

# Data Visualization Toolkit for Targeted Genotyping-By-Sequencing (GBS)

Prasad Siddavatham<sup>1</sup>, Haktan Suren<sup>1</sup>, Krishna Reddy Gujjula<sup>1</sup>, and Jeanette Schmidt<sup>2</sup>, <sup>1</sup>Thermo Fisher Scientific, 2130 Woodward Street, Austin, TX, USA, 78744, <sup>2</sup>Thermo Fisher Scientific, 3450 Central Expressway, Santa Clara, CA, USA, 95051

## ABSTRACT

Traditionally, high-throughput genotyping has been carried out by array based technologies or simplex PCR techniques. AgriSeq™ GBS with Ion Torrent next generation sequencing (NGS) technology offers a faster, flexible, multiplexing, customizable, cost-effective alternative solution to study fifty to five thousand markers. However, the data formats and complexity of NGS can make the scientific interpretation challenging. For the wider adoption and usability, we need a better way of summarizing and presenting the data for easier interpretation. Unfortunately, there are no tools available to comprehensively visualize the genotyping outputs. We developed a unified software tool to provide run summary metrics, genotype matrix table, genotypes in TOP/BOTTOM format, and additional features to view and compare the genotype calls.

Preliminary toolkit consists of the following features:

**Genotype Summary** - A summary report of the sequencing run with the high-level metrics of the sample call rates.

**GBSmatrix** - Actual genotype alleles are displayed in a sample-by-marker matrix of all the samples from a single sequencing run.

**GenotypeTB (TOP/BOTTOM)** - By default, AgriSeq™ reports genotype calls based on the positive strand alleles. To compare different genotyping technologies and calculate concordances, genotype calls are converted and displayed in TOP/BOTTOM format.

The plugin enables researchers to visualize, interpret and troubleshoot the genotyping results better. In doing so, the tool helps them leveraging the informative power of NGS applied to targeted GBS. The data visualization toolkit will be distributed as an Ion Torrent Software Suite Plug-In.

## INTRODUCTION

GBS transcended population genetics to population genomics and aiding breeding endeavors in several important crops [1]. AgriSeq™ GBS allows users to multiplex up to 768 samples per sequencing run that contain fifty to few thousand markers. Multiple variant types can be combined into a single panel. Workflow is optimized to get sample to answers in 2-3 days (Fig. 1). However, no integrated software tool is available to visualize and explore genotype calls from NGS data. We developed the AgriSeq™ Toolkit with the following objectives.

- Generate genotyping calls in multiple formats
- Eliminate reliance on external software for data processing
- Provide overall sequencing run summary metrics
- Opportunity to add custom metadata to the genotype results

