applied biosystems

Development and characterization of a high throughput targeted genotyping-by-sequencing solution for agricultural genetic applications

Michelle Swimley, Angela Burrell, Prasad Siddavatam, Chris Willis, Christina Buchanan-Wright and Rick Conrad, Thermo Fisher Scientific, 2130 Woodward Street, Austin, TX, USA, 78744

ABSTRACT

Genotyping by Sequencing (GBS) is emerging as a powerful and cost-effective method for discovery and genotyping SNPs in agricultural species. Targeted GBS provides a lower-cost alternative to microarrays when analyzing 5000 variants or less, and can dramatically increase sample throughput up to thousands of samples per day.

The Applied Biosystems™ AgriSeq™ targeted GBS solution is a flexible, customizable amplicon resequencing workflow, which uses ultra-high multiplex PCR for targeted sequencing of known SNPs, MNPs, and INDELs with downstream sequencing on the Ion S5™ sequencing system. With 60-80M reads per Ion 540™ chip, the potential to barcode and multiplex numerous samples can significantly reduce the cost and handling requirements of sequencing while increasing throughput. Here we report the development and characterization of a set of 768 barcodes that enables up to 1536 samples to be processed per day, the coverage requirements for optimal genotyping and the performance of this technology across several different species

INTRODUCTION

The study of DNA polymorphisms is the basis of molecular breeding. Single nucleotide polymorphisms (SNPs), which are heritable and generally have a low mutation rate, have emerged as the most widely used molecular marker for genotyping applications. With advances in next generation sequencing technologies and targeted resequencing approaches, genotyping by sequencing (GBS) provides an attractive alternative to arrays for mid-density SNP genotyping. The AgriSeq GBS workflow is a targeted resequencing workflow, developed to provide a high-throughput, low-cost, reproducible and robust solution to deliver consistent plant and animal genotypes.

Historically, only 384 IonCode™ Barcode Adapters have been available, limiting the number of samples that can be sequenced in parallel for smaller panels. In this study, we developed an additional set of 384 IonCode barcode adapters in order to increase the multiplexing capability to 768 samples per chip for the AgriSeq GBS workflow. Each barcode has been validated to ensure similar performance across amplicons, and to ensure consistent coverage across samples and efficient use of sequencing coverage.

MATERIALS AND METHODS

Barcode Design

Barcode candidate sequences <14 bases long were generated in silico based on ternary error correcting algebraic code to ensure synchronization in flow space and error correcting requirements. Candidate sequences were further filtered based on different criteria including length, GC content, homopolymer length and secondary structures to generate 428 sequences that were evaluated in wet lab.

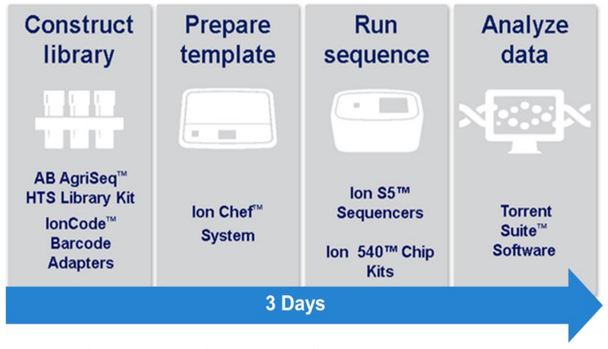
Wet lab evaluation of barcodes

Candidate barcode sequences were evaluated for functional performance with three custom AgriSeq panels; a small bovine parentage panel containing 200 markers, a maize genotyping panel containing 903 markers and a large porcine panel containing 3682 markers.

NGS and Data Analysis

Once constructed, AgriSeq libraries were diluted 2-fold and placed on the Ion Chef overnight, for template preparation and chip loading, followed by sequencing on the Ion S5 XL. Analysis was performed using the Coverage Analysis and Torrent Variant Caller (TVC) plugins using default settings for germline analysis. These plugins are available directly from the Torrent Suite™ Software on the Torrent server. The TVC plugin is a tool designed to detect and call variants (i.e. SNPs). The Coverage Analysis plugin provides statistics that describe the level of sequence coverage produced for targeted genomic regions. Barcode performance was assessed using coverage metrics (mapped reads*, uniformity* and strand bias*) and call rates*. An additional proprietary analysis was used to assess each barcode for ligation bias towards inserted amplicons. Any representation bias, as indicated by variation of the observed versus expected amplicon frequency per barcode, is shown as a Pearson correlation coefficients.

Figure 1. AgriSeq GBS workflow for performing sequencing and analysis



The AgriSeq targeted GBS workflow can be completed from DNA to results in as little as three days.

