appliedbiosystems

End-To-End AgriSeq™ Targeted GBS Long INDELs Solution

Haktan Suren[1], Chris Willis[1], Krishna Reddy Gujjula[1], Prasad Siddavatam[1], Jason Wall[1], Claudio Carrasco[1], Rick Conrad[1], Jeanette Schmidt[2]
[1]Thermo Fisher Scientific, 2130 Woodward Street, Austin, TX, USA, 78744, [2]Thermo Fisher Scientific, 3450 Central Expressway, Santa Clara, CA, USA, 95051.

ABSTRACT

AgriSeq™ targeted genotyping-by-sequencing (GBS) is successfully being used as a high throughput, customizable and cost effective genotyping solution in animal and plant breeding studies, parentage testing and genetic purity. One of the powers of this technology is its capability to support different types of markers including single nucleotide polymorphisms (SNPs), multiple nucleotide polymorphisms (MNPs), insertions and deletions (INDELs), and other structural variants (e.g. inversions, duplications). INDEL markers longer than 50bp can be a challenge for amplicon based GBS. The inclusion of these marker type require a different strategy during the panel design and genotype calling. We employed a three amplicon strategy to facilitate genotype calls from both the alleles and developed a new pipeline to automate the generation of present/absent calls. We successfully developed a custom canine SNP genotyping panel with 22 long INDELs and evaluated the performance with known true genotype samples.

The robustness of this technology has been demonstrated across 384 samples using 22 canine long INDEL markers whose length ranges from 62bp to 6.5Mbp. Overall, 97% call rate across samples and 100% concordance calls with true genotypes were observed. We found that primer design and down-stream analysis were not impacted by the INDEL size.

INTRODUCTION

Genotyping has been widely used in agricultural improvement programs for genomic selection, parentage testing and marker assisted breeding. The advent of next-generation technologies have dramatically dropped the DNA sequencing costs where genotyping-by-sequencing (GBS) is now feasible for high diversity and large genome species [1]. AgriSeq™ targeted GBS is developed to deliver consistent high marker call rates across diverse set of marker types and samples.

In addition to SNP markers, AgriSeq™ targeted GBS solution also allows for non-SNP variants like INDELs and other structural variants to be identified which are applications beneficial for parentage and other genetic defect detection.