## AgriSeq™ WORKFLOW

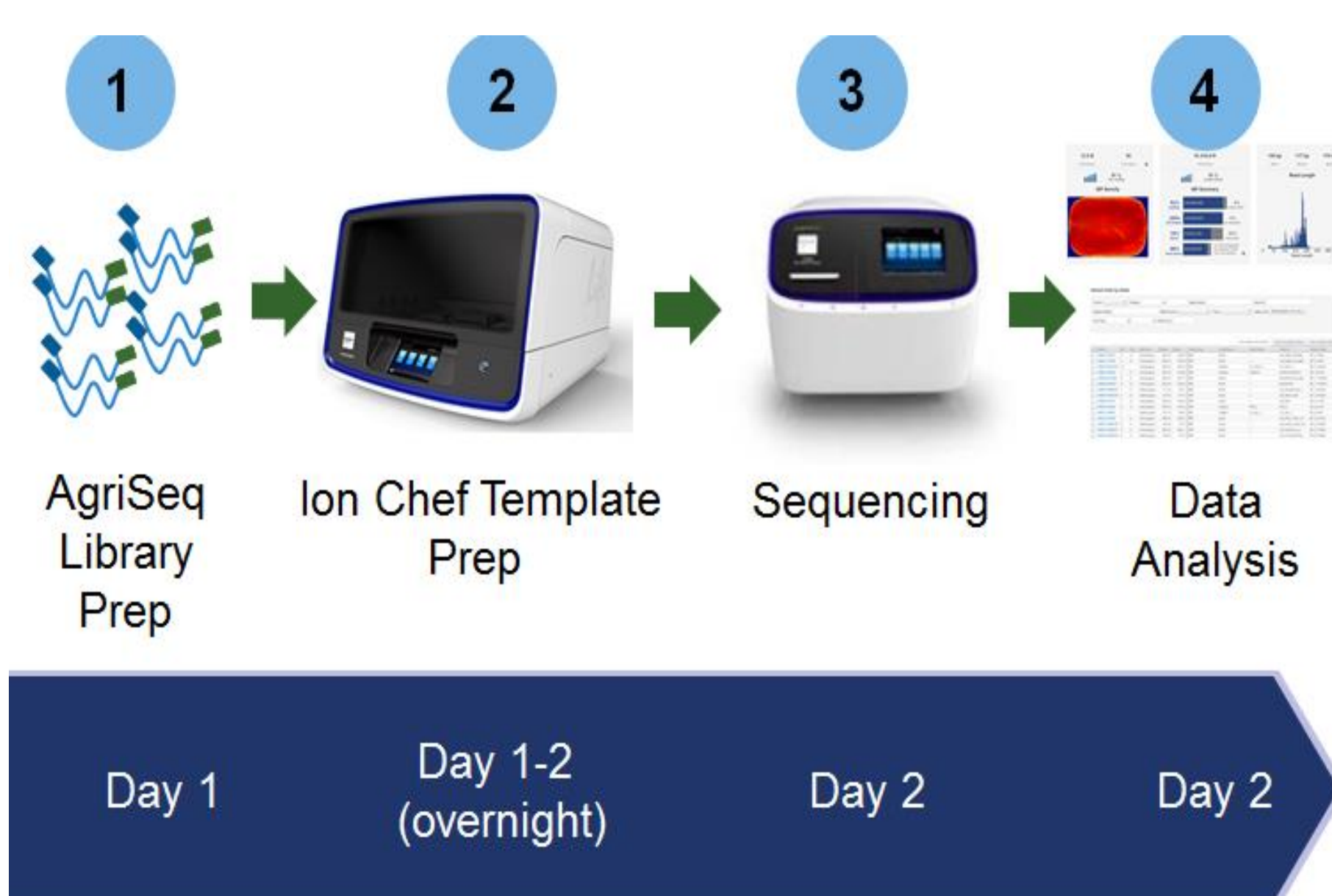


Figure 1. Workflow for performing sequencing and analysis

## MATERIALS AND METHODS

The sequence data is analyzed using Torrent Suite Software (TSS - version 5.10). Per sample and marker mapping statistics are calculated using the coverage analysis plugin and genotype calls are generated with the Torrent Variant Caller (TVC) plugin. AgriSeq™ Toolkit plugin aggregated data from individual plugins. All the plugins are available in the TSS.

