Evolving next generation sequencing for production agriculture: increasing throughput, decreasing effort and delivering more

Jie Lu[1], R.C. Willis[1], A.Burrell[1], M. Swimley,[1] P. Siddavatam[1], C. Buchanan-Wright,[1] Haktan Suren[1],R. Conrad[1],

[1]Thermo Fisher Scientific, 2130 Woodward Street, Austin, TX, USA, 78744

ABSTRACT

The utility of restriction-enzyme genotyping by sequencing (GBS) in production agriculture is challenged because of the technology's limitations in SNP targeting and high rates of allele drop-out between samples. In contrast, targeted Genotyping by Sequencing (GBS) can deliver consistent, high marker call rates for specified SNPs in a high throughput, cost effective manner. AgriSeq™ targeted GBS allows up to 5000 markers to be simultaneously screened across 100s of samples in a single Ion Torrent sequencing run. The robustness of this technology has been demonstrated across 19 agriculturally relevant species, with marker call rates between 88-99%, >99% reproducibility and >99% concordance with orthogonal genotyping technologies. Here we report the expansion of the AgriSeq workflow to accommodate complex structural variants with in-dels ranging between 2-400,000bp.

Enhancements have also been made to the AgriSeq workflow to improve throughput. An additional 384 IonCode barcodes were developed to enable 768 sample multiplexing. Barcodes were validated over 3 different panels generating equivalent performance for ligation efficiency, uniformity and mapped reads. In addition, AgriSeq automation workflows were developed on the Gilson PipetMax to enable 384 samples processing with <1 hour of hands on time. Equivalent call rates, percentage on-target and coverage uniformity were observed between manual and automated processing across two different panels, over 6 runs with 192 samples. The expansion of available barcodes and automation of the AgriSeq library prep enables up to 1536 samples to be processed each day while reducing operator fatigue, the potential for technical errors, and the sequencing cost per sample.

INTRODUCTION

Genotyping has become an essential tool in parentage testing, marker assisted breeding and genomic selection in agricultural improvement programs. While high-density genotyping panels can drive the discovery and linkage new variants [1], their cost is not economically practical when genotyping large sets of individuals where fewer markers are required.

Imputation has enabled the use of fewer markers for breeding allowing for lower cost [2], alternative mid-density technologies to be considered for selective breeding programs. Targeted Genotyping by Sequencing (GBS) is one such alternative technology providing a low cost per sample, high throughput mid-density genotyping platform that can not only robustly and consistently genotype user specified marker, but also can discover novel variants through contextual sequencing.

In addition, sequencing data also allows for non-SNP variants such as INDELs and other structural variants to be identified which are beneficial for parentage and other genetic defect detection applications.

MATERIALS AND METHODS

The AgriSeq panels were designed using an automated process that optimizes a number of oligonucleotide properties (GC content, melting temperature, etc.) and amplicon properties (size, centering a SNP within its amplicon, etc.).

Library prep was performed using the AgriSeq workflow. There are 6 main steps in the AgriSeq workflow: initial amplification, Pre-Ligation Enzyme incubation, IonCode Barcode ligation, Pooling, Ampure cleanup, and Normalization.

AgriSeq library preparation may be automated using the AgriSeq Trilution workflow for the PipetMAX platform From Gilson. The Agriseq Trilution workflow is designed to prepare one, 384-well plate of multiplexed amplicon libraries with up to 384 unique barcodes in a single workday.

An additional 384 IonCode barcodes were developed to enable 768 sample multiplexing. Barcodes were selected and validated over 3 different AgriSeq panels for equivalent ligation efficiency, uniformity and number of mapped reads.

Large indel framework was developed to detect indels that are longer than 30bp. Genotyping calls were validated using 12 canine large indel markers on 300+ samples. Concordance was calculated based on the truth dataset of those samples.

Figure 1: AgriSeq targeted GBS workflow.

