Axiom™ Genotyping Solution Data Analysis

USER GUIDE

Publication Number MAN0018363

Revision A.0

ThermoFisher SCIENTIFIC

Affymetrix Inc. | 3450 Central Expressway | Santa Clara, CA 95051 | USA
For descriptions of symbols on product labels or product documents, go to **thermofisher.com/symbols-definition**.

The information in this guide is subject to change without notice.

DISCLAIMER: TO THE EXTENT ALLOWED BY LAW, THERMO FISHER SCIENTIFIC INC. AND/OR ITS AFFILIATE(S) WILL NOT BE LIABLE FOR SPECIAL, INCIDENTAL, INDIRECT, PUNITIVE, MULTIPLE, OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH OR ARISING FROM THIS DOCUMENT, INCLUDING YOUR USE OF IT.

Revision history: Pub. No. MAN0018363

Revision	Date	Description
A.0	29 July 2020	Initial release in Thermo Fisher Scientific document control system. Supersedes legacy Affymetrix publication number 702961. Updated to the current document template, with associated updates to trademarks, logos, licensing, and warranty. Updated to include new features added to the software such as multi allelic analysis to align with Axiom Analysis Suite 5.0.1 software.

Important Licensing Information: These products may be covered by one or more Limited Use Label Licenses. By use of these products, you accept the terms and conditions of all applicable Limited Use Label Licenses.

TRADEMARKS: All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified.

Contents

1 Introduction to Axiom™ data analysis

About this guide

This guide provides information and instructions for analyzing Axiom™ genotyping array data. It includes the use of Axiom™ Analysis Suite, Applied Biosystems™ Array Power Tools and SNPolisher™ package to perform quality control analysis (QC) for samples and plates, SNP filtering prior to downstream analysis, and advanced genotyping methods. While this guide contains specific information tailored to analyzing Axiom genotyping array data, most principles can be applied to all Applied Biosystems™ genotyping array data with the QC metrics being array specific (e.g., contrast QC for Genome-Wide SNP 6.0 Arrays vs. dish QC for Axiom™ arrays).

Prerequisites

This guide is for scientists, technicians, and bioinformaticians who need to analyze Axiom™ genotyping array data. This guide uses conventions and terminology that assume a working knowledge of bioinformatics, microarrays, association studies, quality control, and data normalization/analysis.

Support

For help or questions, contact your local Thermo Fisher Scientific Field Application Scientist or **thermofisher.com/support**.

Analysis software

Three analysis software systems are used for Axiom™ analysis and described in this document: (1) Axiom™ Analysis Suite version 1.1.1 and later, (2) and Applied Biosystems™ Array Power Tools (APT) version 1.18 and later, (3) the SNPolisher™ R package version 3.0 and later. The workflow using these software systems is shown in the section “Execute the required steps of the workflow” on page 23.

Axiom™ Analysis Suite is a software package that integrates all of the tools necessary to execute the Best Practices Workflow into one program. The software is designed to allow you to set the desired settings and process through all steps with one click. The application eliminates the need for multiple software packages, making the automated analysis of diploid and allopolyploid genomes seamless while also generating various QC metrics. Axiom™ Analysis Suite is the recommended software system for most Axiom™ users.

APT is a set of cross-platform command line programs that implement algorithms for analyzing and working with Applied Biosystems™ arrays (**https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html**). APT programs are for "expert users" who prefer programs that can be used in scripting environments and are sophisticated enough to handle the complexity of extra features and functionality. For more information on the setup and operation of these tools, see the Axiom™ Analysis Suite software user manual, and the APT Help (**thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html**).

SNPolisher™ functions provide visualization tools and advanced analysis methods. All functions necessary for best practices are incorporated into Axiom™ Analysis Suite and APT. Usage of the SNPolisher™ package requires the user to have some familiarity with the programming language R. The files necessary to install the SNPolisher™ package are available on the Thermo Fisher Scientific website (**http://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-devnet-tools.html**).

Download SNPolisher™ installation files

The SNPolisher™ package is available in the Analysis Tools section of the DevNet Tools page (**thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-devnet-tools.html**). You must be registered with Thermo Fisher Scientific to download the necessary files.

1. If needed, select **Register** from the **Sign In** dropdown list the top of the website to register your email address with Thermo Fisher.
2. On the main page under **Applications & Techniques** → **Life Sciences**, click **Microarray Analysis.**
3. In the left column click **Partners & Programs**, then select **Developers' Network.**
4. On the Developer's Network webpage, click **DevNet Tools** on the left side of the menu.
5. Download the zipped SNPolisher™ folder (SNPolisher_package.zip).

The zipped folder contains the R package file (SNPolisher_XXXX.tar.gz, where XXXX is the release number), the user guide, the quick reference card, the help manual, the license, copyright, readme files, a PDF with colors for use in R, and the example R code with example data for running in R. Note that this zipped folder is the package file for installation in R. Users must unzip the file to extract the SNPolisher™ package folder, which contains the tar.gz package file. For instructions on R basics, installation, and usage of the R functions (including additional functions

not discussed in this document), see the *SNPolisher™ Package User Guide* (Pub. No. MAN0017790).

Analysis library files

Axiom™ Analysis Suite and APT software tools require the files (collectively referred to as "analysis library files") listed in Table 1 to appropriately process and interpret the data. For Axiom™ arrays developed through the Axiom™ custom design program, analysis files are made available from a secure file exchange server to the owner of the array. The analysis files for Axiom™ catalog and expert arrays are available from either the array product page (**www.thermofisher.com**) or through direct download via Axiom™ Analysis Suite.

Table 1 Files used for analysis of Axiom™ genotyping arrays. <axiom_array> will be replaced with the actual name of the array. r<#> will be replaced with the version# of the analysis files. For example, when <axiom_array>= Axiom™_BioBank1 and r<#>= r2 then <axiom_array>_96orMore_Step2.r<#>.apt-genotype.axiom.Axiom™GT1.apt2.xml= Axiom™_BioBank1_96orMore_Step2.r2.apt-genotype-axiom.Axiom™GT1.apt2.xml.

Analysis library files	Axiom™ Analysis Suite	APT
<axiom_array>.analysis_settings	Required	N/A
<axiom_array>.ax_package	Required	N/A
<axiom_array>.r<#>.ps2snp_map.ps	Required	Required
<axiom_array>_96orMore_Step1.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml or <axiom_array>_GenericPriors.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml	Required	Required
<axiom_array>_96orMore_Step2.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml or <axiom_array>GenericPriors.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml	Required	Required
<axiom_array>_LessThan96_Step1.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml or <axiom_array>_SNPSpecificPriors.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml	Required for small sample size	Required for small sample size
<axiom_array>_LessThan96_Step2.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml or <axiom_array>_SNPSpecificPriors.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml	Required for small sample size	Required for small sample size

Table 1 Files used for analysis of Axiom genotyping arrays. <axiom_array> will be replaced with the actual name of the array. r<#> will be replaced with the version# of the analysis files. For example, when <axiom_array>= Axiom_BioBank1 and r<#>= r2 then <axiom_array>_96orMore_Step2.r<#>.apt-genotype.axiom.AxiomGT1.apt2.xml= Axiom_BioBank1_96orMore_Step2.r2.apt-genotype-axiom.AxiomGT1.apt2.xml. *(continued)*

Analysis library files	Axiom™ Analysis Suite	APT
<axiom_array>.r<#>.cdf	Required	Required
<axiom_array>.r<#>.qca	Required	Required
<axiom_array>.r<#>.qcc	Required	Required
<axiom_array>.r<#>.step1.ps	Required	Required
<axiom_array>.r<#>.generic_prior.txt	Required	Required
<axiom_array>.r<#>.AxiomGT1.sketch	• Required for human genomes • Optional for non-human genomes	• Required for human genomes • Optional for non-human genomes
<axiom_array>.r<#>.chrXprobes	• Required for mammalian genomes • N/A for non-mammalian genomes	• Required for mammalian genomes • N/A for non-mammalian genomes
<axiom_array>.r<#>.chrYprobes	• Required for mammalian genomes • N/A for non-mammalian genomes	• Required for mammalian genomes • N/A for non-mammalian genomes
<axiom_array>.r<#>.specialSNPs	• Required for human genomes • Required for non-human genomes if gender calling is executed	• Required for human genomes • Required for non-human genomes if gender calling is executed
<axiom_array>.r<#>.AxiomGT1.models	Required for small sample size	Required for small sample size
<axiom_array>.r<#>.apt-geno-**QC.AxiomQC1.xml**	Required	Optional
<axiom_array>r<#>.step2.ps	Required	Optional
<axiom_array>.apt-probeset-**genotype.AxiomSS1.xml**	• Optional for human genomes • N/A for non-human genomes	• Optional for human genomes • N/A for non-human genomes