## RESULTS

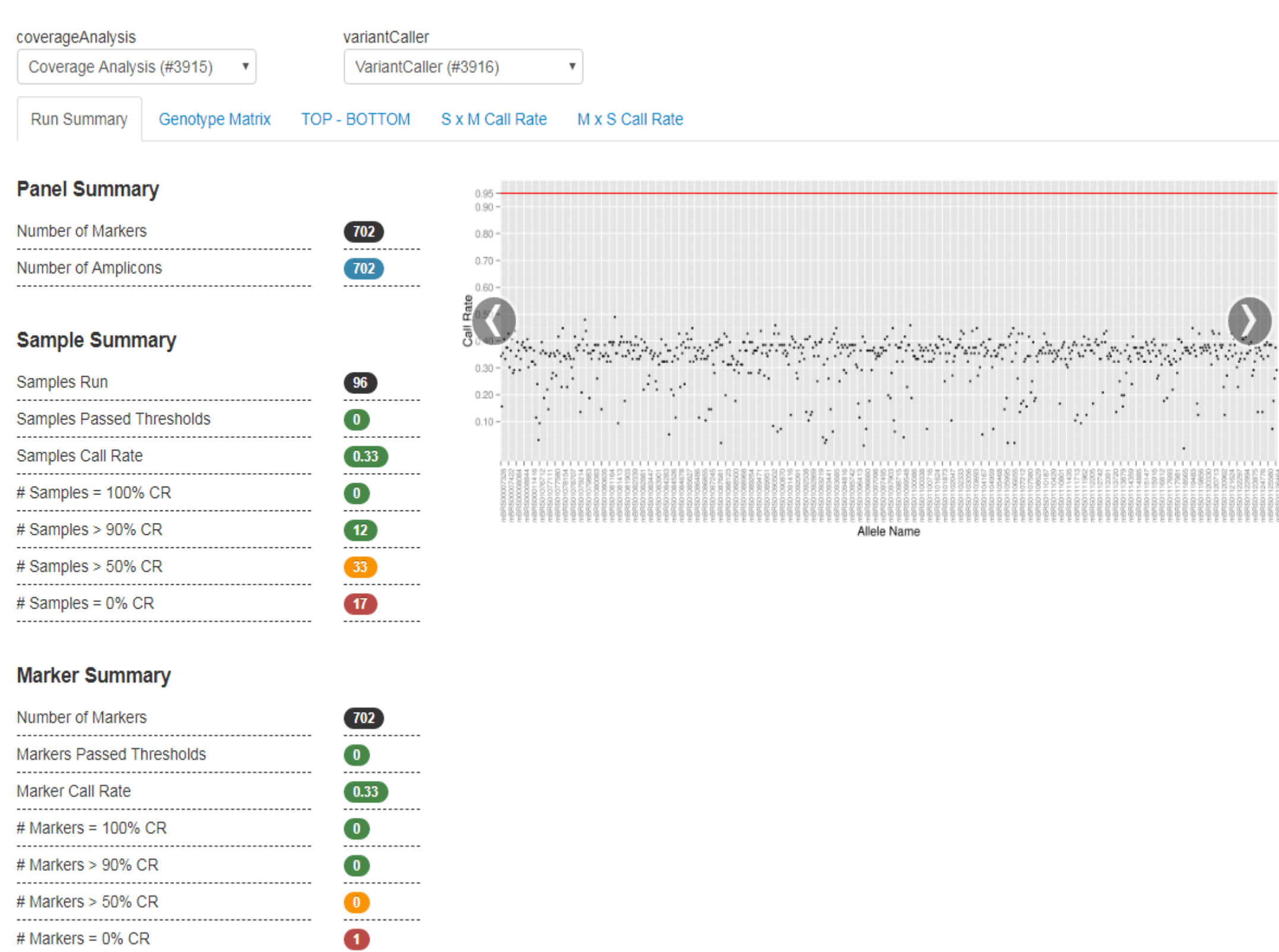


Fig 2. Overview of AgriSeq™ GBS run summary. Panel performance can be evaluated at both sample level and marker level. Based on user defined thresholds, the summary metrics are dynamically calculated and color codes the output.

Sample Name	Marker1	Marker2	Marker3	Marker4	Marker5	Marker6
ABC1	A/G	A/G	A/G	A/G	A/G	A/G
ABC2	T/A	T/A	T/A	T/A	T/A	T/A
ABC3	A/G	A/G	A/G	A/G	A/G	A/G
ABC4	G/G	G/G	G/G	G/G	G/G	G/G
ABC5	A/A	A/A	A/A	A/A	A/A	A/A
ABC6	-/-	-/-	-/-	-/-	-/-	-/-
ABC7	A/G	A/G	A/G	A/G	A/G	A/G
ABC8	A/G	A/G	A/G	A/G	A/G	A/G
ABC9	A/G	A/G	A/G	A/G	A/G	A/G
ABC10	A/G	A/G	A/G	A/G	A/G	A/G
ABC11	A/G	A/G	A/G	A/G	A/G	A/G
ABC12	A/A	A/A	A/A	A/A	A/A	A/A
ABC13	G/G	G/G	G/G	G/G	G/G	G/G
ABC14	-/-	-/-	-/-	-/-	-/-	-/-
ABC15	A/G	A/G	A/G	A/G	A/G	A/G

Sample Name	Marker1	Marker2	Marker3	Marker4	Marker5	Marker6
ABC1	AB	AB	AB	AB	AB	AB
ABC2	AB	AB	AB	AB	AB	AB
ABC3	AB	AB	AB	AB	AB	AB
ABC4	BB	BB	BB	BB	BB	BB
ABC5	AA	AA	AA	AA	AA	AA
ABC6	-	-	-	-	-	-
ABC7	AB	AB	AB	AB	AB	AB
ABC8	AB	AB	AB	AB	AB	AB
ABC9	AB	AB	AB	AB	AB	AB
ABC10	AB	AB	AB	AB	AB	AB
ABC11	AB	AB	AB	AB	AB	AB
ABC12	AA	AA	AA	AA	AA	AA
ABC13	BB	BB	BB	BB	BB	BB
ABC14	-	-	-	-	-	-
ABC15	AB	AB	AB	AB	AB	AB

Table 1. A. The actual genotypes are displayed in sample-by-marker matrix format. This format is used as input in other down stream analysis (e.g. Genome Wide Association Study) B. Genotypes from top table are converted to TOP/BOTTOM format, which enables easy genotyping results comparison between different technologies.