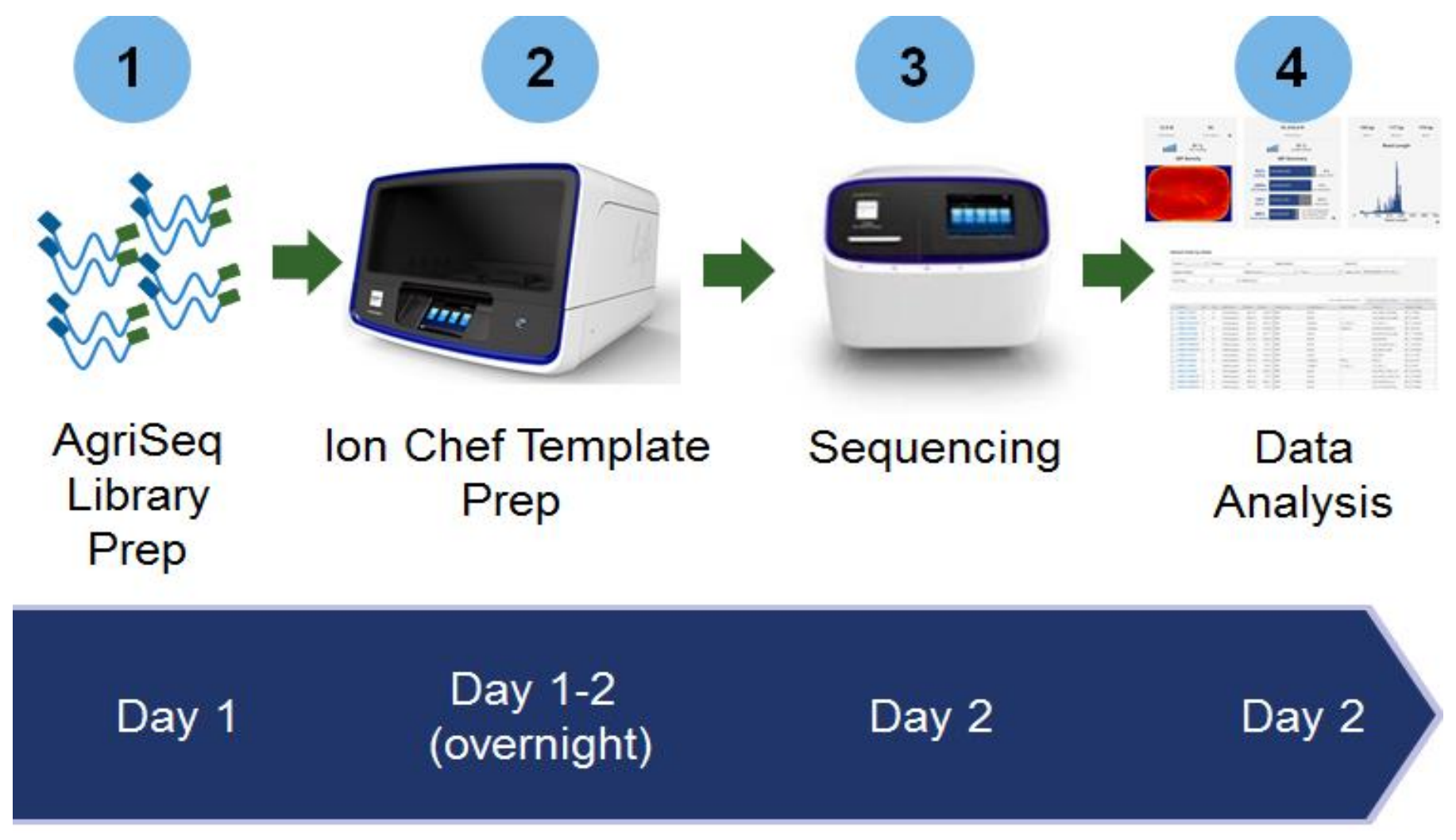


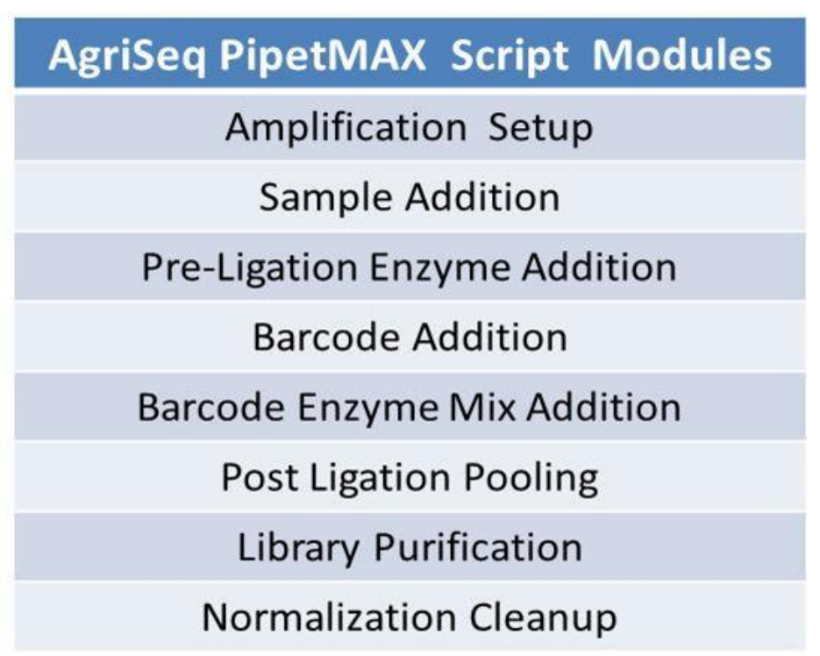
Table 1. The maximum number of samples that can be analyzed at different marker densitities per chip, per day, or per year on an Ion 540 chip.

	Maximum number of samples		
Markers	Per chip	Per day	Per year
5,000	140	280	72,800
3,645	192	384	99,860
1,822	384	768	199,780
1,215	576	1,152	299,580
911	768	1,537	399,560

Assuming an average of 70 million reads/chip to achieve 100X coverage amplicon coverage. Numbers are based on a single Ion Chief and Ion S5 System with one operator working a standard 8-hour shift, 5 days per week.

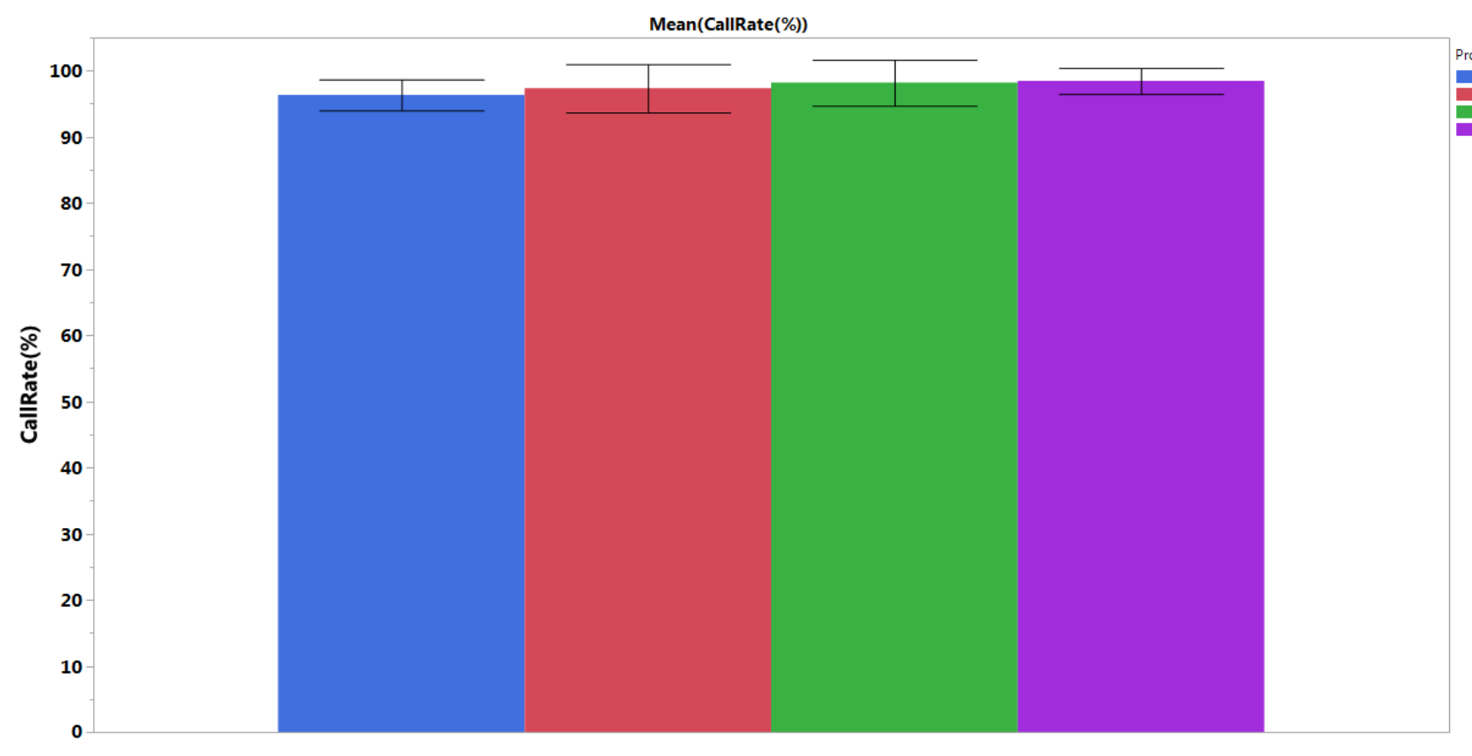
RESULTS

Figure 2. AgriSeq library preparation may be automated using the AgriSeq Trilution workflow for the PipetMAX platform From Gilson The Agriseq Trilution workflow is designed to prepare one, 384-well plate of multiplexed amplicon libraries with up to 384 unique barcodes in a single workday. The AgriSeq Trilution work flow will allow the user to process 96, 192, 270 or 384 samples at a time. The AgriSeq Trilution workflow is divided into 8 individual scripts allowing for a flexible workflow which enables processing of 384 samples with < 1 hour of hands on time.



AgriSeq PipetMAX Script Modules
Amplification Setup
Sample Addition
Pre-Ligation Enzyme Addition
Barcode Addition
Barcode Enzyme Mix Addition
Post Ligation Pooling
Library Purification
Normalization Cleanup

Figure 3: Equivalent call rate and genotype concordance between manual and automated 384-well workflows. A porcine genotyping panel targeting 1500 SNPs and 96 distinct porcine samples were used to compare the Gilson 384 well workflow with manual processing.



Run	Mean Genotyping Call Concordance	Stdev
Manual run	99.72%	0.7%
Gilson Agriseq-96 Run1	99.69%	0.7%
Gilson AgriSeq-96 Run2	99.12%	1.7%
Gilson AgriSeq-96 Run3	99.40%	1.2%

Figure 4. Performance of 768 barcodes generate similar proportion of mapped reads. We validated performance and reproducibility of a second set of 384 IonCode barcode adapters (IonCode 385-768) by preparing 768 uniquely barcoded libraries in triplicate using a small bovine panel (data not shown) and a large maize panel

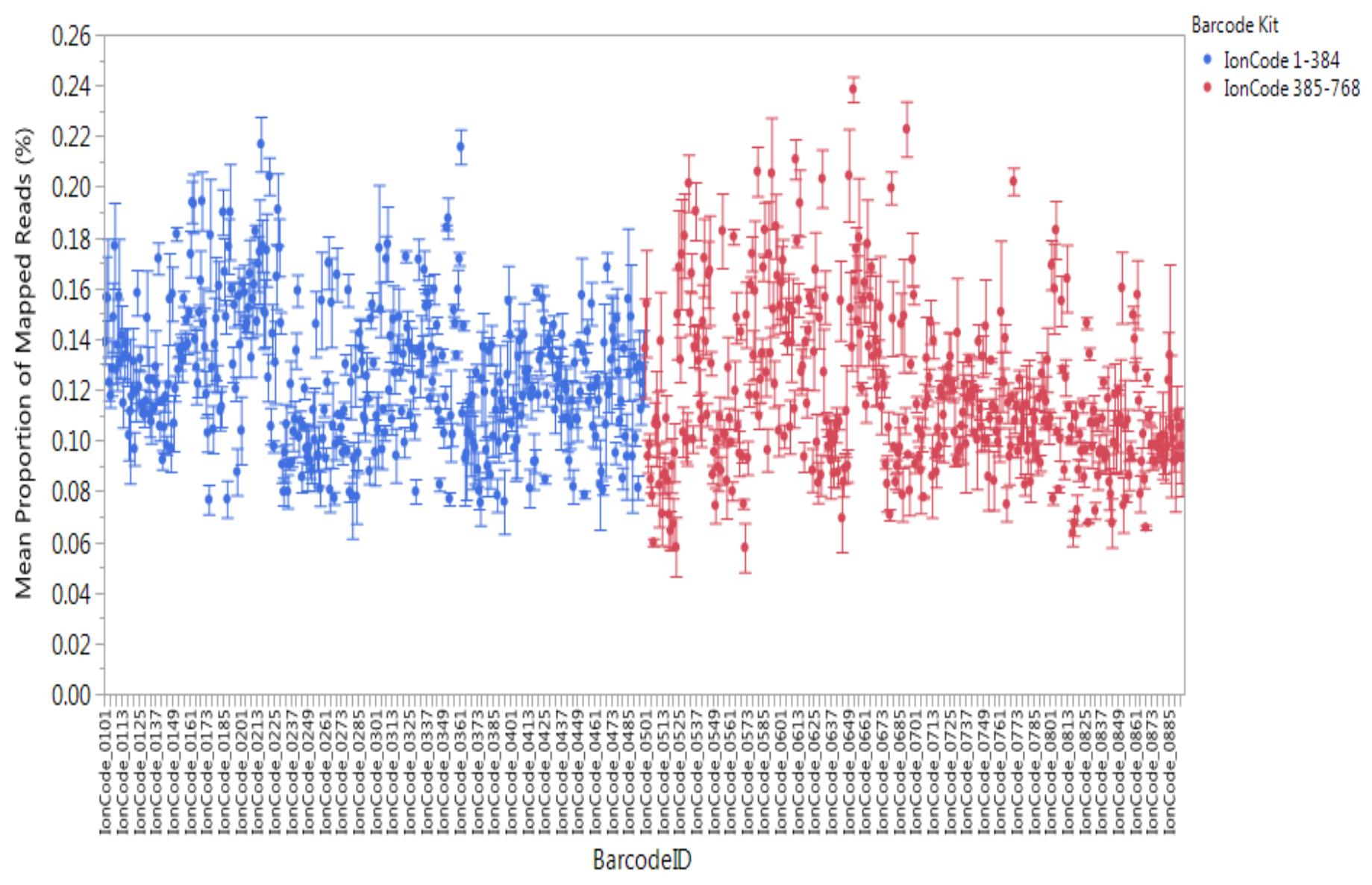


Table 2. The list of the organisms, number of markers per panel, type of variants, call rates, and reproducibility of the AgriSeq^{TM} panels

	Organism	Number of Markers	Variant Types	Call Rates (%)	Reproducibility (%)
1	Eucalyptus	155	SNPs	-	-
2	Equine	204	SNPs, INDELs, Large INDELs and Inversions	99.2	>99
3	Feline	315	SNPs, INDELs and Large INDELs	98.1	>99
4	Bovine	200	SNPs, INDELs	98.5	>99
5	Canine	412	SNPs, INDELs and Large INDELs	93.5	>99
6	Maize	1080	SNPs	87.5	>99
7	Soybean	1134	SNPs	98.3	>99
8	Chicken	1983	SNPs	-	-
9	Porcine	3000	SNPs	97.6	>99
10	Cucumber	3044	SNPs	90.8	>99
11	Salmon	3152	SNPs and INDELs	93.9	>99
12	Canola	5264	SNPs	-	-
13	Tomato	5588	SNPs	-	-

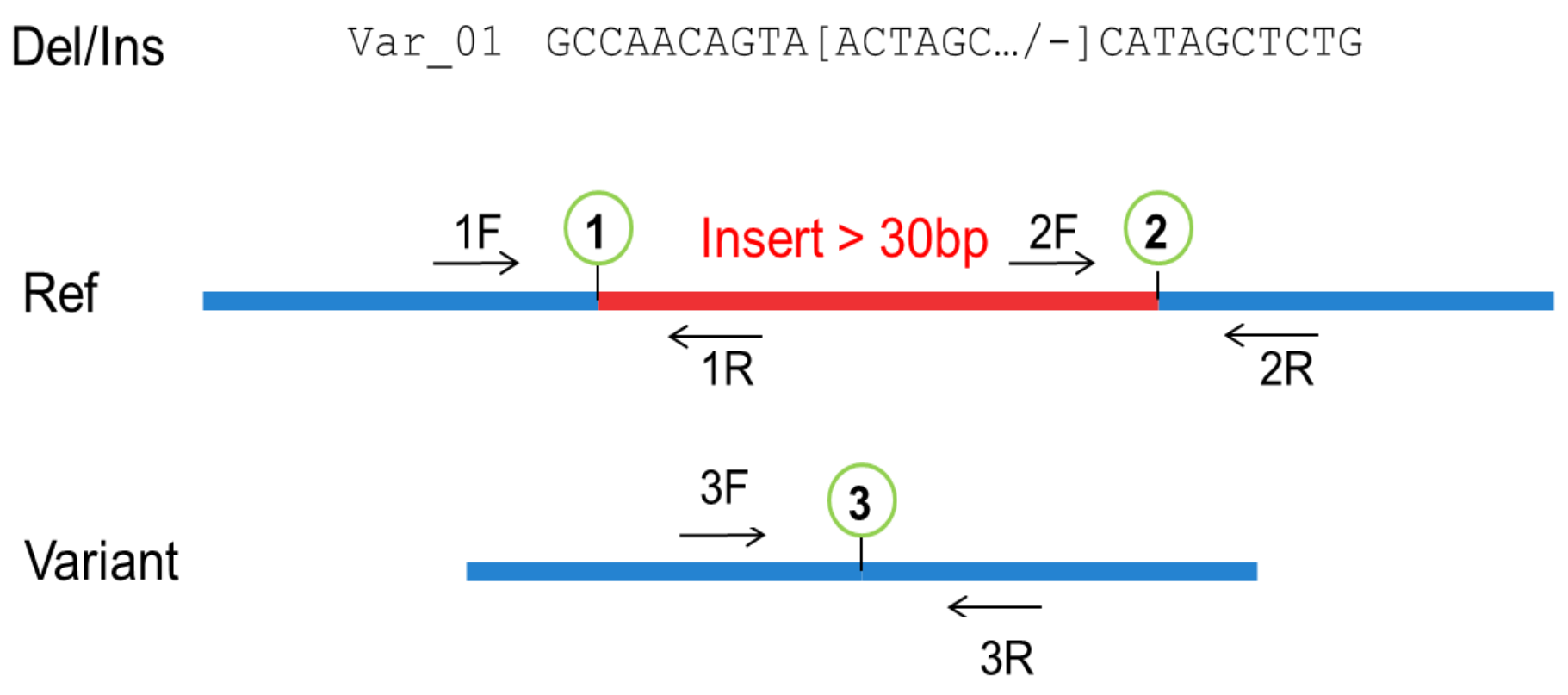
The robustness of this technology has been demonstrated across 13 agriculturally relevant species, with marker call rates between 88-99%, >99% reproducibility.

Table 3. Concordance with array data (ISAG Bovine Panel)

Samples run on both technologies	44
Total number of markers	200
Number of concordant calls	8751
Concordance (%)	99.44318

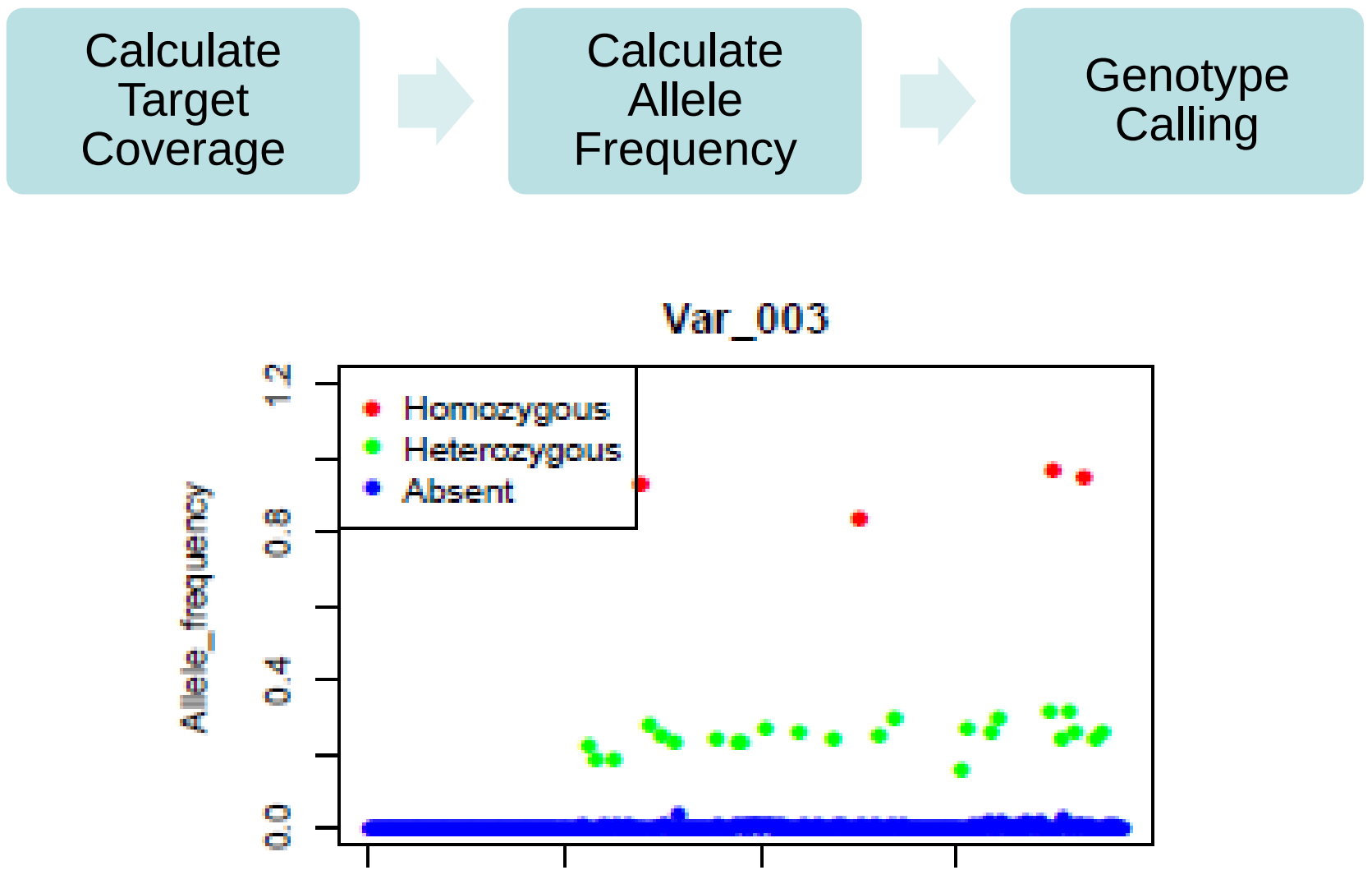
Samples were hybridized to six Illumina arrays in order to obtain consensus genotype calls for the array data. Concordance was calculated as the number of times the genotype call matched between samples run on the two different technologies divided by the total number of markers (200).