Table 1 Files used for analysis of Axiom genotyping arrays. <axiom_array> will be replaced with the actual name of the array. r<#> will be replaced with the version# of the analysis files. For example, when <axiom_array>= Axiom_BioBank1 and r<#>= r2 then <axiom_array>_96orMore_Step2.r<#>.apt-genotype.axiom.AxiomGT1.apt2.xml= Axiom_BioBank1_96orMore_Step2.r2.apt-genotype-axiom.AxiomGT1.apt2.xml. *(continued)*

Analysis library files	Axiom™ Analysis Suite	APT
<axiom_array>r<#>.signatureSNPs.ps	• Required for human genomes • N/A for non-human genomes	• Optional for human genomes • N/A for non-human genomes
<axiom_array>r<#>.psi	Required	N/A

Table 1 lists the names of all analysis files that are used to process Axiom™ genotyping arrays in Axiom™ Analysis Suite or APT. Some arrays have more files than those listed in the table in their library file package. An annotation file is an extra file that is not required for genotyping and is not listed below, but used in Axiom™ Analysis Suite to display SNP annotations in SNP results tables, the cluster graph visualizations, and for some export functionality. Annotation files are available for download through Axiom™ Analysis Suite, the array product page, or the secure file exchange in the same locations as the analysis library files.

Note: Library files for some Axiom™ arrays may use alternative names for the Step1 and Step2 .xml files:

- apt-probeset-genotype.AxiomGT1.apt2.xml
- apt-axiom-genotype.AxiomGT1.apt2.xml

Introduction

The success of a genome-wide association study (GWAS) in finding or confirming the association between an allele and disease and traits in human, plant, and animal genomes is greatly influenced by proper study design and the data analysis workflow, including the use of quality control (QC) checks for genotyping data. Although the number of replicated allele/complex disease associations that are discovered through human GWAS has been steadily increasing, most variants that are detected to date have small effects, and very large sample sizes have been required to identify and validate these findings (Manolio & Collins, 2009; de Bakker, et. al. 2008; Baker, 2010). As a result, even small sources of systematic or random error can cause false positive results or obscure real effects. This reinforces the need for careful attention to study design and data quality (Laurie, et. al. 2010). In addition, most genotyping methods assume three genotype clusters (AA, AB, BB) for two alleles. This assumption does not always hold, especially in plant and animal studies, due to the existence of subpopulation genome structural variation and/or auto-polyploid genomes.

This guide presents the Best Practices Genotyping Analysis Workflow to address these challenges, along with instructions for using Axiom™ software for all Axiom™ Genotyping Arrays (human, plant, and animal). The Axiom™ Genotyping Solution

produces calls for both SNPs and indels (insertions/deletions). For simplicity, in this document, the term SNPs refers to both SNPs and indels. Additional chapters in the document include:

- Chapter 2, “Background” provides information that is needed for understanding the remainder of the document.
- Chapter 3, “Best Practices Genotyping Analysis Workflow” discusses the required eight steps for producing high quality and appropriate genotypes for downstream statistical analysis and guidance on interpreting SNP cluster plots. Instructions for executing the steps and visualizing SNP cluster plots are provided in Chapters 7, 8, and 9.
- Chapter 4, “Additional functions and genotyping methods” discusses methods for changing genotype calls and advanced methods for genotyping more than three genotype clusters.
- Chapter 5, “Additional sample and plate QC” discusses QC considerations for samples, and plates that are in addition to those in the required Best Practices steps (Chapter 3, “Best Practices Genotyping Analysis Workflow”).
- Chapter 6, “SNP QC metrics and classification” describes metrics that are used in the Best Practices workflow (Chapter 3, “Best Practices Genotyping Analysis Workflow”) for SNP classification and additional metrics that are used in the field for SNP QC.
- Chapter 7, “Execute Best Practices steps with Axiom™ Analysis Suite” provides instructions for executing all Best Practices Steps with Axiom™ Analysis Suite. Instructions for visualizing SNP cluster plots with the suite are also provided in this chapter.
- Chapter 8, “Executing Best Practices steps with command line software” provides instructions for executing the Best Practices with APT. Instructions for visualizing SNP cluster plots from the command line are provided in this chapter.

