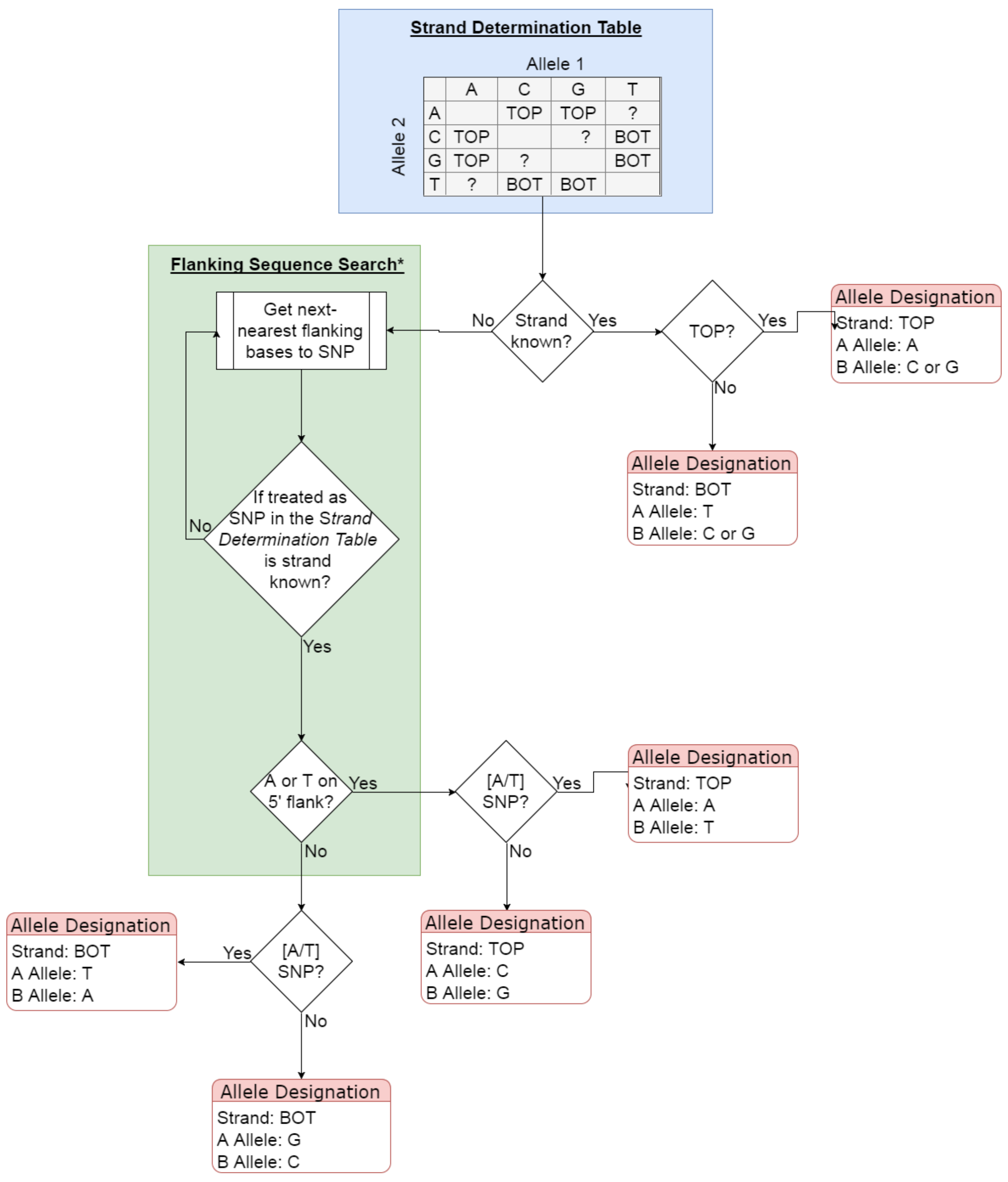
New output formats for Axiom genotyping arrays

Ali Pirani, Joseph M Foster, Alessandro Davassi, Brant Wong, Mohini A Patil, and Luis Jevons
Thermo Fisher Scientific, 3450 Central Expressway, Santa Clara, CA 95051

INTRODUCTION

The high throughput agricultural genotyping landscape encompasses a broad range of applications and technical platforms. One of the major challenges of adopting a new platform or performing meta-analyses is data format congruity. Biallelic genotypes are recorded in one of three ways; "AA", "AB" and "BB" call codes, "0", "1", and "2" numeric call codes and base calls "A", "T", "G" or "C". For call codes and numeric call codes, the A and B alleles must be designated. Historically, two formats have dominated the designation of variant alleles; "Forward" and "TOP". For bi-allelic SNPs this can create a situation where the "A" allele designated by one format differs from the other.

Figure 1. TOP/BOT format allele designation



Strand Determination Table

Allele 1 / Allele 2

	A	C	G	T
A		TOP	TOP	?
C	TOP		?	BOT
G	TOP	?		BOT
T	?	BOT	BOT	

Flanking Sequence Search*

Get next-nearest flanking bases to SNP

If treated as SNP in the *Strand Determination Table* is strand known?

Strand known?

TOP?

A or T on 5' flank?

[A/T] SNP?