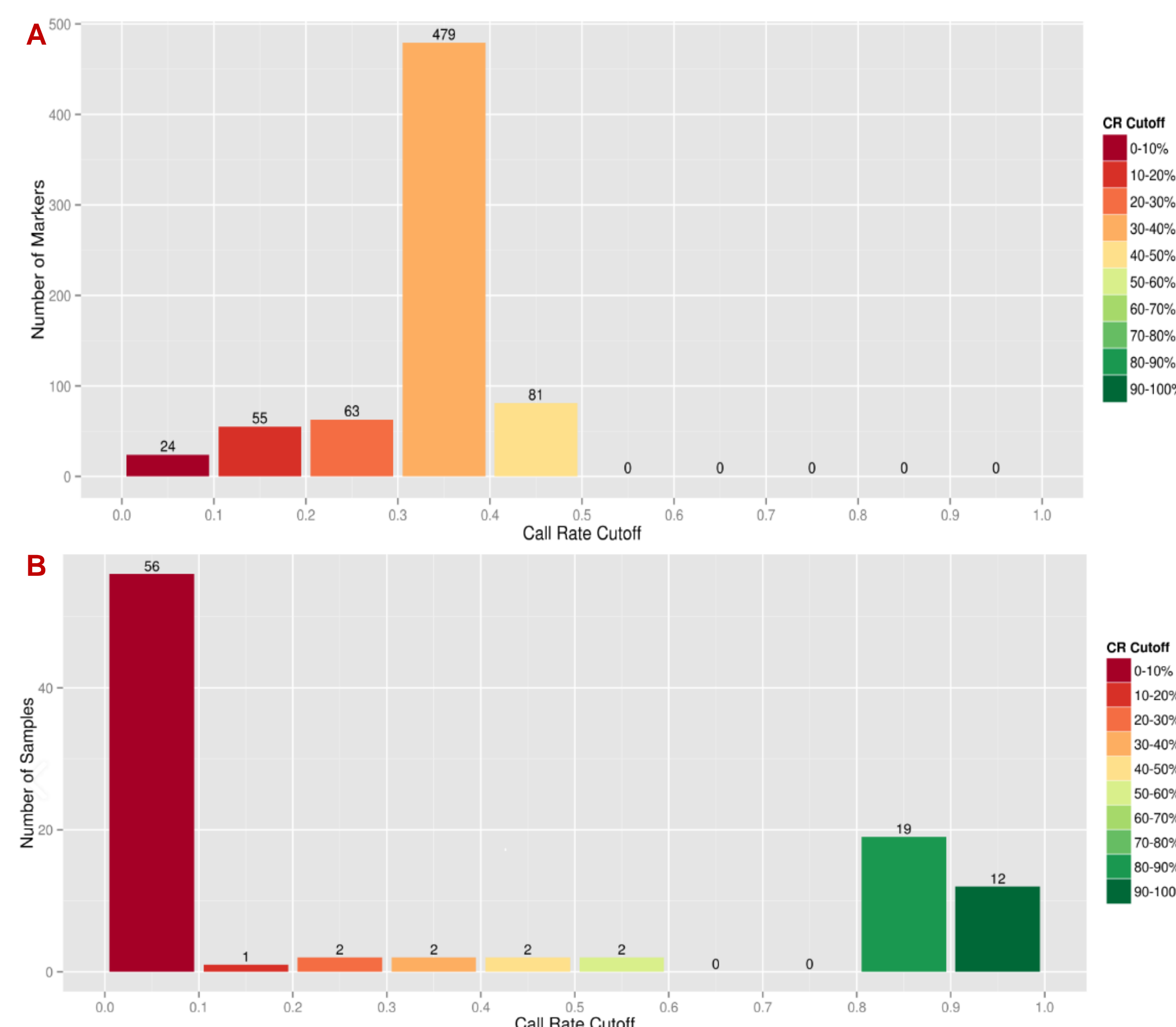


Fig 3. A. Marker call rate - Genotype calls across all the samples from the panel for a specific marker. B. Sample call rate - Genotype calls across all the markers from the panel for a specific sample.

