Capture-based SNP Genotyping with Twist Target Enrichment Panels

ABSTRACT

Arrays have long been the go-to method for the large scale genotyping of single nucleotide polymorphisms (SNPs). This application note demonstrates how Twist Custom Target Enrichment Panels can be designed for the identification of hundreds of thousands of markers by NGS. Variant calling performance is evaluated using genomic genotyping standards and compared directly with arrays, demonstrating accurate genotyping with minimal bias. SNP and indel genotyping can now be performed on the same platform as whole-exome sequencing, reducing costs, time, and effort.

INTRODUCTION

In the past two decades, genotyping arrays have been instrumental in the large-scale characterization of single nucleotide polymorphisms (SNPs) and the genetic makeup of individuals. This key technology has advanced our understanding in diverse areas: from evolutionary genomics, and heritable and complex disease, to personalized genomics and medicine. In recent years, reductions in next-generation sequencing (NGS) cost has made the technology an attractive option for genotyping. NGS expands our ability to genotype beyond SNPs into detection of multi-allelic sites, insertions, deletions, and other structural variants by providing the full sequence information around a variant. NGS also brings increased flexibility compared to the fixed template format of arrays, as probes don't have to be designed for specific variants and genotypes which may or may not actually be present in a certain sample.

However, targeted sequencing has yet to fully replace microarrays due to barriers associated with performance at scale. For this reason, exome sequencing and array-based genotyping are often run independently, as separate workflows for the same samples, to obtain full variant information.

In this application note, Twist's Custom Panel design algorithms are leveraged to generate a ~240,000 SNP target enrichment panel for genotyping by sequencing. Twist Custom Panels can be designed and built to cover a wide range of panel sizes, target regions, and multiplexing requirements all with exceptional and consistent performance. Previously we have shown that our target enrichment panels tolerate mismatches to bait sequences with small reductions in capture efficiency (**Figure 1**). We evaluate panel performance against results from matched array-based genotyping using genomic genotyping standards and show precision and sensitivity for variant calls > 99%. We also carefully evaluate biases, such as GC context, reference allele bias, and applicability to different populations, and show accurate genotyping with minimal bias. In summary, we demonstrate a unified workflow to merge genotyping with exome sequencing, which leads to considerable savings compared to running each individually.

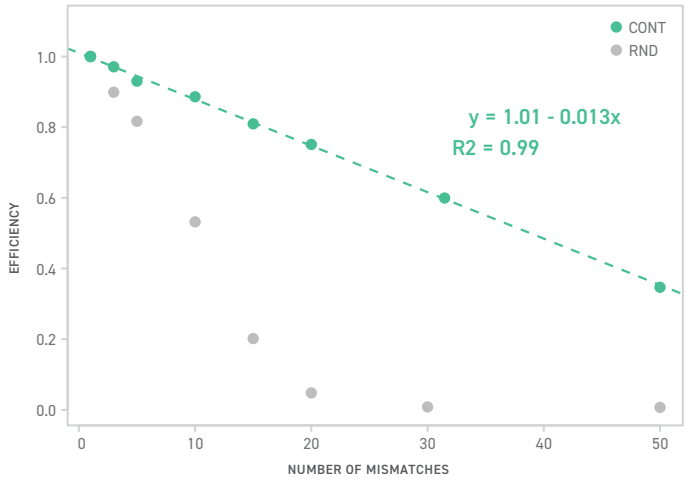


Figure 1: Capture efficiency robustness of Twist Target Enrichment Panels to randomly separated (RND) and contiguous mismatches (CONT) relative to perfect match capture. Additional details are provided in our whitepaper investigating the effect of mismatches on DNA capture by Hybridization.

MATERIALS & METHODS

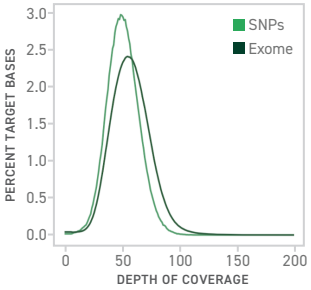
Genotyping Panel Design

To evaluate the applicability of Twist's Custom Target Enrichment Panels for genotyping by sequencing (GBS), we designed a proof-of-concept SNP panel against variants contained in a popular genotyping array: the Illumina Infinium Global Sequencing Array (GSAv2). After removing mitochondrial SNPs and variants that were less than 250 bp from genes to enable measuring GBS performance when run alongside the exome, approximately 240,000 SNPs remained that were amenable to short-read sequencing as determined by the high-quality regions of the Genome in a Bottle Consortium (GiAB).

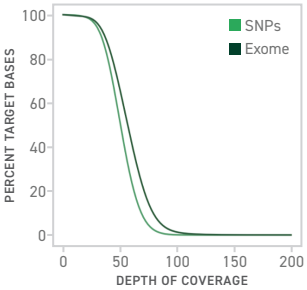
Evaluation of Genotyping Performance

Capture experiments were performed based on the Twist standard hybridization protocol using the SNP panel separately or as a spike-in to the Twist Human Core Exome Panel. All experiments were performed in replicate using genomic DNA samples from Coriell covering European continental, Asian continental, and Ashkenazi ancestry. These consist of cell lines NA12878, NA24694,

COVERAGE DISTRIBUTION



CUMULATIVE COVERAGE



CAPTURE METRICS

Mean target coverage	51
Fold 80 base penalty	1.29
Percent off-bait	33%
Percent 20x coverage	99%
Percent 30x coverage	95%
Zero Coverage	0.005%
Duplication rate	2.5%
AT dropout	0.39
GC dropout	<0.01

Figure 2: Capture performance for targeted SNP genotyping after 150x raw sequencing coverage. Graphs show the distribution of coverage across target bases comparing SNP and core exome panels. Coverage distribution: the percentage of bases at a given level of coverage. Cumulative coverage: the percentage of bases at or above a given level of coverage. The table presents additional capture metrics obtained using Picard.

SNP GENOTYPING (TE SNP PANEL)

SAMPLE	REPLICATE	PRECISION	SENSITIVITY
NA12878	1	99.90	99.79
NA12878	2	99.91	99.81

SNP GENOTYPING (ARRAY BASED)

SAMPLE	PRECISION	SENSITIVITY
NA12878_1	99.59	99.35
NA12878_2	99.59	99.30
NA12878_3	99.58	99.37
NA12878_4	99.58	99.03

INDEL GENOTYPING (TE SNP PANEL)

SAMPLE	REPLICATE	PRECISION	SENSITIVITY
NA12878	Replicate-1	91.07	90.39
NA12878	Replicate-2	90.47	90.32

Table 1: Genotyping performance of Twist SNP panels compared to array-based genotyping for a single sample (NA12878).

and NA24143 which have been comprehensively evaluated by GiAB and included as standards for genotyping by the National Institutes of Standards and Technology.

Sequencing was carried out on the Illumina NextSeq platform, using a NextSeq500/550 High Output kit with 2x75 bp reads. Alignment to the human genome (based on the hg19 assembly, against which the original GSAv2 array was designed) was performed using BWA (Li and Durbin, 2009) with a minimum mapping quality of 20. Variant calling was performed using the best practices workflow for GATK v3.5 (Van der Auwer et al., 2013). Array-based genotyping was performed on aliquots of each of the same samples used for GBS in replicate by a 3rd party provider using the GSAv2 array and Genome Studio 2.0 to produce genotype calls. Conversion from Illumina top/bottom notation to plus/minus strand was performed using the Strand tool (Rayner and McCarthy, ASHG, 2011). Genotyping based on sequencing or arrays for matched targets was compared against the high confidence calls released by GiAB as the gold standard using the benchmarking pipeline and recommendations established by the Global Alliance for Genomics and Health (GA4GH; Kruche et al., 2019).

Evaluation of Reference-Allele Bias

The proportion of reads supporting the alternate allele for heterozygous SNP positions was computed and compared to the expectation of sampling alleles with an equal probability for sites with the same numbers of total reads (binomial with p=0.5 and n= the number of reads mapping at each SNP locus).

RESULTS

Capture Performance and Target Enrichment Metrics

The 240K SNP panel was designed in as similar manner as the Twist Human Core Exome Panel. It was first compared against the Twist exome by capturing SNP targets independently with 150X raw sequencing relative to the number of baited bases for each panel.

The SNP panel demonstrated excellent capture uniformity (1.35 fold 80 base penalty vs. 1.32 in our exome) and low duplication rates (2.5% vs. 2.9%), matching the extremely high-quality standards expected of Twist Custom Panel designs. Additionally, target-specific metrics such as AT and GC dropout are comparable with exome values, and only 0.005% of SNP positions had zero coverage (**Figure 2**).

An increase in off-target capture (33% for SNPs vs. 15% for the exome) can be observed driving a minor shift in the coverage distribution's peak (**Figure 2**). Nevertheless, when focusing only on SNP positions as targets, rather than all bases covered by baits, the metrics of SNP targets matched those of the exome with a percent 20x and 30x coverage of 99% and 95%, respectively, and a modest increase in the uniformity of capture (1.29 fold 80 base penalty and narrow coverage distribution).

Highly Sensitive and Accurate Genotyping

Following the verification of panel capture performance (**Figure 2**), we evaluated GBS metrics compared to array-based call rates for the same SNPs using GiAB calls in each of the three samples

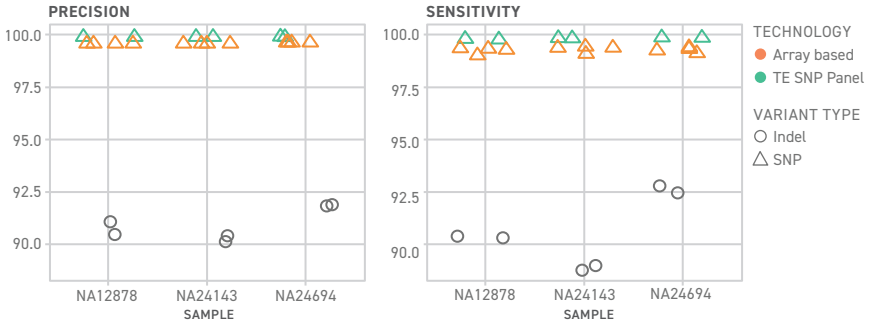


Figure 3: Comparison of performance between the Twist SNP panel and a leading genotyping array across GiAB samples from three different populations.

MEAN COVERAGE	REPLICATE	10X (0.84 GB SEQ)	15X (1.4 GB SEQ)	18X (1.6 GB SEQ)	20X (1.9 GB SEQ)
SNP Precision	Rep 1	99.20	99.82	99.88	99.90
	Rep 2	99.20	99.80	99.86	99.89
SNP Sensitivity	Rep 1	90.39	97.75	98.84	99.33
	Rep 2	90.50	97.72	98.70	99.22

Table 2: SNP Genotyping as a function of mean target coverage (based on NA12878).

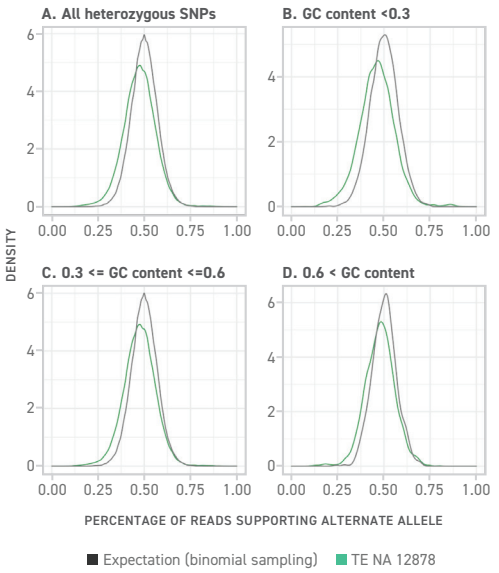


Figure 4: Percentage of reads supporting the alternate allele across A): all heterozygous SNPs or B–D): stratified by GC content.

studied as the gold-standard. Results based on 150X sequencing for GBS are shown in **Table 1**, individually for NA12878, and jointly, in **Figure 3**, across variant types, technologies, and samples from different populations. Precision and sensitivity for the SNP panel matched or exceeded those in arrays, all of which were greater than 99%. Additionally, the genotyping panel allowed us to identify insertions and deletions at rates of over 90% precision and over 88% sensitivity.

We next performed a sub-sampling analysis using NA12878 and observed that metrics were robust until a 20X mean coverage. SNP sensitivity drops noticeably below 15x coverage, but precision remains higher than 99% (**Table 2**). The amount of sequencing required to hit a given mean coverage across SNPs is also shown on the table, with 1.9Gb of sequencing enabling high sensitivity genotyping applications for panels of ~250K SNPs. Noting that the precision of the SNP panel remains >99% even at 10x coverage (**Table 2**), the same amount of sequencing can easily afford panels of 500K SNPs and above for applications that can tolerate some false negatives, or that can statistically integrate weak information across SNPs such as imputation or ancestry inference.

Robustness of Genotyping Performance to Allele and Context Biases

To reduce manufacturing costs, baits for target enrichment panels are usually designed against a single genomic sequence. This leads to a mismatch in polymorphic positions where a sample contains an allele different from the genome reference, but a perfect match otherwise. To quantify the possible biases introduced by this type asymmetry in design, we looked at 70,693 GiAB reported heterozygous SNPs for NA12878 in the GBS panel and found that on average, the percentage of reads supporting the non-reference allele is 47%, deviating only by 3% from the 50% mean expected under no bias (**Figure 4A**). Encouragingly, this small reference bias is within the limits of biases known to be introduced by read alignment algorithms alone, which inherently will prefer reads that match the reference genome perfectly, compared to reads containing variants (Lunter and Goodson 2011, Degner et al. 2015). When stratifying by GC content (**Figure 4B–D**), a modest increase was observed to a mean value of 46% for baits with very low GC (**< 30%, panel B**). This led to a 6% increase in the number of positions captured (**green curve, panel B vs. green curve panel A**) falling below what is expected from sampling variance with perfect 50:50 probability for both alleles (**grey curve in all panels**, see methods).

Genotyping and High Coverage Whole Exome Sequencing

In addition to testing the capture performance of the SNP panel on its own, we also tested the combined capture of SNP targets alongside high coverage whole-exome sequencing (WES) using our core exome and a spike-in titrated to yield mean 0.5X fold coverage for SNPs relative to coding sequences (Exome + 0.5x SNPs). **Table 3** shows a comparison of key capture metrics, obtained after 150X raw sequencing, for each set or combination of targets (SNPs or Exome only, or both Exome and SNPs) for the combined capture experiment. Metrics obtained for captures using each panel independently, and varying targets between SNPs and baits for the GBS panel are provided for comparison.

Analyses across all targets (illustrated in **Figure 5** for one locus) showed maintenance of both panels' key capture metrics after combined capture. AT/GC dropout rates and fold 80 remained the same, and mean coverage responded as expected relative to the 0.5X coverage selected for SNPs. The only exception was a small increase in fold 80 base penalty for SNP targets, but to a value still below 1.4, representing best in class uniformity for targeted capture panels. Note that the fold 80 base penalty for the Exome + SNP targets is undefined in the combined capture experiment given the coverage distribution of the combined panel is no longer unimodal by design (with the mean of one part of the panel at ~0.5x of the other).

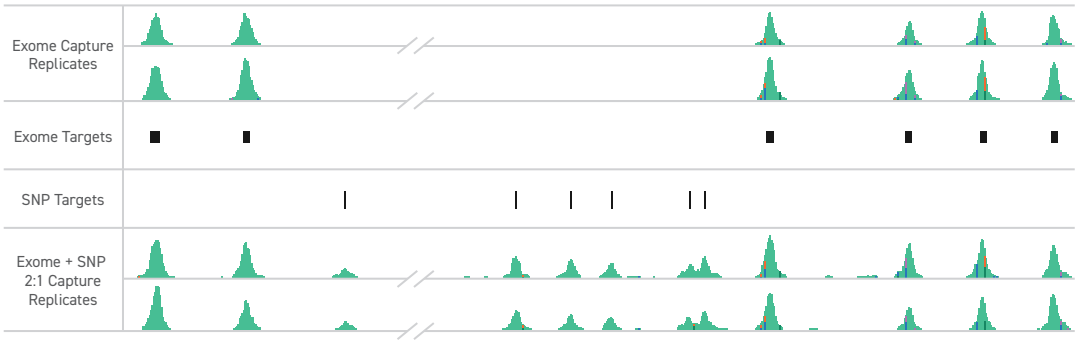


Figure 5: Coverage for the 0.5X SNP to 1X exome combined capture panel for a small region between chr1 between 165,320,712 and 165,379,000 bp in hg19 (broken up for visualization to exclude a small segment in the 165,330,000 to 165,362,000 bp range). Capture peaks correspond to the first two exons of the LIM Homeobox Transcription Factor 1 alpha (LMX1A gene) on the left, followed by coverage for SNP targets, and the last three exons of the retinoid receptor X gamma (RXRG) on the right.

CAPTURE EXPERIMENT	REGIONS EVALUATED	MEAN COVERAGE	% TARGET BASES 20X	% TARGET BASES 30X	AT DROPOUT	GC DROPOUT	FOLD 80
Exome + 0.5x SNPs	Exome + SNPs	71	98%	97%	2.26	0.68	—
	SNPs only	37	94%	71%	0.53	<0.01	1.38
	Exome only	72	98%	97%	2.24	0.53	1.37
Exome	Exons	57	98%	95%	1.78	0.41	1.32
SNPs	Baited SNP Bases	44	98%	86%	4.48	0.05	1.35
	SNPs	51	99%	95%	0.39	<0.01	1.29

Table 3: Capture performance of a blended panel comprising both SNP and exome targets.

DISCUSSION AND SUMMARY

In this study, we have focused on testing the performance of Twist's unique manufacturing and Custom Target Enrichment Panel design capabilities for genotyping by sequencing using a set of SNPs predefined within a popular genotyping array.

Out-of-the-box results across all samples showed excellent performance with exceptional uniformity and low duplicates, high coverage of target SNPs, and call rates matching or exceeding array-based genotyping. The SNP panel also provided the ability to detect more complex variation sources, such as a broader range of indels.

Although we also observed an increase in off-target capture for the custom GBS panel relative to the Twist Core Exome panel, coverage metrics matched or exceeded those in exons due to the narrower target profile of SNPs relative to the sequencing coverage peaks generated by baits. While a very modest reference bias was appreciable at low GC, the custom GBS panel enabled genotyping with sensitivity and precision of >99%, with as little as 20X mean coverage across SNPs. Additionally, and of particular interest for genotyping applications where information is integrated across SNPs (such as imputation or ancestry decomposition), 97% sensitivity at 15X and 90% at 10X mean coverage across SNPs is maintained along with a precision >99%.

The high performance of our panels also opens the door to standalone NGS-based genotyping at a scale. Examples include genomic ancestry decomposition, imputation driven whole-genome association studies, and variable depth genotyping applications. Together with the excellent compatibility of Twist Target Enrichment workflows and custom GBS panels, Twist provides an exciting alternative to genotyping arrays for a variety of standalone or combined applications where array-based genotyping is still run alongside sequencing.

REFERENCES

Degner J., et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. 2009. Bioinformatics, 25(24), 3207–3212 https://dx.doi.org/10.1093/bioinformatics/btp579

Heng L, Richard D. Fast and accurate short read alignment with Burrows-Wheeler transform. 2009. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Krusche, P, Trigg, L, Boutros, PC et al. Best practices for benchmarking germline small-variant calls in human genomes. 2019.Nat Biotechnol, 37, 555–560. https://doi.org/10.1038/s41587-019-0054-x

Lunter G., Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. 2011. Genome Res, 21(6), 936-9. https://doi.org/10.1101/gr.111120.110

Rayner NW, McCarthy MI. Development and Use of a Pipeline to Generate Strand and Position Information for Common Genotyping Chips. ASHG Conference Poster. https://www.well.ox.ac.uk/~wrayner/tools/

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. 2013. Curr Protoc Bioinformatics, 43(1110). https://doi.org/10.1002/0471250953.bi1110s43

Published: November 2, 2020 | DOC-001171 | Revision: 1.1