Figure 5. AgriSeq large indel solution



Indels longer than 30bp are split into two regions where two pairs of primers are designed to encompass the proximal junction (Target 1: 1F and 1R) and the distal junction (Target 2: 2F and 2R). Amplicons from two junctions, 1F+1R and 2F+2R, are expected if the insert is present while amplicons from only 1F+2R is expected if the insert is absent. If either Target 1 or 2 cannot be primed, an extra Target 3 which encompasses the breakpoint will be created and primed for amplification. In this scenario, only one junction, the amplicon from 1F+1R (or 2F+2R), will be detected if the insert is present while 3F+3R will be detected if the insert is absent.

Figure 6. Large indel variant calling workflow



To call the large indel genotype, a customized workflow was developed for AgriSeq^{TM}. Read counts/coverage per target and allele frequency per marker can be calculated. Final genotypes can be inferred based on allele frequency and sequencing depth.

Table 4. Concordance of canine large indel calls with truth dataset

Var Name	Type	Insert Length	Expected Positives	Observed Positives	Concordance
Var_1	INS	62	0	0	100%
Var_2	INS	78	0	0	100%
Var_3	INS	159	0	0	100%
Var_4	INS	179	0	0	100%
Var_5	DEL	180	0	0	100%
Var_6	INS	236	0	0	100%
Var_7	DEL	317	0	0	100%
Var_8	INS	4228	0	0	100%
Var_9	DEL	7799	28	28	100%
Var_10	DEL	15721	5	5	100%
Var_11	DEL	129788	4	4	100%
Var_12	DEL	405248	5	5	100%

We have tested 12 real canine long indel markers on 300+ samples. 100% concordance was observed when compared to truth data. It is not unusual that 2/3 markers show no variations among 300+ samples. There is no upper limit on indel size using this method.

.

CONCLUSIONS

We have demonstrated our enhanced AgriSeq GBS solution in a wide variety of agriculturally relevant species. We expanded our design framework to enable the detection of large indels (>30bp). Automation workflow and additional barcodes greatly increases the number of samples that could be processed each day while reducing sequencing cost per sample, operator hands-on time, and the potential for human errors.

REFERENCES

[1] Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature reviews. Genetics, 12(7): 499.

[2] Zhang Z, Ding X, Liu J, Zhang Q, de Koning DJ., 2011. Accuracy of genomic prediction using low-density marker panels. J Dairy Sci. 2011 Jul;94(7):3642-50.

For Research Use Only. Not for use in diagnostic procedures. AgriSeq is restricted for use with plants, agricultural animals or companion animals only. This product is not for use with human samples and/or in commercial applications.