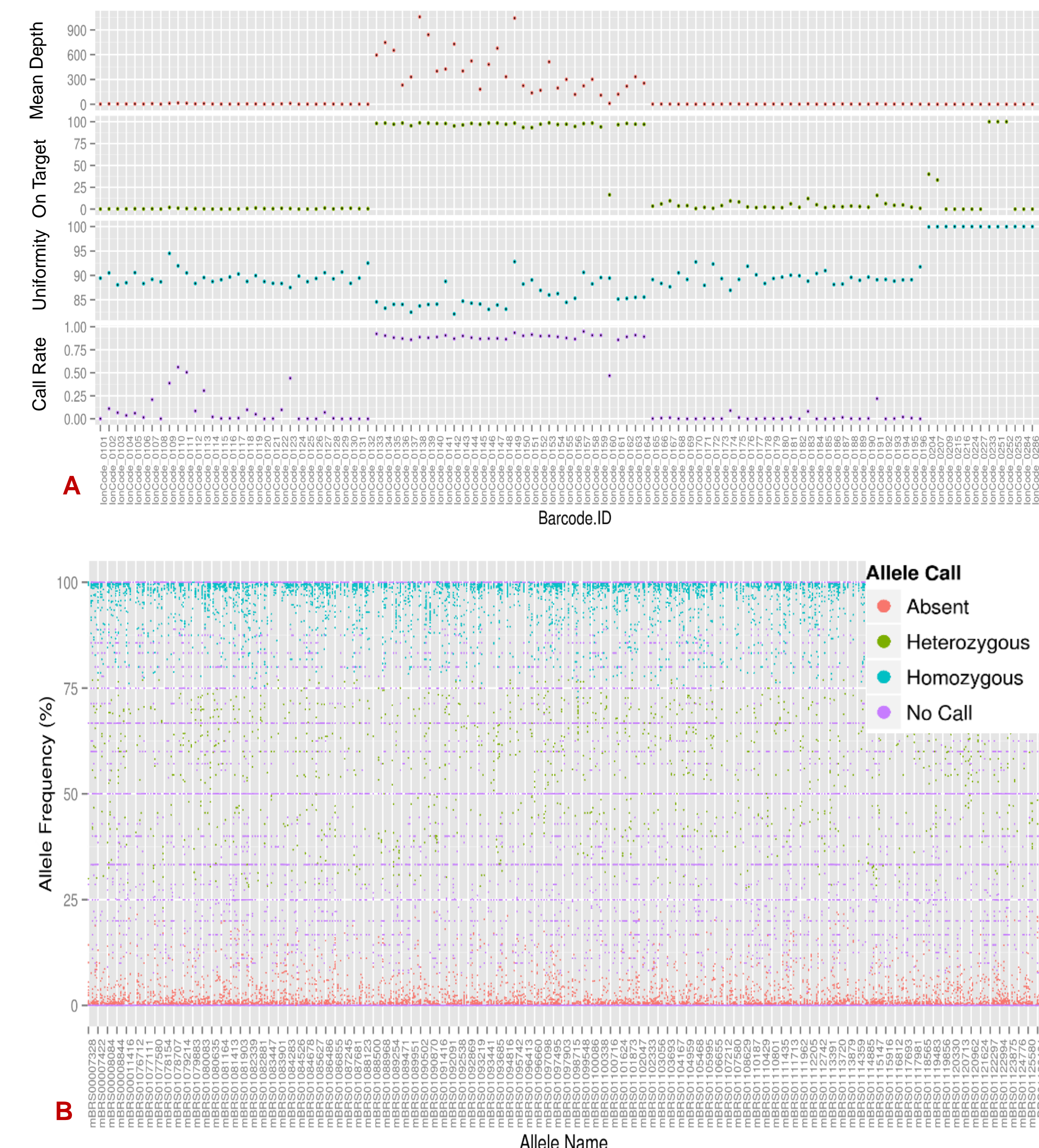


Figure 4. Deeper insights into the data can be further visually explored at global level with A. sample specific metrics (Call rate, Uniformity, On Target mapped reads and Mean depth) and B. marker specific metrics (Allele frequency). These plots help to visualize the potential reasons for under-performance of samples and markers.

## CONCLUSIONS

- We developed the AgriSeq™ Toolkit to provide GBS run summary metrics, genotype matrix table, genotypes in TOP/BOTTOM format, and additional features to view and compare the genotype calls.
- Data exploration to understand the reasons for poor-performance of some samples/markers is feasible.
- The AgriSeq™ Toolkit will be distributed as an Ion Torrent Software Suite Plug-In.

## REFERENCES

1. Voss-Fels K. and Snowden R.J. (2016) Understanding and utilizing crop genome diversity via high-resolution genotyping. Plant Biotechnol. J. 14, 1086–1094.

## 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. © 2019 Thermo Fisher Scientific Inc. All rights reserved. All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified.