1827	2044
30	–	1794	2126
50	–	1735	2028
TOTALS	29,176		

Table 1. Number and distribution of mismatches per probe in the Variant probe panel. The Variant probe panel contained 120-nt probes that either matched their target sequences perfectly (Control) or had varying numbers of mismatches distributed randomly along the probe (Random) or along one continuous stretch (Continuous).

not affect average capture efficiency (**Figure 2A**), independent of the nucleotide substituted (**Figure 2B**) or the position of the mutation within the probe (**Figure 2C**).

Figure 3 shows that when multiple mismatches occur, their distribution has a significant impact on capture efficiency: a long stretch of mismatches does not reduce efficiency as much as mutations that are randomly distributed along the probe. For example, the efficiency of capture by probes with a continuous mismatch of 50 nt is 0.4 (**Figure 3H**), and probes with only 15 mismatches randomly distributed along the probe yield an efficiency of 0.33 (**Figure 3E**). Probes with 30 or 50 randomly placed mismatches are completely ineffective (mean efficiencies of <0.01, **Figures 3G, H**).

Interestingly, the decay in efficiency for probes with randomly distributed mismatches is bimodal — the probe efficiency breaks, rather than decaying gradually (**Figure 3E, F**). Moreover, the breakpoint depends on the GC content of the probe and the number of mismatches — of the probes with 15 mismatches, only those with the lowest GC (<40%) show an average efficiency of <2%, but with 20 mismatches, probes with <50% GC are ineffective. The reduction in efficiency from a continuous stretch correlates linearly with the length of mismatch and so can be extrapolated to predict the efficiency for any continuous mismatch of 1–50 nt. In contrast, the decay in efficiency for probes with randomly distributed mismatches is more rapid and less predictable (**Figure 3I**).

Length of complementarity, GC content, and hybridization temperature modulate capture efficiency

Other factors that modulate capture efficiency include the length of complementarity between the probe and target (Öhrmalm et al. 2010), the GC content of the probe, and the temperature at which hybridization is performed (Liu et al. 2007). We examined how these factors mitigate or exacerbate the effects of mismatch by repeating our experiment with hybridization at 60°C and grouping the probes into bins of GC content or length of longest perfect match. We show the effects of these modulating factors on probes with 15 mismatches because these showed an intermediate reduction in mean capture efficiency and a large, homogeneous dynamic range (**Figure 4**). The conclusions, however, are applicable to other numbers of mutations, as well.

Figure 4 shows that overall, these three factors had a greater impact on hybridization with probes with randomly distributed, singular mutations than on capture with probes with a continuous 15-nt mismatch. For probes with 15 random mismatches, reducing the hybridization temperature from 70°C to 60°C did not affect median capture rates, but it did improve capture by probes that performed poorly in 70°C, reducing overall variability (**Figure 4A**). Including longer stretches of complementarity had a similar effect, in that it yielded more consistent hybridization results but did not affect median capture rates (**Figure 4B**). The effect of probe GC content, however, was the most significant: GC content affected both the median capture rate and variance. Probes with randomly distributed mismatches and low GC content (35–40%) were approximately 50 times less efficient than those with high GC content (55–60%) (**Figure 4C**).

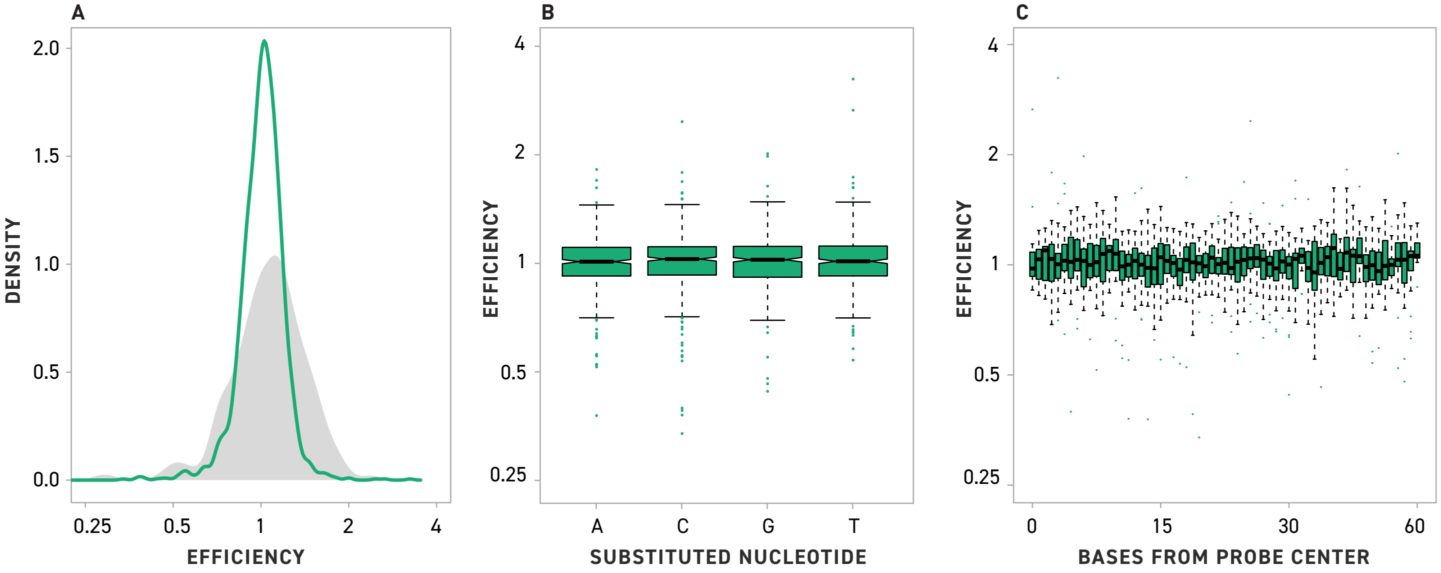

Figure 2. Single mismatches are the most frequent, but they do not affect capture. A. Relative efficiency of hybridization using fully complementary probes (no mismatches, gray) and probes containing a single mismatch (green solid line). Notably, the distribution for fully complementary probes was wider than that for the single mismatch probes. This is because the distances between most of the fully complementary probes were in close proximity (<250 nt) to other probes, while all single mismatch probes were designed to be at least 250 nt apart from other probes **B**, **C.** Boxplots of the capture efficiency of probes with a single mismatch as a function of the identity of the mutated base (B) or the distance from the probe center (C).

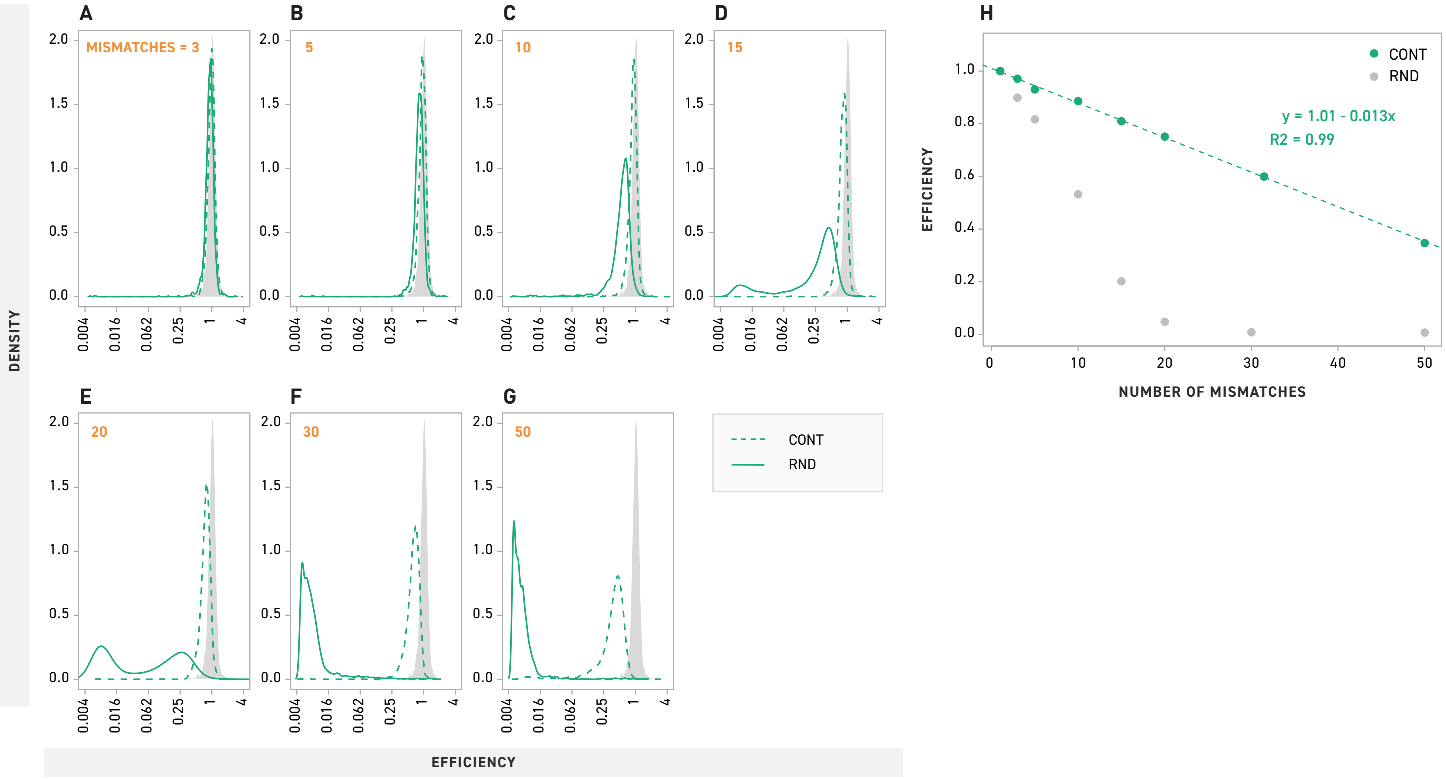

Figure 3. The effects of randomly distributed versus continuous stretches of mismatches. Panels A–G depict the distribution of relative capture efficiency for probes with a single mismatch (gray) and probes with multiple mismatches (green lines; the number of mismatches is indicated in the left top corner). In each panel, the 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). Panel H plots median probe efficiency (y-axis) as a function of the number of mismatched nucleotides (x-axis). Probes with continuous mismatches (CONT, green dots) follow a linear pattern (R^2=0.99), whereas the efficiency of probes with randomly distributed mismatches (RND, gray) drops rapidly.

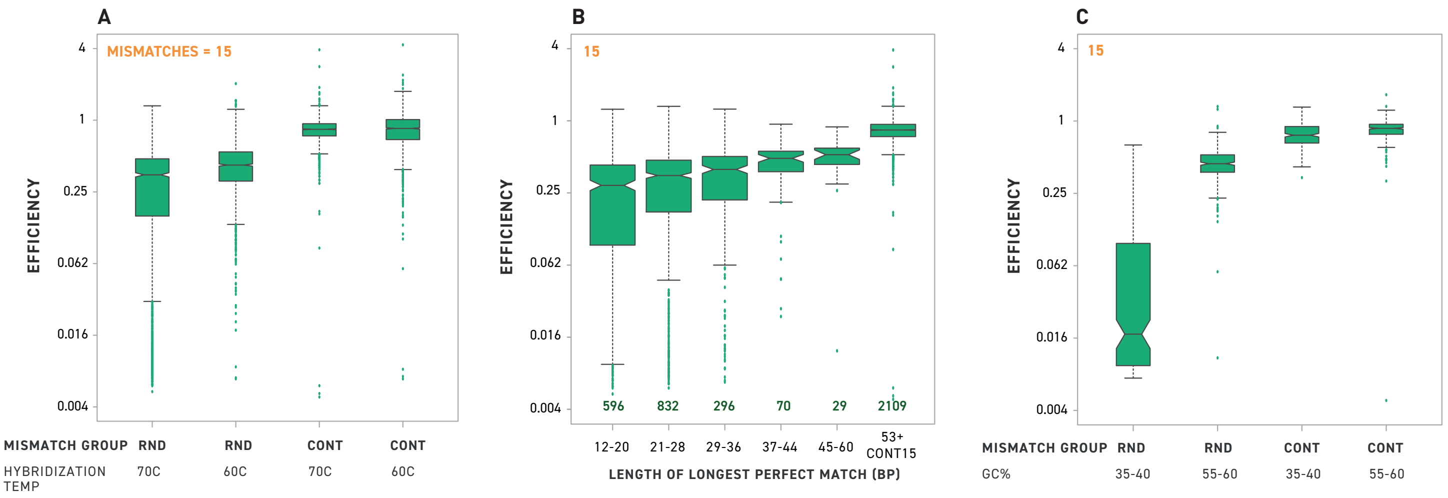

Figure 4. Effects of hybridization temperature, length of the longest perfect match, and GC content on capture efficiency. Panels A–C show the effects of hybridization temperature (A), length of the longest perfect match (B), and GC content (C) on capture efficiency of probes with 15 mismatches distributed either randomly (RND) or in a continuous stretch (CONT) across the probe.

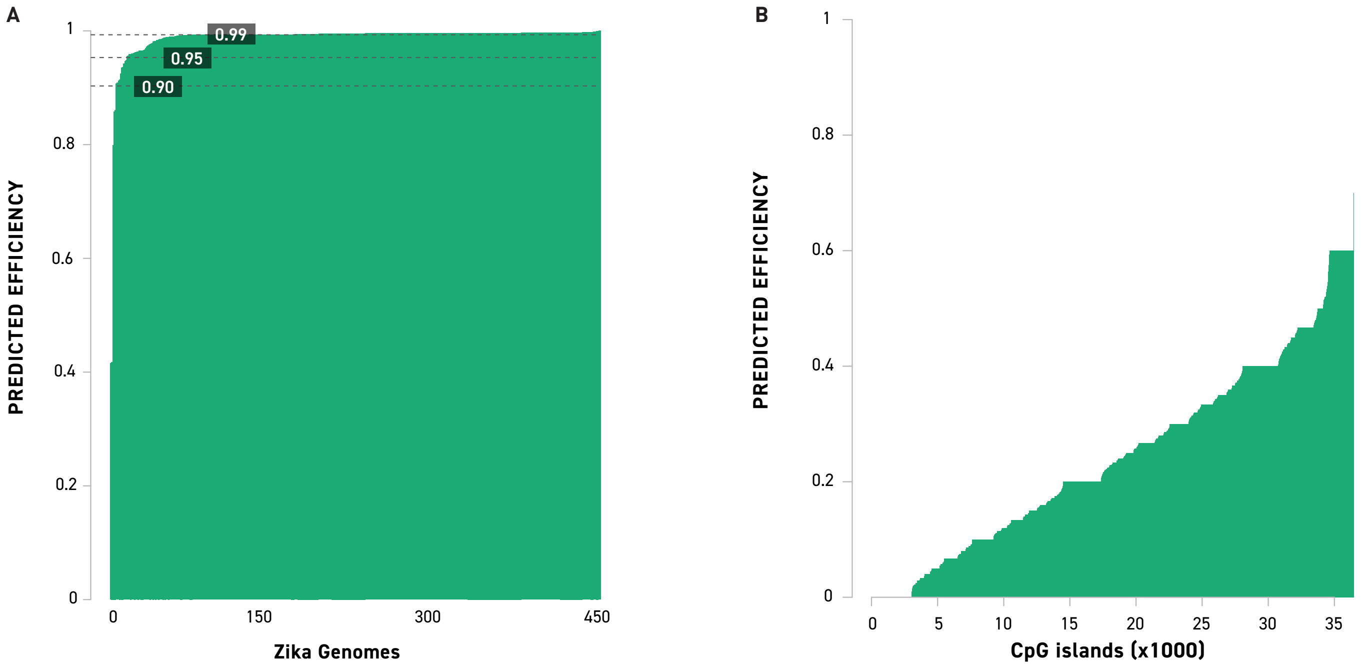

Figure 5. Potential applications of mismatch data to probe design. A. The efficiency prediction for the design of 450 whole-genome Zika isolates from human samples indicates >98% of the viruses would be captured with >90% efficiency. **B.** CpG islands downloaded from the UCSC annotation track for human genome hg38 and designed using standard Twist design rules, which assumed all regions are converted with bisulfite treatment. The distribution of C nucleotides and number of CpG sites per probe suggest that many of the CpG islands would not be captured with probes designed for fully converted regions when methylated.

DISCUSSION

We investigated the effects of probe-target sequence mismatches on target enrichment performance. Expanding on work by other groups, our study applied a large number of probes to enable a thorough exploration of the relative influences of both the degree and type of mismatch, as well as the context and position along the sequence in which they occur. Our results demonstrate that certain aspects of probe mismatches do impact capture efficiency, and these insights can be applied to the design of probes to optimize capture efficiency in a range of applications.

Single mismatches did not affect capture performance by our 120-nt probes, a result that should reassure those using target enrichment to capture SNPs. Multiple mismatches, however, negatively impact capture in a mismatch distribution and probe GC-dependent manner. The most dramatic aspect of our results was the high tolerance of long, continuous mismatches (up to 40% of the probe length) as opposed to the rapid decline for distributed variants.

Potential applications

Together with the sequence information available in genomic databases, our data can help guide probe design for applications focused on capturing genomic diversity (**Figure 5**). For example, metagenomics or clinical applications often aim to detect and characterize the proportions of a variety of microorganisms in a single sample. Using mismatch data, one could evaluate which sequences (and therefore, which subtypes of microorganisms) would be efficiently captured with particular probe designs and identify those sequences that might require additional probes to augment their capture.

Figure 5 shows a simulation of information for two types of targets, all full-length genomes of the Zika virus isolated from human samples (obtained from NCBI viral variation database, taxid 64320, **Figure 5A**) and all CpG islands in the human genome (obtained from UCSC table browser, Figure 5B). In both cases, we designed probes to match a reference genome sequence and then calculated the expected number of mismatches per probe for all other imperfect targets, assigning an expected efficiency for each probe-target pair. We then calculated the average efficiency per target as mean efficiency of all probes covering it.

In some cases, such as for the Zika virus (**Figure 5A**), the basic design captures >90% of the Zika subtypes with >98% efficiency. In contrast, bisulfite-treated human CpG islands would have on average too many mismatches and require optimization of probe positioning as well as augmentation of the design with methylation-specific probes **(Figure 5B**). The ability to quantitatively predict capture efficiency in mixed sample scenario, allows for rational and effective panel design for these complicated sets.

Conclusion

Target enrichment relies on the ability to hybridize probes with samples that do not have perfect sequence complementarity. Having more clarity on the behavior of capture efficiency with respect to mismatches in the context of your capture can significantly influence design strategies. We have used the results of this study and expanded on similar studies to improve assay sensitivity, reduce off-target rates, and build robust probe design approaches. In addition, having an accessible process and panels in place to evaluate capture efficiency expedites optimization of novel applications and standard implementations.

METHODS

Target Enrichment and Sequencing

A gDNA library was prepared from 50 ng NA12878 template gDNA (Coriell Institute) using the Twist Library Preparation EF Kit (96 samples, 101058) with full-length combinatorial dual index TruSeq-compatible Y-adapters (Illumina) according to the Twist Bioscience Library Protocol. All samples were sequenced on an Illumina NextSeq instrument.

Data Analysis

All samples were sequenced with 2x151nt reads. Fastq read files were curated with FastQC and aligned to the hg38 genome reference without the alternative chromosomes using BWA-mem with default parameters. Reads that overlapped each of the probes were counted with BEDtools coverage. All downstream analysis was carried out using custom R scripts: individual counts were scaled to total library size and then log2-transformed. GC effect was modeled on the Control panel using the polynomial model *GCfit = lm(y~poly(x,3))*, where y is mean probe count across samples, and *x* is GC content of the probe. That model was then applied to both the control and the Variant panel, and its residuals were used in further analysis.

After accounting for GC effect, probes counts were normalized to the mean counts of probes with a single mismatch. The single mismatch probes performed similarly to those with no mismatches, and the number and GC distribution of probes with a single mismatch represented the number and GC spectrum of probes in the examined groups more closely. The GC distribution of probes with no mismatch, on the other hand, was skewed toward lower GC and the number of probes in that category was significantly lower.

REFERENCES

Eberle MA, et al. (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Research 27: 157–164

Gotoh M, Hasegawa Y, Shinohara Y, Shimizu M, Tosu M (1995) A new approach to determine the effect of mismatches on kinetic parameters in DNA hybridization using an optical biosensor. DNA Res 2: 285–293

Havrilla JM, Pedersen BS, Layer RM, Quinlan AR (2019) A map of constrained coding regions in the human genome. Nature Genetics 51: 88–95

Ke SH, Wartell RM (1993) Influence of nearest neighbor sequence on the stability of base pair mismatches in long DNA; determination by temperature-gradient gel electrophoresis. Nucleic Acids Res 21(22): 5137–5143

Liu F, Tøstesen E, Sundet JK, Jenssen T-K, Bock C, Jerstad GI, Thilly WG, Hovig E (2007) The human genomic melting map. PLOS Computational Biology 3: 874–886

Naiser T, Kayser J, Mai T, Michel W, Ott A (2008) Position dependent mismatch discrimination on DNA microarrays – experiments and model. BMC Bioinformatics 9: 509

Öhrmalm C, Jobs M, Eriksson R, Golbob S, Elfaitouri A, Benachenhou F, Strømme M, Blomberg J (2010) Hybridization properties of long nucleic acid probes for detection of variable target sequences, and development of a hybridization prediction algorithm. Nucleic Acids Res 38:e195

Piao X, Sun L, Zhang T, Gan Y, Guan Y (2008) Effects of mismatches and insertions on discrimination accuracy of nucleic acid probes. Acta Biochim Pol 55:713–7s20

Zook J, et al., (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls Nature Biotechnology volume 32, pages 246–251 (2014)

Zook J, et al., (2018) Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. https://www.biorxiv.org/content/10.1101/281006v2
doi: https://doi.org/10.1101/281006

These Products are subject to certain use restrictions as set forth in Twist's Supply Terms and Conditions www.twistbioscience.com/supply-terms-and-conditions