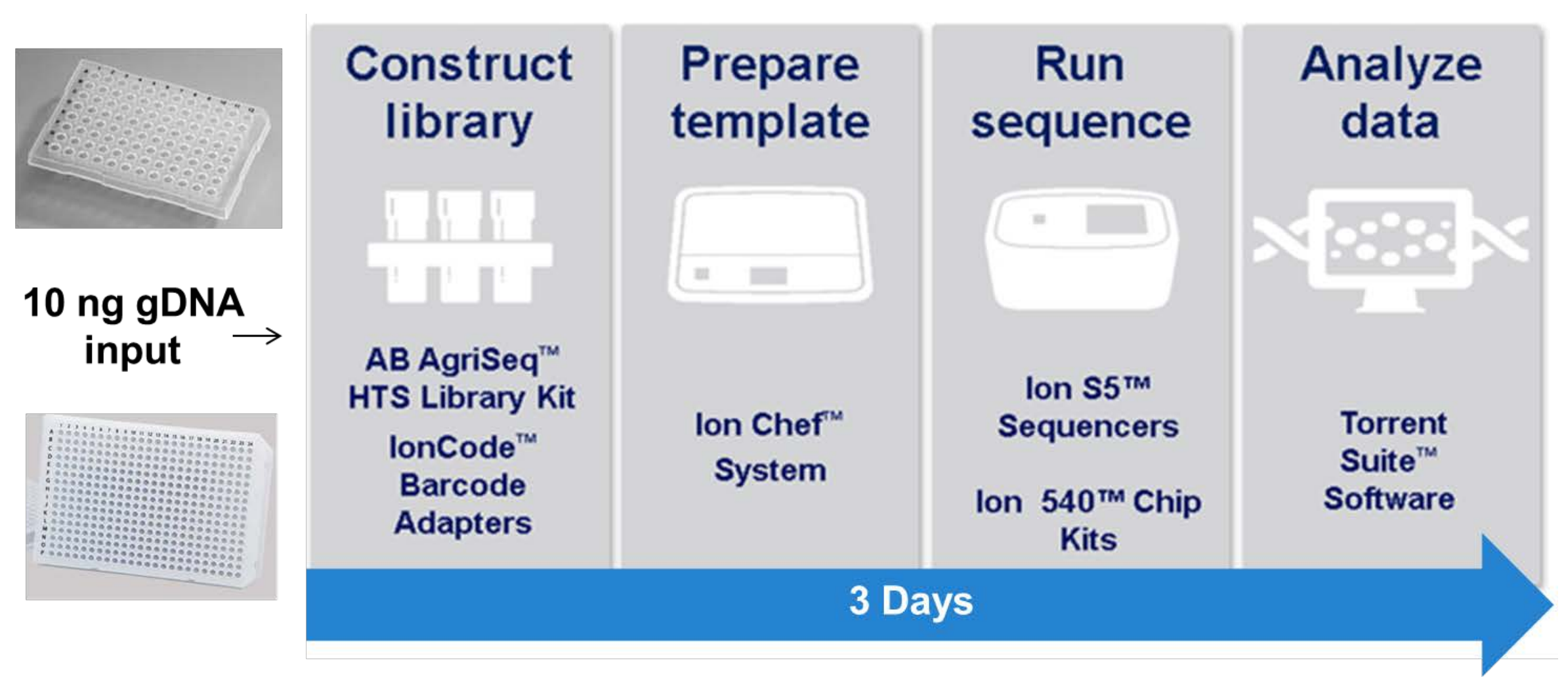
MATERIALS AND METHODS

The AgriSeq™ panels were designed using an automated process that optimizes a number of oligonucleotide properties (GC content, melting temperature (Tm), secondary structure, uniqueness etc.), amplicon properties (length, SNP position etc.) and their arrangements in the genome.

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.

Long INDEL framework was developed to detect INDELs that are longer than 50bp. Robustness of genotyping were validated using 22 canine long INDEL markers on 384 samples. Concordance was calculated based on the truth dataset (if available) of those samples.