Yes

No

Allele Designation
Strand: TOP
A Allele: A
B Allele: C or G

Allele Designation
Strand: BOT
A Allele: T
B Allele: C or G

Allele Designation
Strand: TOP
A Allele: A
B Allele: T

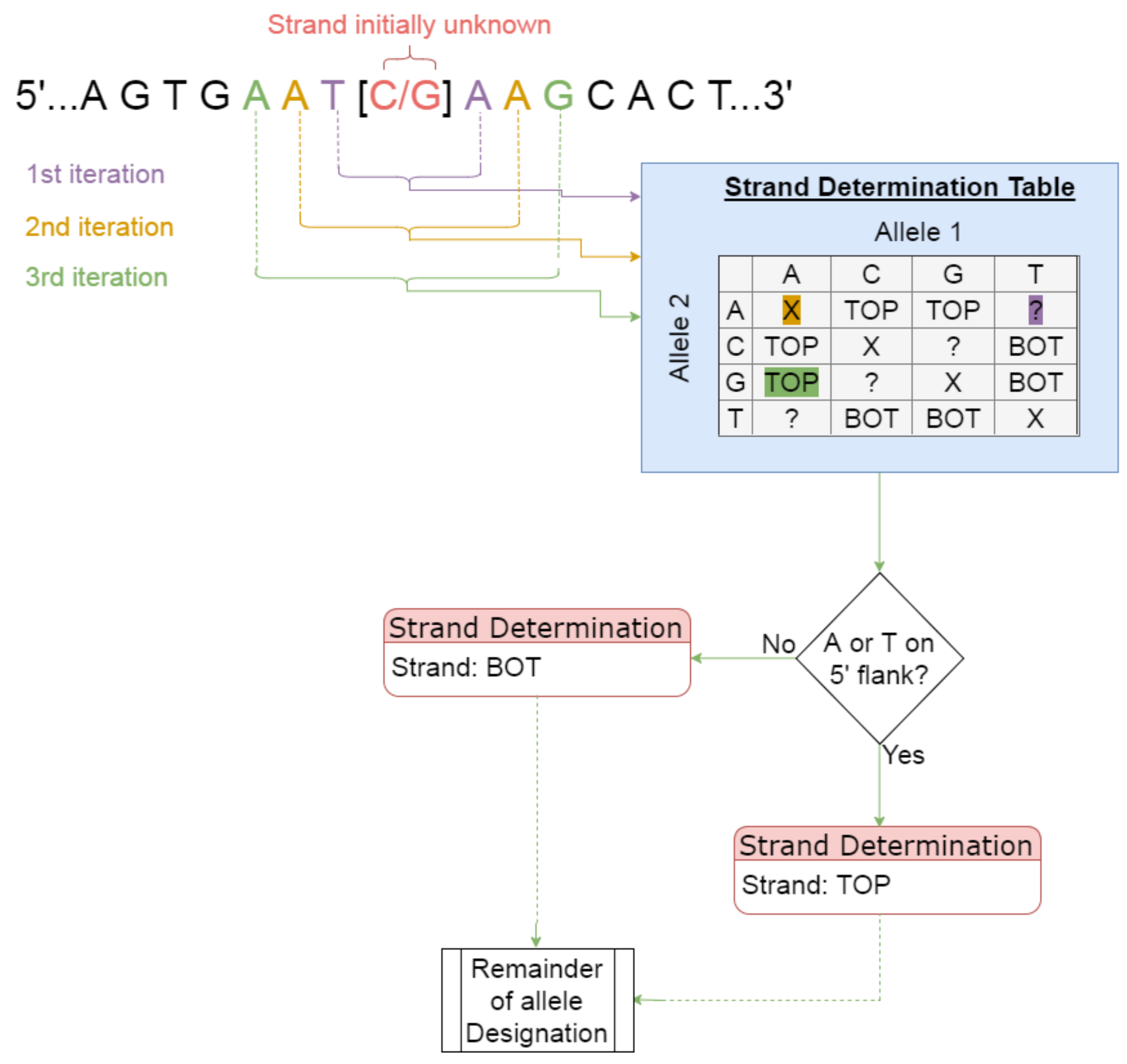
Allele Designation
Strand: TOP
A Allele: C
B Allele: G

Allele Designation
Strand: BOT
A Allele: T
B Allele: A

Allele Designation
Strand: BOT
A Allele: G
B Allele: C

The flow diagram in *Figure 1* describes the process for allele designation of bi-allelic SNPs for the TOP/BOT format. Initially strand determination is done by the *Strand Determination Table*. For [A/C] and [A/G] SNPs the strand is defined as TOP, for [T/C] and [T/G] SNPs the strand is defined as BOT. Where strand determination is possible in this manner, allele designation follows such that any A or T base is considered the A allele and the other base that constitutes the SNP is designated as the B allele. For [A/T] and [G/C] SNPs the strand is unknown and the flanking sequence is used to determine strand by the *Flanking Sequence Search* process (more detailed view in *Figure 2*). For SNPs determined to be on the TOP strand, the A or C base is designated as the A allele and the T or G base as the B allele for [A/T] and [G/C] SNPs respectively. For the SNPs determined to be on the BOT strand, the reverse is true; the T or G base is designated as the A allele and the A or C base as the B allele for [A/T] and [G/C] SNPs respectively.

Figure 2. TOP/BOT strand determination for [A/T] and [G/C] SNPs



Strand initially unknown

5'...A G T G A A T [C/G] A A G C A C T...3'

1st iteration

2nd iteration

3rd iteration

Strand Determination Table

Allele 1 / Allele 2

	A	C	G	T
A	X	TOP	TOP	?
C	TOP	X	?	BOT
G	TOP	?	X	BOT
T	?	BOT	BOT	X

A or T on 5' flank?

No

Yes

Strand Determination
Strand: BOT

Strand Determination
Strand: TOP

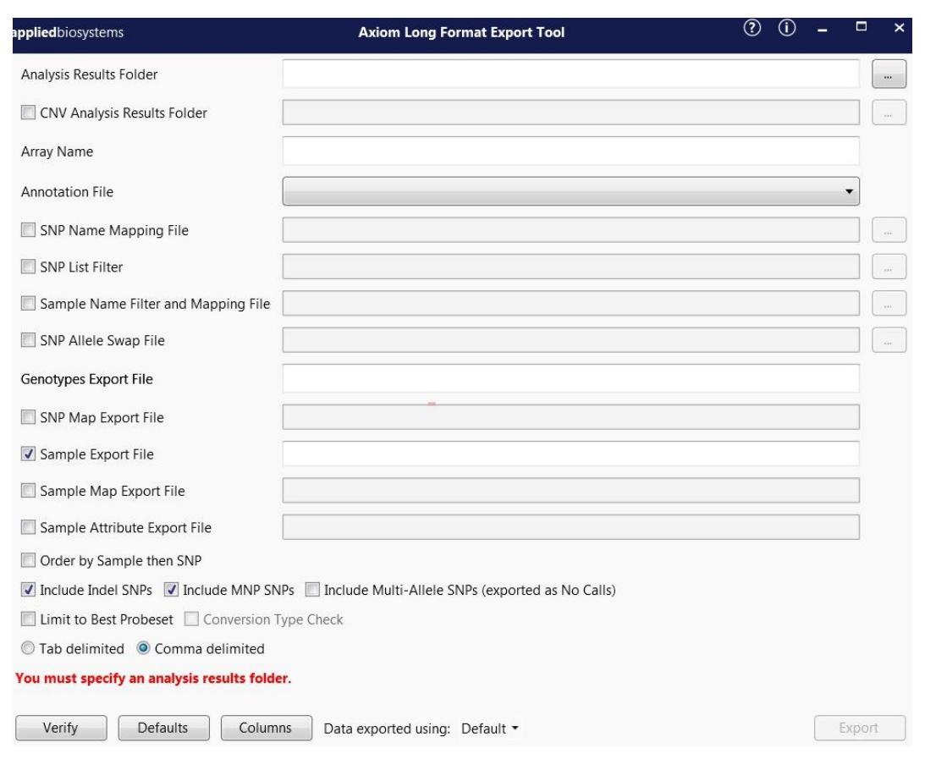
Remainder of allele Designation

Figure 2 shows the strand determination process for [A/T] and [G/C] SNPs. Starting with the nearest pair of bases either side of the SNP, the *Strand Determination Table* is checked to see if a strand determination can be made. For each time a strand determination cannot be made the algorithm increments 1 position further away from the SNP in both directions and re-checks the strand determination table until a strand determination would be made. Once this iterative process is complete, if the A or T base is in the 5' flanking sequence the strand is determined as TOP. Conversely, if the A or T allele is in the 3' flanking sequence the strand is determined to be BOT.

AXIOM LONG FORMAT EXPORT TOOL (AxLE)

To support cross-platform high throughput genotyping analysis, Thermo Fisher Scientific has developed the Axiom Long format Export Tool (AxLE)[1]; a companion application to Axiom™ Analysis Suite[2]. The tool converts Axiom genotype data from native "Forward" format to the "TOP" format based on the polymorphism itself, or the contextual surrounding sequence and designates the A/B allele. The tool also converts the standard Axiom output into a format that is similar to the long format options for other platforms. This makes Axiom genotyping easier to integrate with existing downstream analysis pipelines and large scale meta-analysis of several cross-platform datasets.

Figure 3. AxLE Tool Usage



appliedbiosystems — Axiom Long Format Export Tool

Analysis Results Folder
□ CNV Analysis Results Folder
Array Name
Annotation File
□ SNP Name Mapping File
□ SNP List Filter
□ Sample Name Filter and Mapping File
□ SNP Allele Swap File
Genotypes Export File
□ SNP Map Export File
☑ Sample Export File
□ Sample Map Export File
□ Sample Attribute Export File
□ Order by Sample then SNP
☑ Include Indel SNPs ☑ Include MNP SNPs □ Include Multi-Allele SNPs (exported as No Calls)
□ Limit to Best Probeset □ Conversion Type Check
○ Tab delimited ● Comma delimited

You must specify an analysis results folder.

Verify | Defaults | Columns | Data exported using: Default | Export

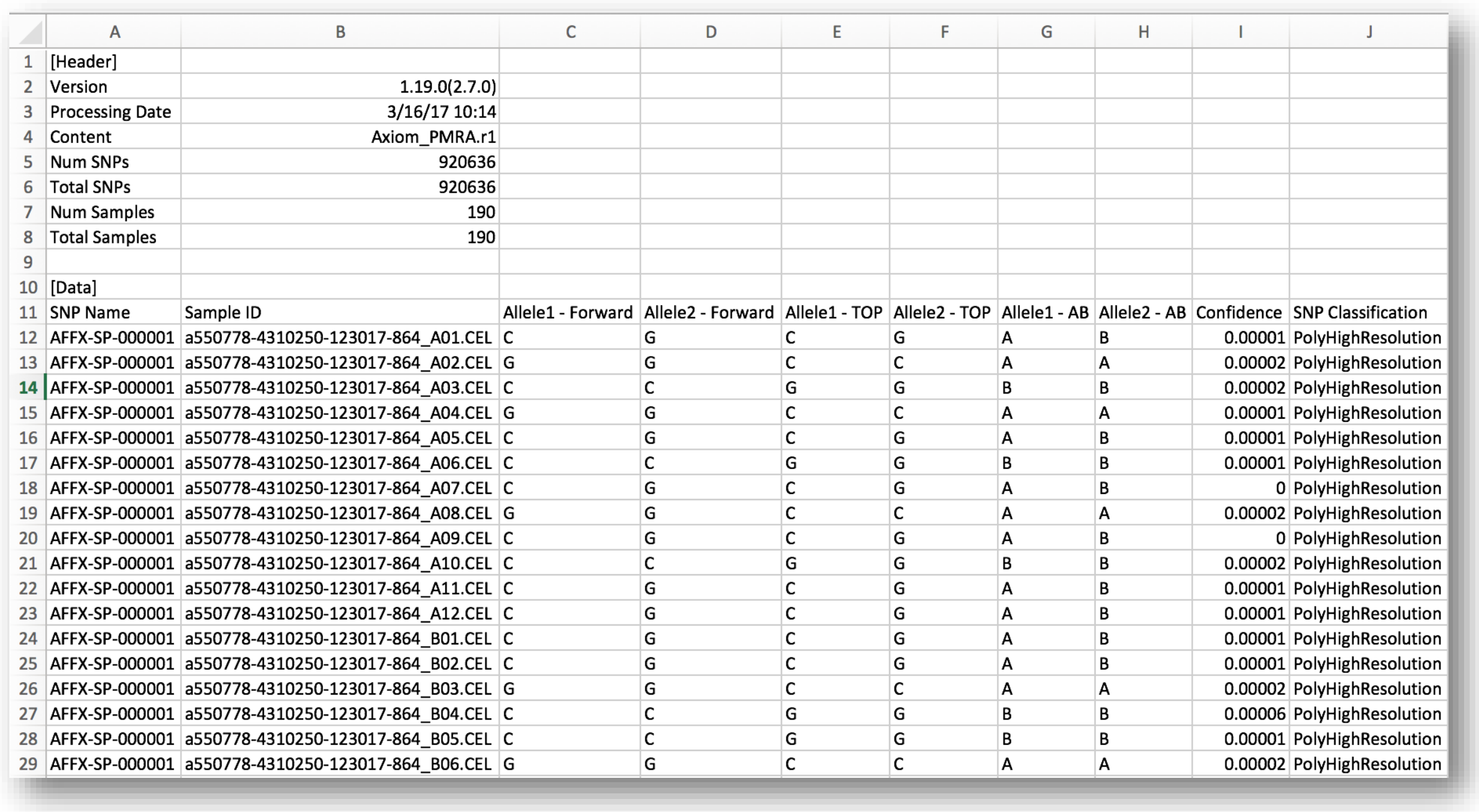
Figure 3 shows a screen shot of the AxLE tool. After executing a Best Practices Workflow using Axiom Analysis Suite, the AxLE tool can be accessed via the "External Tools" menu of Axiom Analysis Suite. Some of the following steps are taken to generate Long format:

1. "Analysis Results": Select the appropriate results folder of the analysis to be converted to Long format.
2. The "Array Name" and "Annotation File" will automatically be populated.
3. Optionally a "SNP Name Mapping File" can be selected. If selected it will replace the native probeset_ids with alternate SNP identifiers in the Long format output.
4. Optionally the analysis results can be filtered by a "SNP List" to return only those SNPs in the Long format output.
5. A name must assigned to the Long format output file.
6. Review the other options and click the "Export" button.

AxLE output consists of:

- A descriptive header
- SNP Name: the default is probeset_id e.g. AX-123456789, but these can be modified to a user-defined value by using a "SNP Name Mapping File"
- Sample ID: the default is the CEL file name
- Allele 1/2 – Forward: Base call relative to Forward Strand
- Allele 1/2 – Top: Base call normalized to TOP strand
- Allele 1/2 - A/B: Axiom designated A/B allele call
- Confidence: AxiomGT1 algorithm confidence score for this genotype assignment
- SNP Classification: SNPolisher[3] conversion type (category)

Figure 4. AxLE Example Output



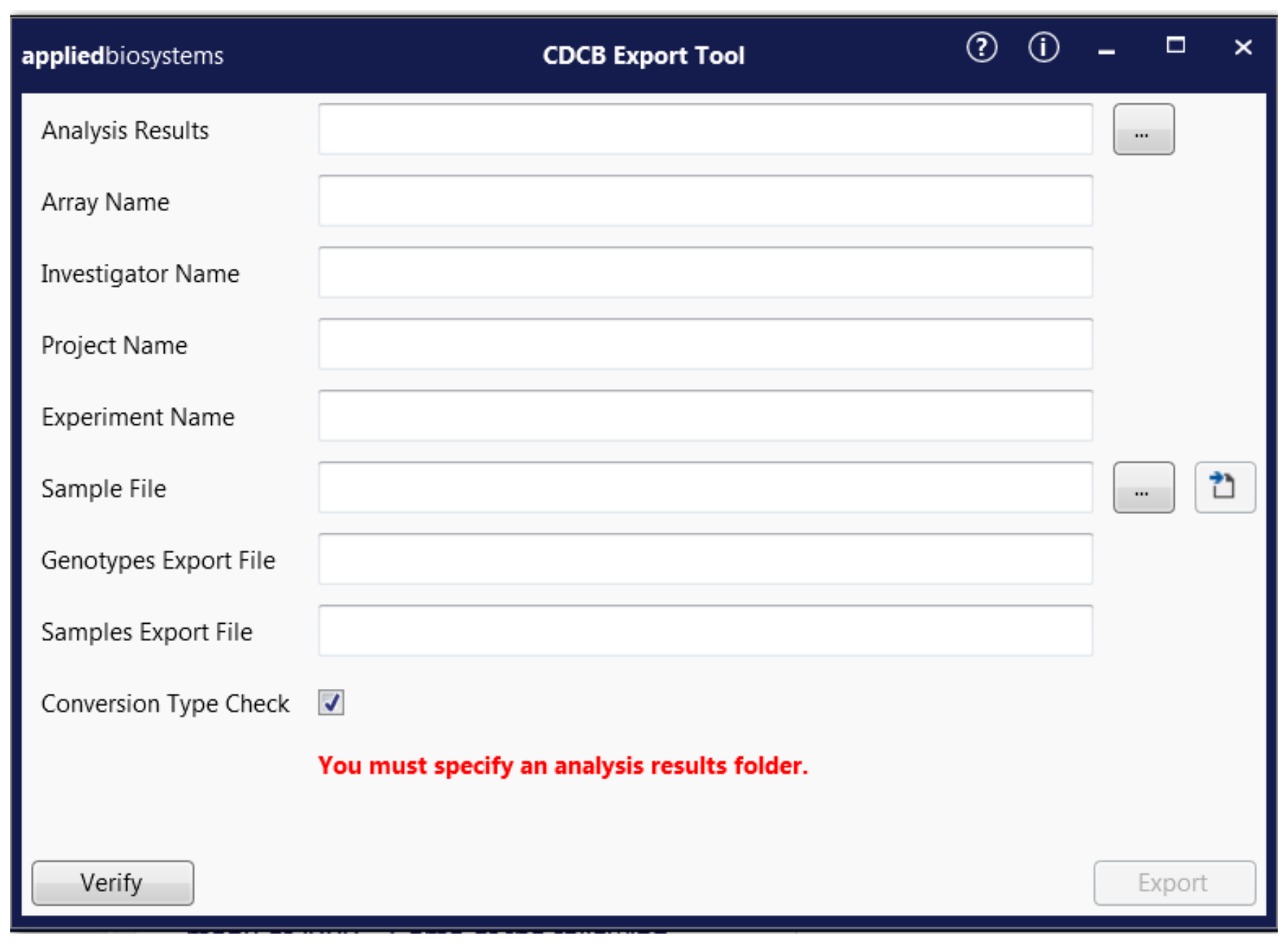
	A	B	C	D	E	F	G	H	I	J
1	[Header]									
2	Version	1.19.0(2.7.0)								
3	Processing Date	3/16/17 10:14								
4	Content	Axiom_PMRA.r1								
5	Num SNPs	920636								
6	Total SNPs	920636								
7	Num Samples	190								
8	Total Samples	190								
9										
10	[Data]									
11	SNP Name	Sample ID	Allele1 - Forward	Allele2 - Forward	Allele1 - TOP	Allele2 - TOP	Allele1 - AB	Allele2 - AB	Confidence	SNP Classification
12	AFFX-SP-000001	a550778-4310250-123017-864_A01.CEL	C	G	C	G	A	B	0.00001	PolyHighResolution
13	AFFX-SP-000001	a550778-4310250-123017-864_A02.CEL	G	G	C	C	A	A	0.00002	PolyHighResolution
14	AFFX-SP-000001	a550778-4310250-123017-864_A03.CEL	C	C	G	G	B	B	0.00002	PolyHighResolution
15	AFFX-SP-000001	a550778-4310250-123017-864_A04.CEL	G	G	C	C	A	A	0.00001	PolyHighResolution
16	AFFX-SP-000001	a550778-4310250-123017-864_A05.CEL	C	G	C	G	A	B	0.00001	PolyHighResolution
17	AFFX-SP-000001	a550778-4310250-123017-864_A06.CEL	C	C	G	G	B	B	0.00001	PolyHighResolution
18	AFFX-SP-000001	a550778-4310250-123017-864_A07.CEL	C	G	C	G	A	B	0	PolyHighResolution
19	AFFX-SP-000001	a550778-4310250-123017-864_A08.CEL	G	G	C	C	A	A	0.00002	PolyHighResolution
20	AFFX-SP-000001	a550778-4310250-123017-864_A09.CEL	C	G	C	G	A	B	0	PolyHighResolution
21	AFFX-SP-000001	a550778-4310250-123017-864_A10.CEL	C	C	G	G	B	B	0.00002	PolyHighResolution
22	AFFX-SP-000001	a550778-4310250-123017-864_A11.CEL	C	G	C	G	A	B	0.00001	PolyHighResolution
23	AFFX-SP-000001	a550778-4310250-123017-864_A12.CEL	C	G	C	G	A	B	0.00001	PolyHighResolution
24	AFFX-SP-000001	a550778-4310250-123017-864_B01.CEL	C	G	C	G	A	B	0.00001	PolyHighResolution
25	AFFX-SP-000001	a550778-4310250-123017-864_B02.CEL	C	G	C	G	A	B	0.00001	PolyHighResolution
26	AFFX-SP-000001	a550778-4310250-123017-864_B03.CEL	G	G	C	C	A	A	0.00002	PolyHighResolution
27	AFFX-SP-000001	a550778-4310250-123017-864_B04.CEL	C	C	G	G	B	B	0.00006	PolyHighResolution
28	AFFX-SP-000001	a550778-4310250-123017-864_B05.CEL	C	C	G	G	B	B	0.00001	PolyHighResolution
29	AFFX-SP-000001	a550778-4310250-123017-864_B06.CEL	G	G	C	C	A	A	0.00002	PolyHighResolution

The table in *Figure 4* demonstrates the output of the Axiom Long format Export tool (AxLE). Each row represents a genotype call for a single SNP in single sample, described by both the Axiom native Forward format and the "TOP" format. In addition, the genotype call confidence as determined by the AxiomGT1 algorithm and the SNP classification by SNPolisher[3] is reported.

COUNCIL ON DAIRY CATTLE BREEDING (CDCB) EXPORT TOOL

A clear requirement for the standardisation of allele designation is in the downstream application of genotyping data to genetic evaluation systems where mixing of the formats could be disastrous to the prediction of economically important traits. To support this specific use case in dairy cattle, Thermo Fisher Scientific has developed the CDCB (Council on Dairy Cattle Breeding) export tool[4]. Once an analysis has been completed in Axiom Analysis Suite[2], the CDCB export tool[4] performs three operations. Firstly, the "A/B" allele designations are swapped where the native "Forward" strand annotation differs to "TOP" based on a predefined list of affected markers. Secondly, the native SNP identifiers are mapped to the CDCB approved SNP identifiers. This occurs when a SNP has previously been submitted to the CDCB as part of another supported array and that name takes priority. Finally, it formats the calls and generates a sample sheet to enable direct upload to the Council on Dairy Cattle Breeding website. The tool is capable of consuming data from any Axiom catalogue bovine array and also custom bovine designs and is freely available to download.

Figure 5. CDCB Export Tool Usage



appliedbiosystems — CDCB Export Tool

Analysis Results
Array Name
Investigator Name
Project Name
Experiment Name
Sample File
Genotypes Export File
Samples Export File
Conversion Type Check ☑

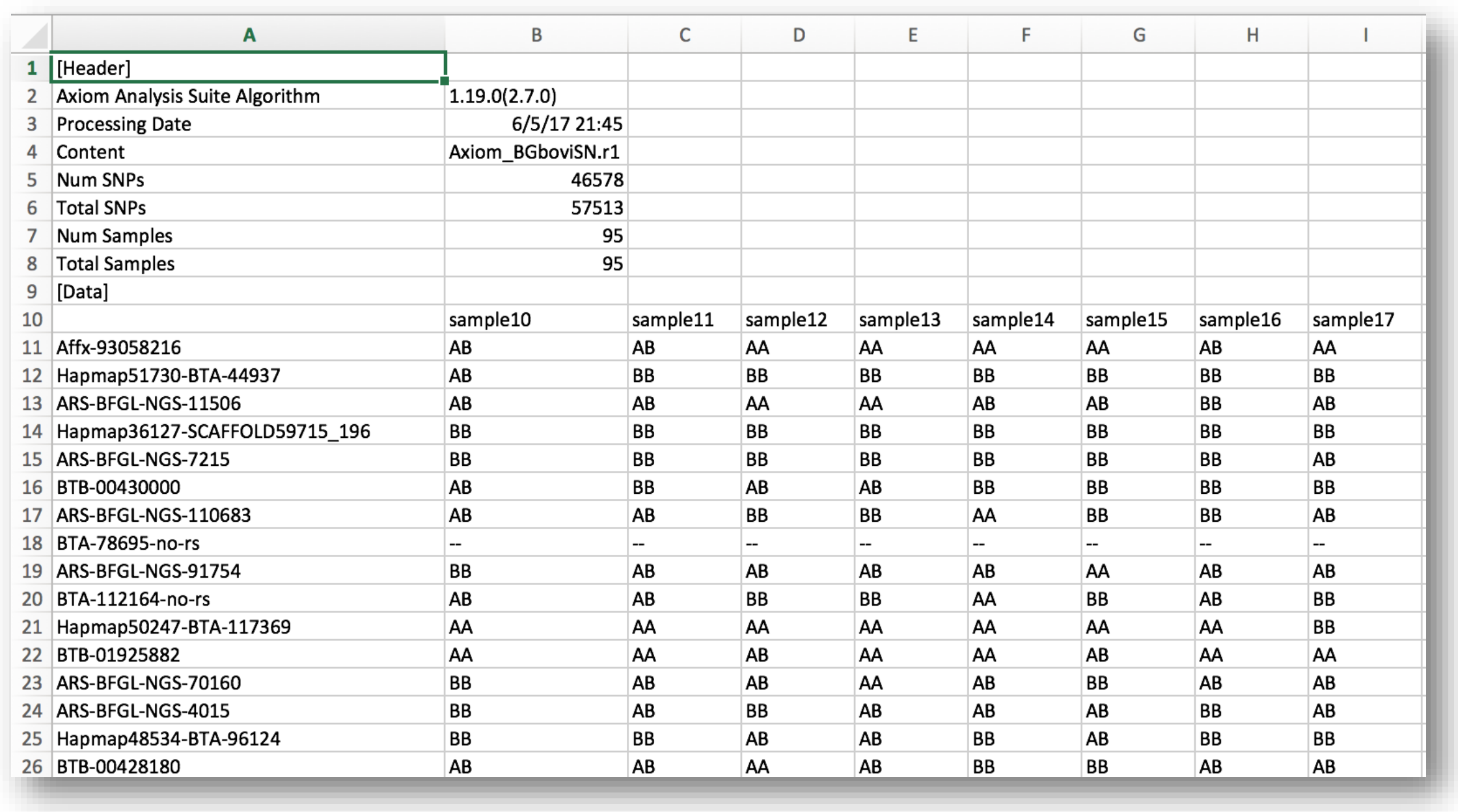
You must specify an analysis results folder.

Verify | Export

Figure 5 shows a screen shot of the CDCB export tool. After executing a Best Practices Workflow using Axiom Analysis Suite, the CDCB Export Tool can be accessed via the "External Tools" menu. The following steps are taken to generate CDCB compliant files for upload directly into the CDCB Genetics Evaluations system:

1. "Analysis Results" – navigate to the Axiom results folder to be exported.
2. "Array Name" will be automatically populated.
3. Complete the "Investigator Name", "Project Name" and "Experiment Name" fields. These will be used to populate the Sample export.
4. Generate a "Sample File" template with the button and populate it with your sample information.
5. Set the "Genotypes Export File" and "Samples Export File" values.
6. Click Export to generate the output files.

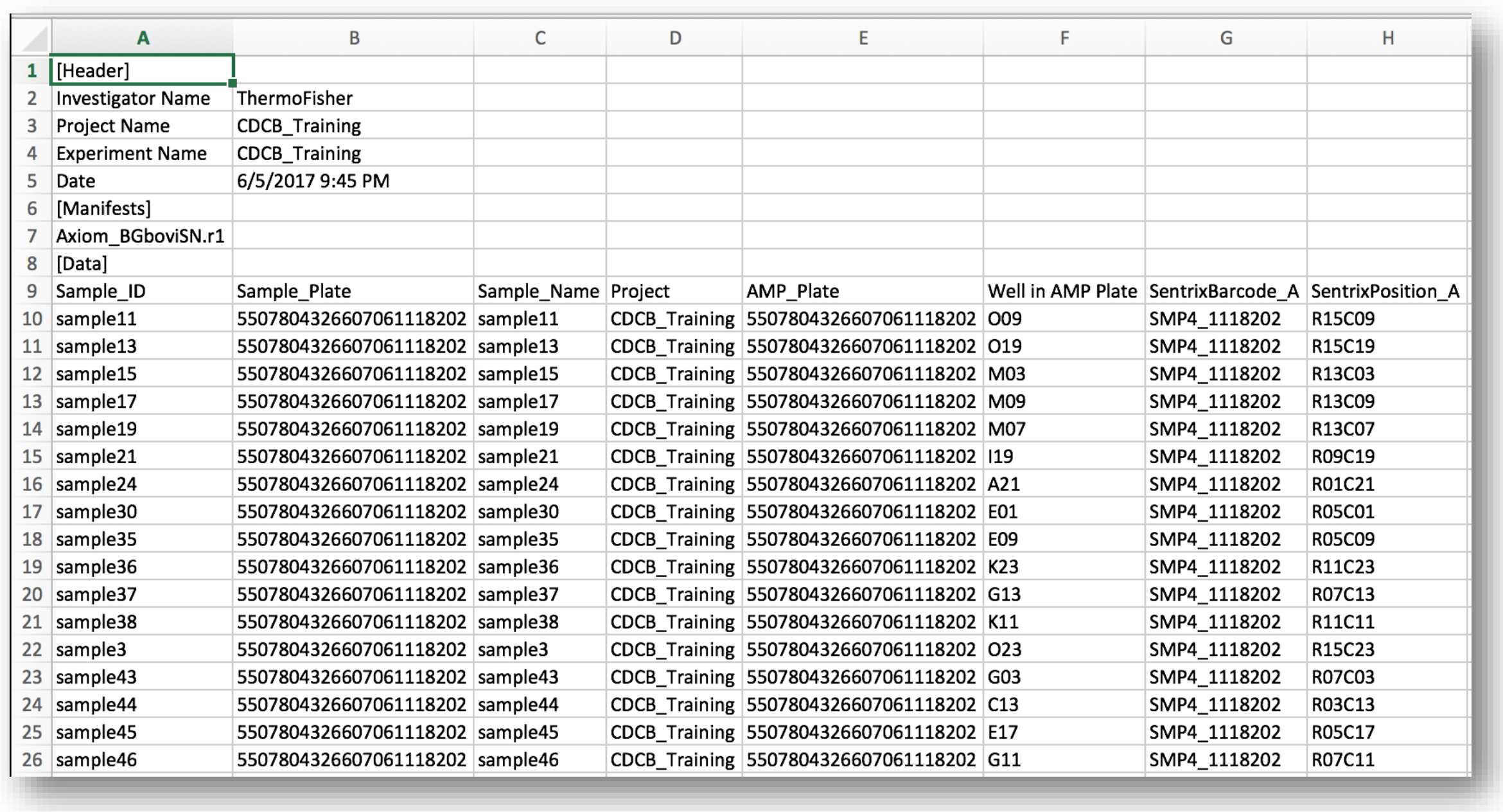
Figure 6. CDCB Export Tool Usage Example Genotyping Output



	A	B	C	D	E	F	G	H	I
1	[Header]								
2	Axiom Analysis Suite Algorithm	1.19.0(2.7.0)							
3	Processing Date	6/5/17 21:45							
4	Content	Axiom_BGboviSN.r1							
5	Num SNPs	46578							
6	Total SNPs	57513							
7	Num Samples	95							
8	Total Samples	95							
9	[Data]								
10		sample10	sample11	sample12	sample13	sample14	sample15	sample16	sample17
11	Affx-93058216	AB	AB	AA	AA	AA	AA	AB	AA
12	Hapmap51730-BTA-44937	AB	BB	BB	BB	BB	BB	BB	BB
13	ARS-BFGL-NGS-11506	AB	AB	AA	AA	AB	AB	BB	AB
14	Hapmap36127-SCAFFOLD59715_196	BB	BB	BB	BB	BB	BB	BB	BB
15	ARS-BFGL-NGS-7215	BB	BB	BB	BB	BB	BB	BB	AB
16	BTB-00430000	AB	BB	AB	AB	BB	BB	BB	BB
17	ARS-BFGL-NGS-110683	AB	AB	BB	BB	AA	BB	BB	AB
18	BTA-78695-no-rs	--	--	--	--	--	--	--	--
19	ARS-BFGL-NGS-91754	BB	AB	AB	AB	AB	AA	AB	AB
20	BTA-112164-no-rs	AB	AB	BB	BB	AA	BB	AB	BB
21	Hapmap50247-BTA-117369	AA	AA	AA	AA	AA	AA	AA	BB
22	BTB-01925882	AA	AA	AB	AA	AA	AB	AA	AA
23	ARS-BFGL-NGS-70160	BB	AB	AB	AA	AB	BB	AB	AB
24	ARS-BFGL-NGS-4015	BB	AB	BB	AB	AB	AB	BB	AB
25	Hapmap48534-BTA-96124	BB	BB	AB	AB	BB	AB	BB	BB
26	BTB-00428180	AB	AB	AA	AB	BB	BB	AB	AB

The table in *Figure 6* demonstrates the genotyping output of the CDCB Export tool. Meta data describing the processing, array, total SNPs on the array, reported SNPs and samples resides at the top, followed by a table of A/B genotyping calls in TOP format. SNP identifiers, where already present in the CDCB database prior to a new Axiom array being added are reported with the original SNP name. Where an Axiom array contains novel SNPs the SNP identifier native to that array is used. A mapping file between native SNP ID and CDCB SNP ID is provided with the array library files.

Figure 7. CDCB Export Tool Usage Example Sample Output



	A	B	C	D	E	F	G	H
1	[Header]							
2	Investigator Name	ThermoFisher						
3	Project Name	CDCB_Training						
4	Experiment Name	CDCB_Training						
5	Date	6/5/2017 9:45 PM						
6	[Manifests]							
7	Axiom_BGboviSN.r1							
8	[Data]							
9	Sample_ID	Sample_Plate	Sample_Name	Project	AMP_Plate	Well in AMP Plate	SentrixBarcode_A	SentrixPosition_A
10	sample11	5507804326607061118202	sample11	CDCB_Training	5507804326607061118202	O09	SMP4_1118202	R15C09
11	sample13	5507804326607061118202	sample13	CDCB_Training	5507804326607061118202	O19	SMP4_1118202	R15C19
12	sample15	5507804326607061118202	sample15	CDCB_Training	5507804326607061118202	M03	SMP4_1118202	R13C03
13	sample17	5507804326607061118202	sample17	CDCB_Training	5507804326607061118202	M09	SMP4_1118202	R13C09
14	sample19	5507804326607061118202	sample19	CDCB_Training	5507804326607061118202	M07	SMP4_1118202	R13C07
15	sample21	5507804326607061118202	sample21	CDCB_Training	5507804326607061118202	I19	SMP4_1118202	R09C19
16	sample24	5507804326607061118202	sample24	CDCB_Training	5507804326607061118202	A21	SMP4_1118202	R01C21
17	sample30	5507804326607061118202	sample30	CDCB_Training	5507804326607061118202	E01	SMP4_1118202	R05C01
18	sample35	5507804326607061118202	sample35	CDCB_Training	5507804326607061118202	E09	SMP4_1118202	R05C09
19	sample36	5507804326607061118202	sample36	CDCB_Training	5507804326607061118202	K23	SMP4_1118202	R11C23
20	sample37	5507804326607061118202	sample37	CDCB_Training	5507804326607061118202	G13	SMP4_1118202	R07C13
21	sample38	5507804326607061118202	sample38	CDCB_Training	5507804326607061118202	K11	SMP4_1118202	R11C11
22	sample3	5507804326607061118202	sample3	CDCB_Training	5507804326607061118202	O23	SMP4_1118202	R15C23
23	sample43	5507804326607061118202	sample43	CDCB_Training	5507804326607061118202	G03	SMP4_1118202	R07C03
24	sample44	5507804326607061118202	sample44	CDCB_Training	5507804326607061118202	C13	SMP4_1118202	R03C13
25	sample45	5507804326607061118202	sample45	CDCB_Training	5507804326607061118202	E17	SMP4_1118202	R05C17
26	sample46	5507804326607061118202	sample46	CDCB_Training	5507804326607061118202	G11	SMP4_1118202	R07C11

The table in *Figure 7* demonstrates the sample output of the CDCB Export tool.

SOFTWARE REFERENCES

[1] Axiom Long format Export tool (AxLE): *https://bit.ly/2R5JYr0*
[2] Axiom Analysis Suite: *http://bit.ly/2uqHODo*
[3] SNPolisher: *http://bit.ly/2tr8NC4*
[4] Council on Dairy Cattle Breeding export tool: *http://bit.ly/2soJRLq*

ThermoFisher
SCIENTIFIC

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