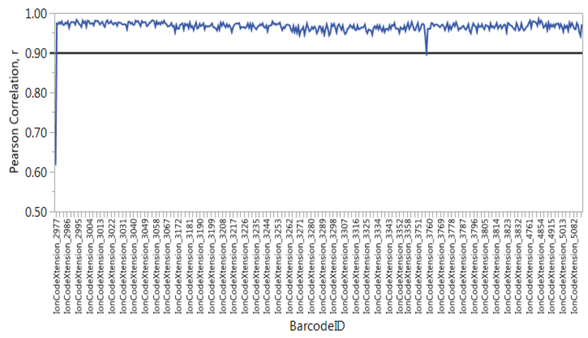
RESULTS

Table 1. Theoretical Genotyping Scalability

Maximum Recommended Markers	Samples per chip	Samples per day
5000	140	280
3645	192	384
1822	384	768
1215	576	1152
911	768	1537

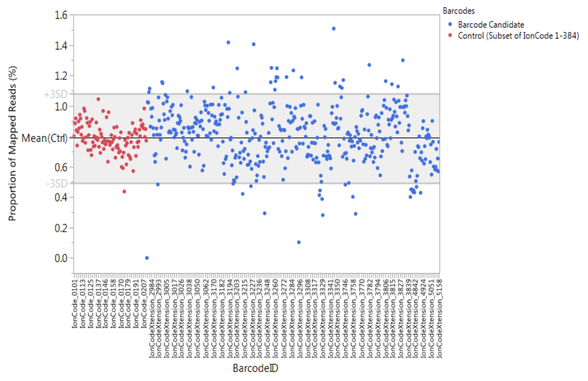
The maximum number of samples that can be analyzed at different marker densities per chip or per day on an Ion 540 chip, assuming an average of 70 million reads/chip, to achieve 100x average amplicon coverage.

Figure 2. Pearson correlation coefficients for each barcode candidate to identify sequence ligation bias



No intrinsic sequence bias observed for majority of the candidate barcodes (evaluated on 428 IonCode barcodes with a 3682 amplicon panel).

Figure 3. Proportion of mapped reads (%) for each barcode candidate to identify barcodes showing representation bias

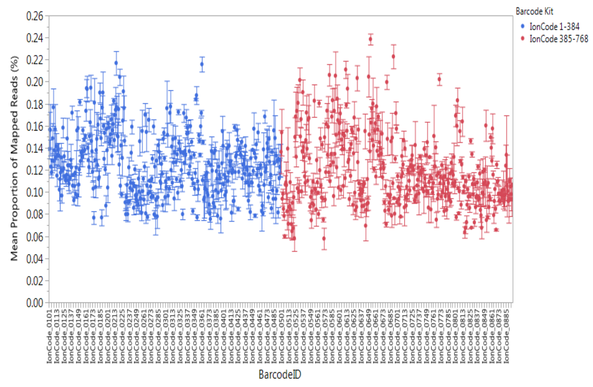


The shaded area represents a range of acceptable performance with the upper and lower limits defined by ±3 standard deviations of the mean proportion of mapped reads (%) for the control. A maximum of 44 barcode candidates outside of this range were eliminated from the final set of 384.

HIGHLIGHTS

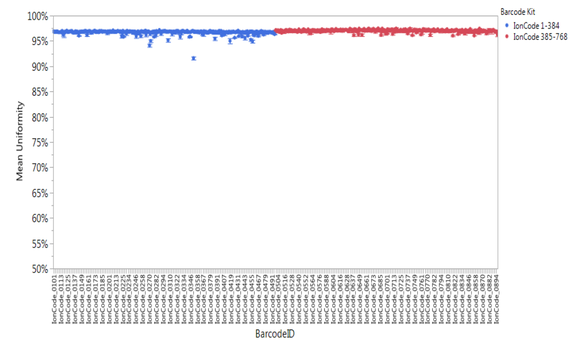
- Required only 10 ng of DNA
- Samples to results in as little as three days

Figure 4. Mean proportion of mapped reads (%) for all 768 IonCode barcode adapters



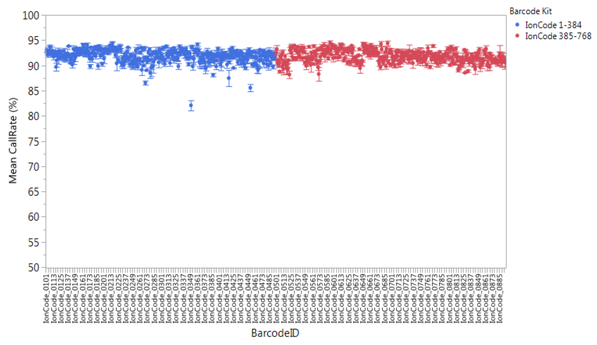
There was a very small difference in the mean proportion of mapped reads (%) between barcode kits 1-384 and the new barcodes 385-768 (difference = 0.008, t-test p-value <0.0001) but the difference was so small it's not of practical significance.