Figure 1. Complete AgriSeq Sequencing Workflow



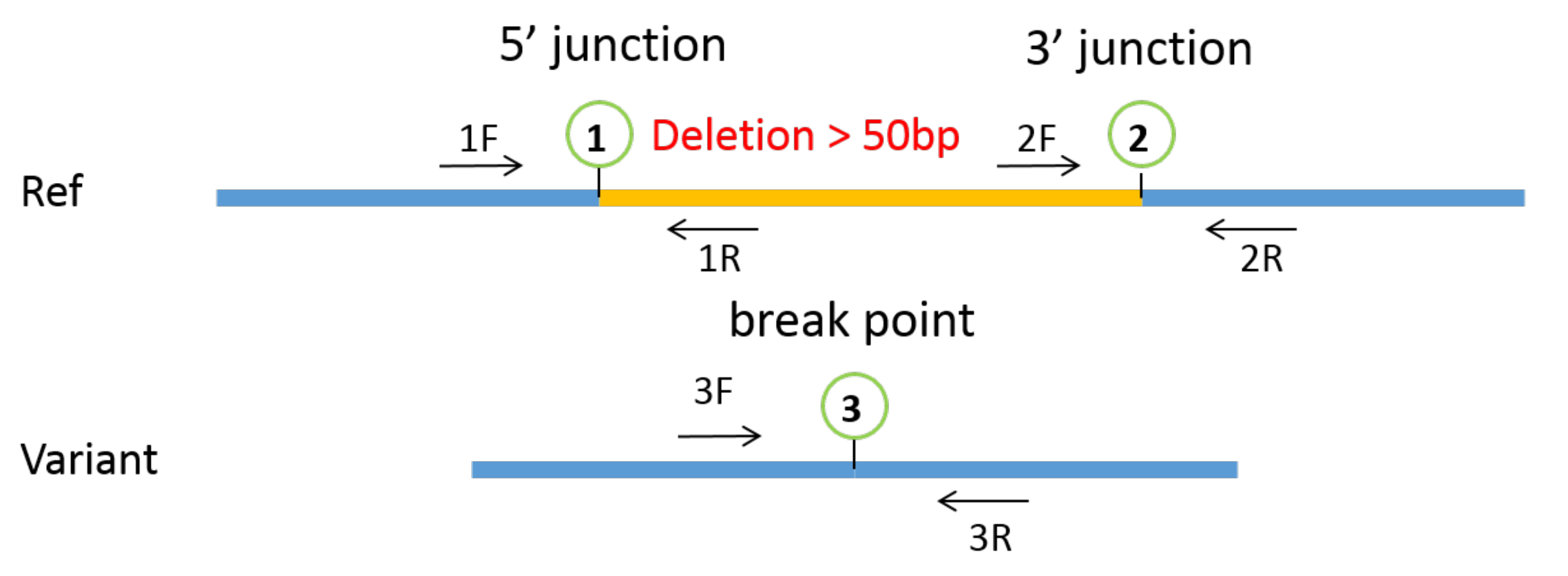
Libraries are constructed the AgriSeq™ HTS Library kit in either 96-well or 384-well format using 10 ng of genomic DNA. Different barcoded adapters are used for each library to allow simultaneous sequencing of hundreds of samples. Automated liquid handling platforms can be used for faster processing and compatibility. Template prep is performed on the Ion Chef™ System and samples are sequenced on the Ion S5™ XL System. Data is automatically analyzed and genotype calls generated using the Torrent Variant Caller plugin for all markers (long INDEL variants are called using customized workflow as explained below) across samples. The complete GBS workflow takes as little as three days from DNA to results.

Table 1. Sample Throughput Capability

	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

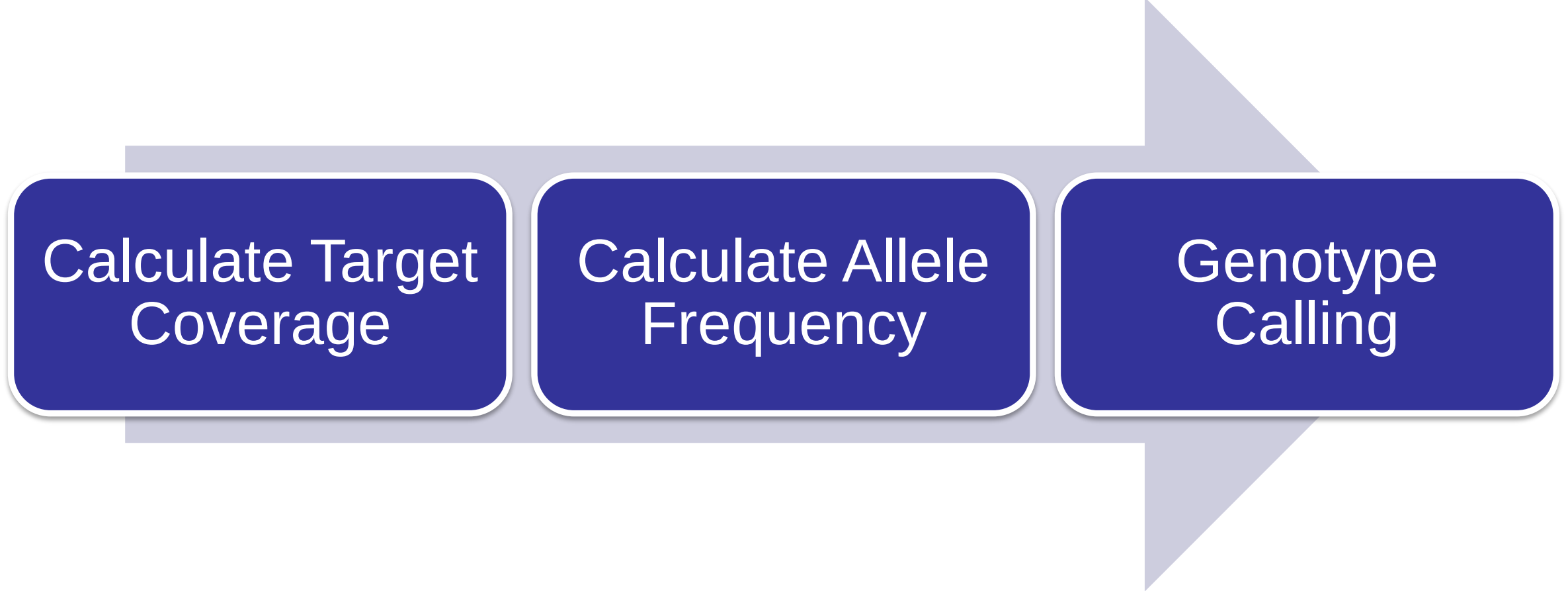
The maximum number of samples can be analyzed simultaneously varies based on the chip type used and the number of markers in the panel. This table shows the maximum recommended number of samples that can be analyzed at different marker densities per Ion 540 chip, per day and per year, assuming an average of 70 million reads/chip to achieve 100X average base coverage with one operator working a standard 8-hour shift, 5 days per week.

Figure 2. AgriSeq™ Long INDELs Design Solution



INDELs longer than 50bp are split into two targets where two pairs of primers are designed to cover the 5' junction (Target 1: 1F and 1R) and the 3' junction (Target 2: 2F and 2R). Amplicons from 1F+1R and 2F+2R are expected to amplify if the deletion is absent (reference allele) while amplicon from 1F+2R, is expected to amplify if the deletion is present (variant) . If a design is not possible for either Target 1 or 2, an extra Target 3 which includes the breakpoint will be created and primed for amplification. In this scenario, only one junction, the amplicon from 1F+1R (or 2F+2R), and 3F+3R will be used to make a call for wild type and variant respectively.

Figure 3. Long INDELs Variant Calling Pipeline



A customized workflow was developed for AgriSeq™ to make the call for the long INDEL genotypes. Genotypes calls are made using allele frequency and sequencing depth for each marker.

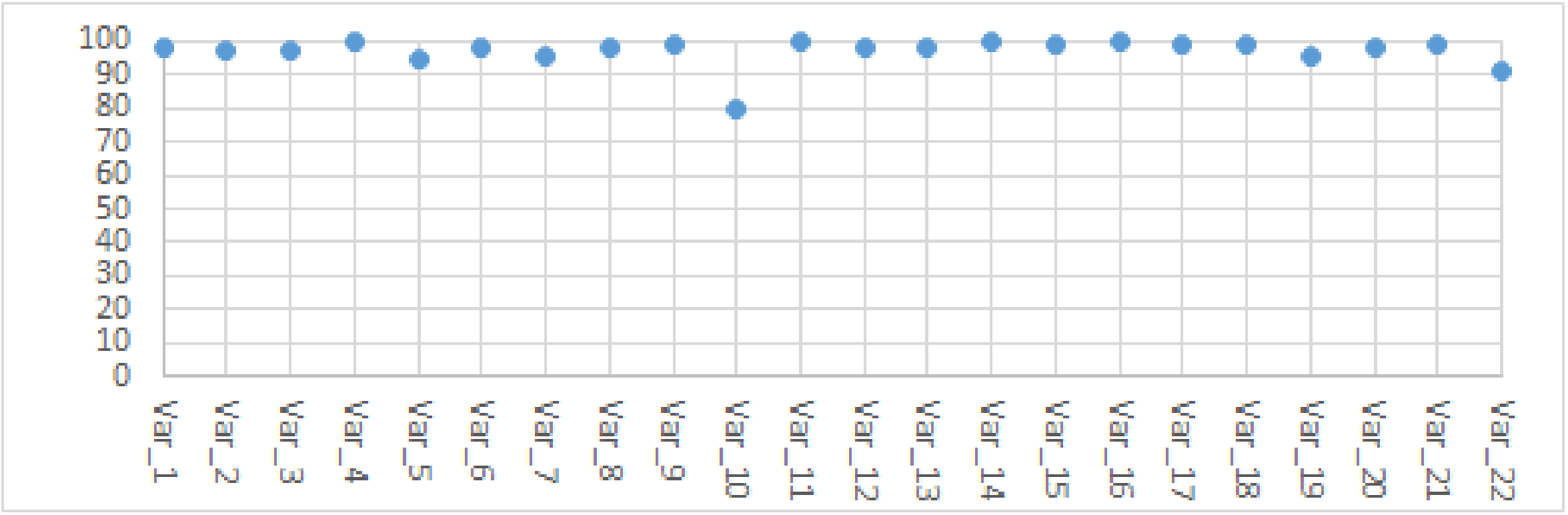
RESULTS

Table 2. Concordance (with truth data) and average call rate of canine long INDEL markers

Name	Type	Length	Expected Positives	Observed Positives	Call Rate (%)	Concordance (%)
Var_1	INS	62	0	0	98	NA
Var_2	INS	78	0	0	97.4	NA
Var_3	INS	159	0	0	96.9	NA
Var_4	INS	167	0	0	100	NA
Var_5	INS	179	0	0	94.3	NA
Var_6	DEL	180	0	0	98.4	NA
Var_7	INS	236	0	0	95.3	NA
Var_8	DEL	317	0	0	97.9	NA
Var_9	INS	3209	0	0	99	NA
Var_10	DEL	3,697	0	0	80	NA
Var_11	DEL	4,083	0	0	100	NA
Var_12	INS	4,228	0	0	98.4	NA
Var_13	DEL	7,799	29	29	97.9	100
Var_14	DEL	9,952	0	0	100	NA
Var_15	DEL	14,399	0	0	99	NA
Var_16	DEL	15,721	7	7	99.5	100
Var_17	DEL	17,932	0	0	99	NA
Var_18	DEL	39,800	0	0	99	NA
Var_19	DEL	129,788	4	4	95.3	100
Var_20	DEL	405,248	5	5	97.9	100
Var_21	INV	5,437,511	0	0	99	NA
Var_22	DEL	6,470,713	1	1	91	100

22 long INDEL canine markers were tested across 384 samples. 100% concordance comparing to truth data and 97% average call rate observed. Primer design and down-stream analysis were not impacted by the INDEL size.

Figure 4. Average Call Rate For Long INDEL Markers



Average call rates for each marker is calculated across all the samples. High call rate observed across samples with subtle variation in markers overall. The average call rate of all markers is 97%.

CONCLUSIONS

We expanded our AgriSeq™ design framework to enable the detection of large INDELs (>50bp). And we demonstrated the capability of long INDEL variant calling pipeline using a wide range of variant length (62 to 6.5Mb) with high concordance comparing to truth data.

High call rate across multiple samples with varying INDEL size indicates the reproducibility and flexibility of the method. AgriSeq™ targeted GBS offers customers end to end solution for genotyping diverse marker types simultaneously using same workflow.

REFERENCES

1. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6(5): e19379. https://doi.org/10.1371/journal.pone.0019379

TRADEMARKS/LICENSING

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.

ThermoFisher
SCIENTIFIC