Targeted Sequencing-based Genotyping as a Competitive Alternative to Genotyping Arrays

Adriana Arneson, Leonardo Arbiza, Kristin Butcher, Siyuan Chen, Sabina Gude, Brenton Graham, Richard Gantt, Rebecca Liao, Rebecca Nugent, Christina Thompson

1. Abstract

Arrays have long been the go-to method for high throughput genome-wide genotyping single nucleotide polymorphisms (SNPs) at a large scale throughout the genome. In this poster, a case study is presented for the design of a Twist Custom Target Capture Panel 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, indel genotyping and whole-exome sequencing can now be performed under one platform, reducing costs, time, and effort.

2. 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. They have advanced our understanding of areas from evolutionary genomics, and heritable and complex disease, to personalized genomics and medicine. In recent years, reductions in the cost of next generation sequencing (NGS) have made it an attractive option for genotyping, expanding our ability to detect multi-allelic sites, insertions, deletions and other structural variants with increased flexibility compared to the fixed template format of arrays.

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.

Here we leveraged Twist's Custom Panel design algorithms 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 and compare precision and sensitivity directly against results from matched array-based genotyping using genomic genotyping standards and show precision and sensitivity > 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 . Overall we demonstrate a unified workflow to merge genotyping with exome sequencing, which leads to considerable savings in money and effort compared to running each individually.

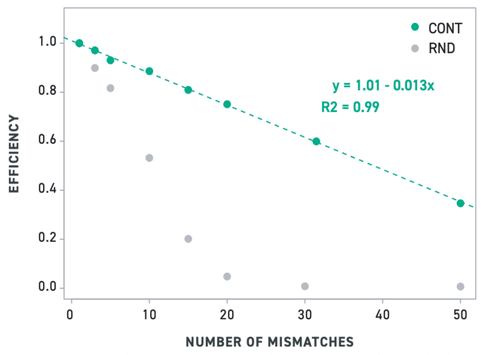


Figure 1: Hybridization Efficiency of Twist probes as a function of mismatches to the captured sequence. CONT points represent mismatches in a continuous stretch of a certain length. RND points represent a random number of mismatches in a probe sequence.

3. Materials and Methods

Genotyping Panel Design

To evaluate the applicability of Twist's custom target enrichment panels for Genotyping by Sequencing (GBS), a proof-of-concept SNP panel was designed in a manner complementary to the Twist Human Core Exome Panel, based on variants contained in a leading genotyping array containing ~600K SNPs. After removing mitochondrial SNPs and variants that were less than 250 bp from genes,~240K SNPs remained that were amenable to short read sequencing as determined by the high quality regions from 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. These consist of cell lines NA12878, NA24694, and NA24143 which have been comprehensively evaluated by GiAB and included as standards for genotyping by the National Institutes of Standards and Technology, covering European continental, Asian continental and Ashkenazi ancestry.

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 done using BWA[1] with a minimum mapping quality of 20. Variant calling was performed using the best practices workflow for GATK v3.5[2]. 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 done using 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[3].

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).

4. Capture Performance

Metric (Picard)	
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

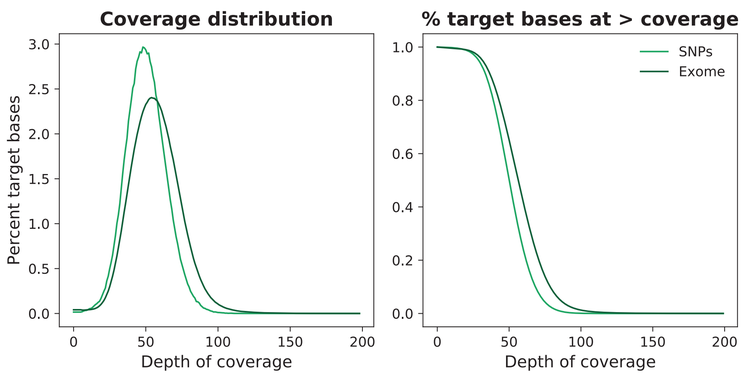
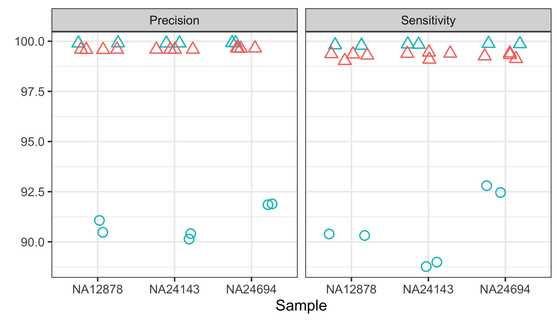


Figure 2: Capture performance for targeted SNP genotyping. The table presents Picard metrics based on 150x sequencing after capture. Graphs show the distribution of coverage across target bases comparing SNP and core exome panels.

5. Genotyping performance



Sequencing based				Array based			
SNP Genotyping NA12878							
Sample	Replicate	Precision	Sensitivity	Sample	Replicate	Precision	Sensitivity
NA12878	1	99.90	99.79	NA12878	1	99.59	99.35
NA12878	2	99.91	99.81	NA12878	2	99.59	99.30
				NA12878	3	99.58	99.37
				NA12878	4	99.58	99.03
Indel Genotyping NA 12878							
Sample	Replicate	Precision	Sensitivity				
NA12878	1	91.07	90.39				
NA12878	2	90.47	90.32				

Mean Coverage	Replicate	10x 0.84Gb seq	15x 1.4Gb seq	18x 1.6Gb seq	20x 1.9Gb seq
SNP Precision	1	99.20	99.82	99.88	99.90
	2	99.20	99.80	99.86	99.89
SNP Sensitivity	1	90.39	97.75	98.84	99.33
	2	90.50	97.72	98.70	99.22

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

Following the verification of panel capture performance (**Figure 2**), we evaluated GBS metrics as compared to array-based call rates for the same SNPs using GiAB calls in each of the three samples studied as the gold-standard. Results based on 150X sequencing for GBS are shown in Figure 3. 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.

Table 1: SNP Genotyping as a function of mean target coverage (NA12878)

We next performed a sub-sampling analysis using NA12878 and observed that genotyping metrics were robust up until a 20X mean coverage. SNP sensitivity drops sharply below 15x coverage, but precision remains higher than 99% (**Table 1**). 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 in the order of ~250K SNPs. Noting that the precision of the SNP panel remains >99% even at 10x coverage (**Table 1**), the same amount of sequencing can easily afford panels of 500K SNPs and above for applications that can tolerate some false negatives or can statistically integrate weak information across SNPs such as imputation or ancestry inference.

6. Adressing Allele and Context Bias

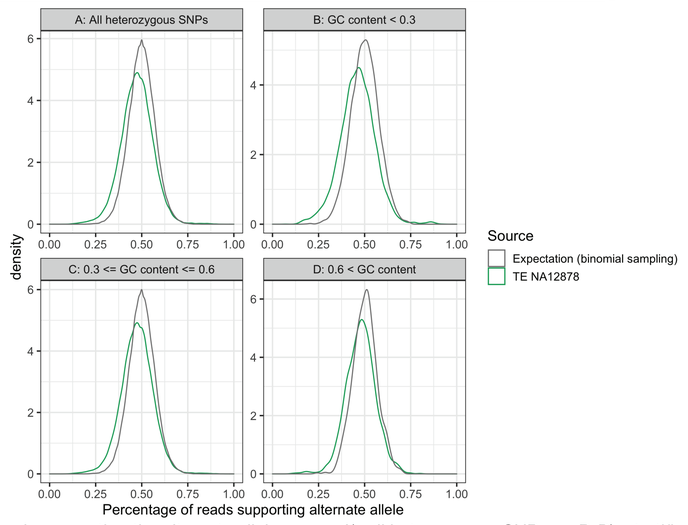


Figure 4: Percentage of reads supporting the alternate allele across A): all heterozygous SNPs, or B-D): stratified by GC content

To reduce 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 that is 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) which lead to a 6% increase in the positions captured (green curve, panel C) relative to SNPs across all GC bins (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).

7. Genotyping and Exome Sequencing

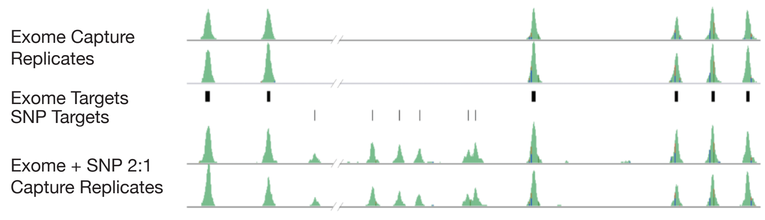


Figure 5: Coverage for the 0.5X SNP to 1X exome combined capture panel for a small region between chr1 between 165,320,712 and 165,379,000 bp in hg19 (broken up for visualization to exclude a small segment in the 165,330,000 to 165,362,000 bp range). Capture peaks correspond to the first two exons of 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 + 0.6x SNPs	Exome + SNPs	71	0.98	0.97	2.26	0.68	-
	SNPs only	37	0.94	0.71	0.53	<0.01	1.38
Exome	Exome only	72	0.97	0.96	2.24	0.53	1.37
	Exome only	57	0.98	0.95	1.78	0.41	1.32
SNPs	Baited SNP Bases	44	0.98	0.86	4.48	0.05	1.35
	SNPs only	51	0.99	0.95	0.39	<0.01	1.29

Table 2: Capture performance of a blended panel comprising both SNP and exome targets

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). Analyses across all targets (illustrated in **Figure 5** for one locus) showed that key capture metrics for both panels were maintained 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.

References

[1]Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754-1760
[2]Van de Auwera, G. et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics, 43(1110):11.10.1-11.10.33
[3]Krusche, P. et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. Nature Biotechnology, 37(5), 555-560

Financial disclosure

Nothing to disclose