Figure 5. Mean uniformity for all 768 IonCode barcode adapters



The grand mean proportion of mapped reads for the maize panel across all barcodes was 0.12 ±0.03. There was a small difference in mean uniformity between barcode kits 1-384 and the new barcode kit 385-768 (difference = 0.003, t-test p-value <0.0001) but the difference was so small as to not be practically significant.

Figure 6. Mean call rate (%) for all 768 IonCode barcode adapters



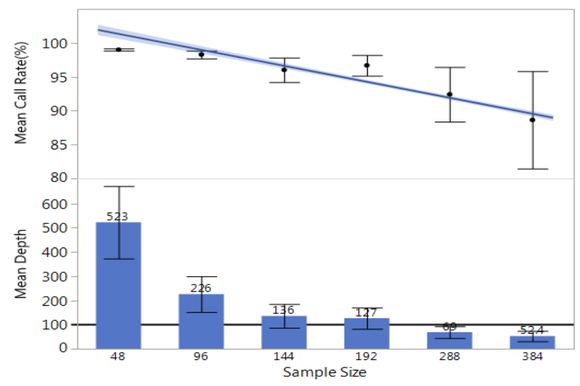
The grand mean uniformity for the maize panel across all barcodes was 96.8% ±0.003. There was no statistical difference in mean call rate between barcode kits 1-384 and 385-768 (difference = 0.10, t-test p-value 0.06). The grand mean call rate for this maize panel across all barcodes was 91.85% ±1.35%

Table 2. Performance metrics of AgriSeq panels

Species	Markers	Mean % Sample Call Rate	Mean % Uniformity	Mean % on Target reads
Bovine	190	99.7	98.0	97.6
	215	98.5	97.3	85.7
Canine	229	99.2	99.2	98.8
Feline	62	99.8	98.6	96.6
Porcine	1500	96.3	99.7	95.8
	3000	96.2	98.2	99.3
Soybean	1134	98.3	96.7	98.9
Cucumber	2804	91.4	96.8	99.7
Maize	1079	87.5	87.2	97.7
Salmon	3153	94.0	92.0	99.4

AgriSeq technology has been demonstrated to generate high marker call rates and performance across multiple agricultural species and panel sizes

Figure 7. Impact of coverage depth on sample call rate



At 120X mean depth, the average call rate is still very high and acceptable for most applications. With a target mean depth ≥100X, this technology has been demonstrated to generate high marker call rates and performance across multiple agricultural species and panel sizes

DEFINITIONS

- **Mean proportion of mapped reads:** The average number of mapped reads for a specific barcode to the reference, divided by the total number of reads per chip. It is a measure of how well balanced the ligation efficiencies are for each barcode to ensure balanced read depth between libraries.
- **Mean read uniformity:** The percentage of target bases covered by at least 0.2x the average base read depth. This metric shows how evenly the target amplicons are being covered by reads. Poor ligation efficiencies can impact uniformity (<90%) and lead to decreased call rates.
- **Call rates:** The average proportion of markers that generate a genotyping call.

CONCLUSIONS

Next generation sequencing offers great potential to fundamentally change the way plant and animal genotyping is delivered. With the availability of 768 high performing barcodes, AgriSeq targeted GBS can deliver up to 1536 sample genotypes with ~800 markers per day. The technology is flexible to scale to higher markers with fewer samples and has demonstrated performance across a breadth of species tested.

REFERENCES

Heaton MP, Harhay GP, Bennett GL, Stone RT, Grosse WM, Casas E, et al. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. Mamm. Genome. 2002;13:272–81.

He J., Zhao X., Laroche A., Lu Z.X., Liu H. and Li Z. (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front. Plant Sci. 5, 484.

Beissinger, T. M., Hirsch, C. N., Sekhon, R. S., Foerster, J. M., Johnson, J. M., Muttoni, G., Vaillancourt, B., Buell, C.R., Kaeppler, S.M., and de Leon, N. (2013). Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics, 193(4), 1073-1081. DOI: 10.1534/genetics.112.147710

TRADEMARKS/LICENSING

For Research Use Only. Not for use in diagnostic procedures.

ThermoFisher SCIENTIFIC