applied biosystems

# Targeted Next Generation Sequencing Approaches in Corn, Cucumber and Soy for High Throughput Genotyping

**M. Swimley[1], A. Burrell[1], R. C. Willis[1], L. Baselgia[2], P. Siddavatam[1], R. Conrad[1], and C. Buchanan-Wright[1]**
**[1]Thermo Fisher Scientific, 2130 Woodward Street, Austin, TX USA 78744, [2]Thermo Fisher Scientific, Wagistrasse 27A, 8952 Schlieren-Zurich, Switzerland**

## ABSTRACT

With advances in plant phenotyping approaches for quantitative genetic analysis and increasing complexity of gene pyramiding schemes, the number of markers required for successful molecular breeding programs in agriculture is increasing. Historically, technology has been polarized between high marker, high cost microarrays or low cost singleplex approaches that are not easily scalable. Targeted genotyping by sequencing (GBS) is emerging as a powerful alternative for mid-density genotyping of 100s to thousands of markers in a high throughput and cost-effective manner.

We have applied AgriSeq targeted GBS, a high throughput amplicon based sequencing workflow performed on the Ion S5 system, to three economically important crops. A 2800 marker cucumber panel, an 1100 marker soybean panel and two independent corn panels targeting 900 and 1000 markers were designed for the AgriSeq workflow. The average genotyping marker call rate ranged between 91%-98% for these panels, with >94% average uniformity and >99% on-target reads. >99.4% reproducibility has been demonstrated for this workflow over multiple independent library preparations and sequencing runs, with genotype concordance to orthogonal array technologies of 99.4%. Compatible with high throughput processing, the AgriSeq approach can multiplex up to 768 samples simultaneously and generate up to 1.6M genotypes per day. In addition, this approach allows for the discovery of additional SNPs and micro-haplotypes around the targeted markers, which enable further traceability and association in downstream studies.

These results demonstrate the utility of the AgriSeq targeted Genotyping by Sequencing workflow for plant molecular breeding programs.

## INTRODUCTION

Single nucleotide polymorphisms (SNPs), which are heritable and generally have a low mutation rate, have emerged as the most widely used genotyping markers in agricultural applications such as trait monitoring, marker-assisted breeding selection or germplasm identification. With advances in next generation sequencing technologies and targeted resequencing approaches, genotyping by sequencing (GBS) provides an attractive alternative to traditionally more costly arrays for mid-density SNP genotyping.

The AgriSeq GBS workflow is a targeted resequencing high-throughput workflow, designed to amplify and sequence up to 5000 genetic markers in a single multiplexed reaction. It offers a low-cost, reproducible and robust solution to deliver up to 1.6M genotypes per day.

Here we demonstrate the utility of the AgriSeq GBS workflow to genotype three different economically important crops, Maize, Cucumber and Soybean.

## MATERIALS AND METHODS

Four crop genotyping primer panels were designed using a reference-based GBS automated pipeline, which selects primers based on optimized parameters such as amplicon length, melting temperature and GC content, for multiplexing 100s to 1000s of oligonucleotides in a single PCR reaction. Primers were also designed to avoid overlapping nearby SNPs and prevent the formation nonspecific PCR products, characteristics which were assessed *in-silico* (Figure 1).

| Crop | No. of Markers |
|---|---|
| Maize Panel 1 | 900 |
| Maize Panel 2 | 1000 |
| Soybean | 1100 |
| Cucumber | 2800 |

These panels were tested using the AgriSeq GBS workflow using either the 96 or 384-well AgriSeq™ HTS Library prep protocol with 10 ng of genomic DNA input. Barcoded amplicon libraries were pooled 1:1 and loaded onto an Ion Chef™ System for template prep and chip loading onto an Ion 540™ chip, and then sequenced on an Ion S5™ XL System. Data were analyzed using the Torrent Variant Caller plugin available as part of the Torrent Suite software package, to determine the genotype calls for each marker and sample tested with each panel (Figure 2).

**Figure 1. Sequencing Panel Design Process**

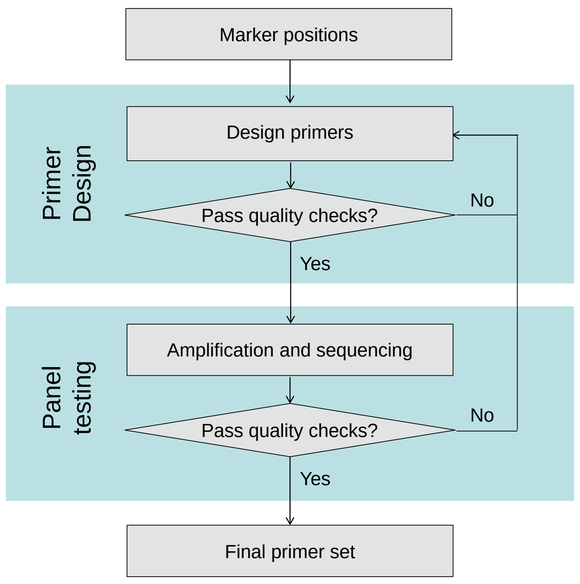



**Figure 1.** Overview of the primer design pipeline for targeted sequencing panels has multiple quality checkpoints to ensure primer specificity within the genome. Poor performing primers identified during *in-silico* or wet lab evaluation are removed from the panel or redesigned.

**Figure 2. AgriSeq GBS Workflow**

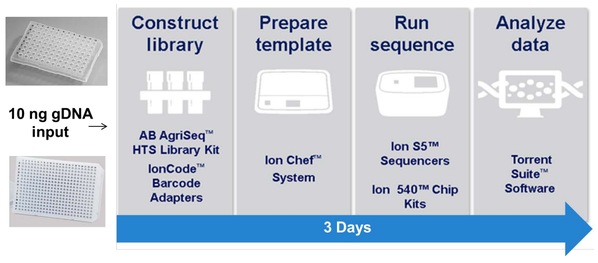



**Figure 2.** Requiring only 10 ng of genomic DNA, amplicon libraries can be constructed using the AgriSeq HTS Library Kit reagents in either 96-well or 384-well format for faster processing and compatibility with automated liquid handling platforms. Compatible with DNA barcoding, different barcoded adapters are used for each library, which allows them to be pooled for simultaneous sequencing of hundreds of samples on the Ion S5 sequencing platform. Once constructed, AgriSeq libraries are placed on the Ion Chef overnight, for template preparation and chip loading, followed by sequencing on the Ion S5 system the next day. The complete GBS workflow takes as little as three days from DNA to results.

**Table 1. Sample Throughput Capability**

| No. of Markers | Maximum no. of Samples per chip | No. of Samples per day |
|---|---|---|
| 5000 | 140 | 280 |
| 3645 | 192 | 384 |
| 1822 | 384 | 768 |
| 1215 | 576 | 1152 |
| 911 | 768 | 1537 |

**Table 1.** Sample scalability depends on the density of the chip used and the number of markers in the panel tested. This table shows the maximum recommended number of samples that can be analyzed at different marker densities per Ion 540 chip or per day, assuming an average of 70 million reads/chip to achieve 100X average base coverage.

## RESULTS

**Table 2. Summary of Panel Performance**

| Panel | Number of Markers | Libraries per Chip | Total Reads | Mean Coverage | Mean Uniformity | Mean On-Target | Mean Call Rate |
|---|---|---|---|---|---|---|---|
| Maize Panel 1 | 900 | 768 | 91.6 M | 119X | 96.8% | 99.0% | 91.9% |
| Maize Panel 2 | 1000 | 192 | 86.7 M | 302X | 87.0% | 97.7% | 90.8% |
| Soybean | 1100 | 48 | 71.4 M | 1167X | 97.0% | 98.9% | 98.3% |
| Cucumber | 2800 | 96 | 61.0 M | 187X | 97.0% | 99.7% | 91.5% |

**Table 2.** This table contains coverage metrics that are used to evaluate panel performance. All four panels evaluated show good performance metrics:

- Mean coverage is >100X, which is usually sufficient to generate good genotyping results, but this may vary by panel and GBS application.
- High uniformity (>90%) indicates even coverage of amplicons from end-to-end.
- High on target (>90%) indicates that majority of mapped reads are aligned over a target region. This metric is influenced by panel design.
- Mean call rate is the percent of markers within each sample that generate a genotype, across all samples tested. Low DNA quantity, poor DNA quality, PCR inhibitors, incorrect thermocycling parameters, carryover ethanol during cleanup steps etc. can all impact sample call rate making this the best indicator of overall workflow performance.

**Figure 3. Mean depth >100X provides sufficient coverage to generate high sample call rates.**

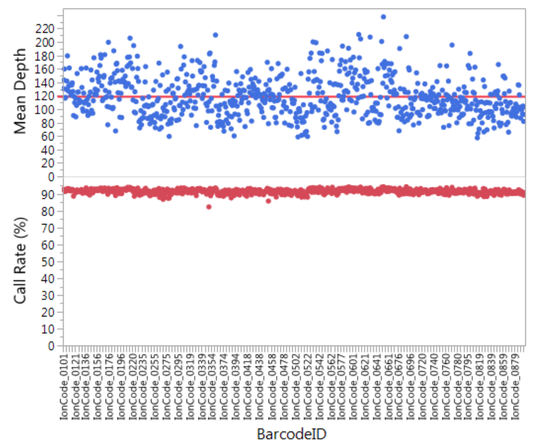



**Figure 3. Mean Depth & Call Rate results from Maize Panel 1.**
Mean Depth is the average base coverage depth over all bases targeted in the reference. It is used to determine if samples have sufficient coverage to generate genotype calls with a certain degree of confidence. Mean depth can vary significantly based on panel size and number of barcoded samples. This graph shows the mean depth and corresponding call rates for 768 barcoded libraries using Maize panel 1 with 900 targeted markers that were sequenced on an Ion 540 chip (60-80M reads). The grand mean depth for this panel was 119X (±30X) across all samples. The sample call rate was 91.7% (±1.3%) for this panel and number of samples tested per chip. This technology offers flexibility between sample throughput and marker density depending on the application requirements.

**Figure 4. Poor performing markers identified with low or no coverage.**

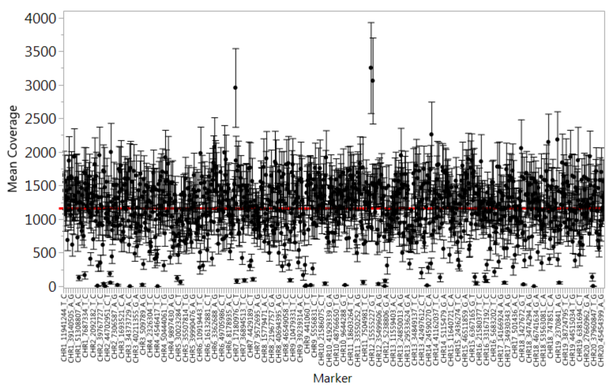
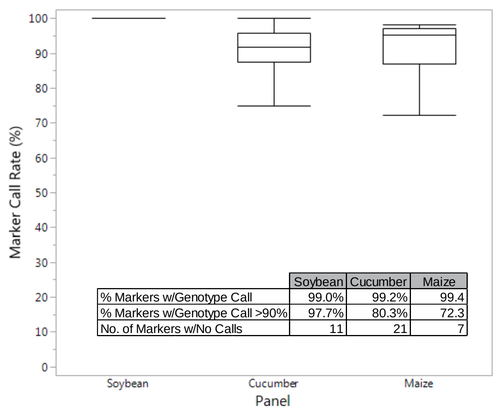



**Figure 4. Marker Coverage results from the Soybean Panel.**
Mean marker coverage is the average of the total coverage at a specified marker position. It is used to identify markers that may be over or under-represented in a panel design. This graph shows the mean coverage by marker position across 48 replicate samples. In this example, the Soybean panel has a mean marker coverage of 1167X (±504X). Overall, majority of the markers show consistent coverage with a few markers showing little to no coverage.

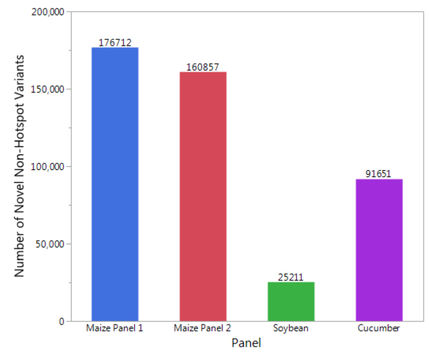
**Figure 5. Majority of markers from initial panel designs perform well**




| | Soybean | Cucumber | Maize |
|---|---|---|---|
| % Markers w/Genotype Call | 99.0% | 99.2% | 99.4 |
| % Markers w/Genotype Call >90% | 97.7% | 80.3% | 72.3 |
| No. of Markers w/No Calls | 11 | 21 | 7 |

**Figure 5.** Marker call rate is the percent of samples that generate a genotype call for a specific marker. It is also used to identify any under-performing markers in the panel design. Majority of markers generated a genotype call for samples tested with each panel. 95.9% of the markers for the Soybean panel had 100% marker call rate, which is excellent panel performance. Less than 1% of markers resulted in a call rate of 0% for all panels. No calls for these markers may be a result of low or no coverage, overlapping SNPs in the primer regions or lack of representation in the samples tested. Redesign or removal of poor performing markers will increase overall call rates for both markers and samples.

**Figure 6. Novel Non-Hotspot Variant Discovery**




| Panel | VariantType | N |
|---|---|---|
| Maize Panel 1 | DEL | 11840 |
| | INS | 81973 |
| | MNP | 6193 |
| | SNP | 76706 |
| Maize Panel 2 | DEL | 13625 |
| | INS | 14252 |
| | MNP | 5291 |
| | SNP | 127689 |
| Soybean | DEL | 953 |
| | INS | 884 |
| | MNP | 432 |
| | SNP | 22942 |
| Cucumber | DEL | 7144 |
| | INS | 4716 |
| | MNP | 2282 |
| | SNP | 77509 |

**Figure 6.** One of the advantages of using this technology is that it allows for the discovery of additional variants that fall within the targeted amplicon regions, which include SNPs, multiple nucleotide polymorphisms (MNP), insertions and deletions. This enables large-scale SNP discovery that can be used for further marker development and construction of genetic linkage maps, which serve as important resources to enable marker-assisted selection in breeding programs.

## CONCLUSIONS

Next generation sequencing offers great potential to fundamentally change the way plant genotyping is delivered. With the availability of 768 barcodes, the AgriSeq GBS workflow can generate genotypes on up 1536 samples per day. The technology is flexible to scale to high markers with fewer samples and has demonstrated performance across several agriculturally important plant species.

Custom GBS panel designs allow the user to find the right balance between cost and quality of results, which when combined with the AgriSeq library prep offers a robust and efficient workflow for SNP genotyping.

## REFERENCES

1. Patel DA, Zander M, Dalton-Morgan J, Batley J. Advances in plant genotyping: where the future will take us. Methods Mol Biol. 2015;1245:1-11. doi: 10.1007/978-1-4939-1966-6_1.
2. 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.

## TRADEMARKS/LICENSING

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



ThermoFisher SCIENTIFIC