2 Background

Axiom™ array terminology

Marker

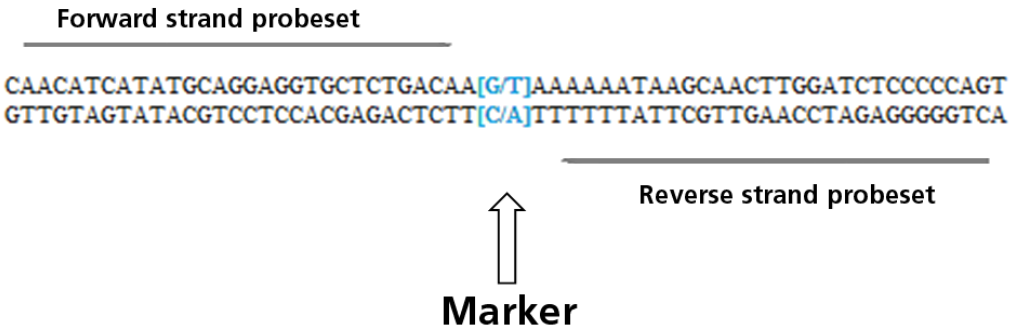


Figure 1 Illustration of a marker interrogated by forward and reverse strand probe sets.

A marker refers to the genetic variation at a specific genomic location in the DNA of a sample that is being assayed by the Axiom™ Genotyping Solution. Both SNPs and indels can be genotyped.

The Applied Biosystems™ unique identifier for a marker is referred to as an affy_snp_id. An affy_snp_id is comprised of the prefix Affx followed by an integer, for example Affx-19965213.

A set of one or more probe sequences whose intensities are combined to interrogate a marker site is referred to as a probeset.

Most Axiom™ markers are interrogated with one or two probesets, one derived from the forward strand sequence and/or one derived from the reverse strand sequence.

The Axiom™ identifier for a probe set is referred to as a probeset_id. A probeset_id is comprised of the prefix AX followed by an integer, for example AX-33782819.

For simplicity, in this document, the term SNP is used to refer to both SNPs and indels. In addition the term SNP is often used to as shorthand for the “probeset used to interrogate the SNP or indel”.

APT file formats

Applied Biosystems™ Applied Biosystems™ Array Power Tools has been updated for use with copy-number and multiallelic SNPs. Due to this, there are more genotyping files with different file formats than in previous versions. The major formatting updates involve call code assignments and multiallelic priors and posteriors.

Dynamic call code assignments

Numeric call codes were previously static assignments, and only a certain set of genotypes were assigned call codes: OTV (–2), NoCall (-1), AA (0), AB (1), and BB (2). With the introduction of copy-number and multiallelic calls, numeric call codes are assigned dynamically during the genotyping process. This means that the number call code 7 may indicate the genotype CD in one calls file and the genotype BF in another calls file. The dynamically assigned call codes are listed in the header comments of each calls file. The previous set of assigned call codes listed above (–2 to 2) have been reserved and will always be assigned to the listed genotype calls. Older calls file will not have the call codes listed, and ps-metrics and ps-classification will use the reserved call codes for these files. ps-metrics and ps-classification cannot process any calls files that have more numeric call codes than the reserved call codes but which do not have the call codes listed in the header. ps-metrics will stop with an error message in this case.

```
#%call-code-1=OTV_1:-4:1
#%call-code-2=NoCall_1:-3:1
#%call-code-3=OTV:-2:2
#%call-code-4=NoCall:-1:2
#%call-code-5=AA:0:2
#%call-code-6=AB:1:2
#%call-code-7=BB:2:2
#%call-code-8=ZeroCN:3:0
#%call-code-9=A:4:1
#%call-code-10=B:5:1
#%call-code-11=C:6:1
#%call-code-12=AC:7:2
#%call-code-13=BC:8:2
#%call-code-14=CC:9:2
#%call-code-15=D:10:1
#%call-code-16=AD:11:2
#%call-code-17=BD:12:2
#%call-code-18=CD:13:2
#%call-code-19=DD:14:2
#%call-code-20=E:15:1
#%call-code-21=AE:16:2
#%call-code-22=BE:17:2
#%call-code-23=CE:18:2
#%call-code-24=DE:19:2
#%call-code-25=EE:20:2
```

Figure 2 The raw numeric call code assignment in the header of a calls file.

Multiallelic SNPs

Multiallelic SNPs are available on some Axiom™ arrays. These are SNPs that have more than one possible genotype base as an alternate allele for a SNP. Multiallelic SNPs can potentially have many more genotypes called than biallelic SNPs. This change in genotyping options affects the calls, summary, posteriors, and priors files. The calls files now have dynamic call code assignments and have as many call codes as there are genotypes across a set of SNPs. The summary files contain one row per allele per probeset. In previous versions, this meant that each probeset had two rows of intensity values in a summary file. Now, a probeset has as many rows as there are alleles genotyped for that probeset. There are still two rows per probeset for a biallelic SNP, and there is a now a variable number of rows for multiallelic SNPs. If a multiallelic SNP has alleles A, B, and C, then there will be three rows in the summary file. If a multiallelic SNP has alleles A, B, C, and D, there will be four rows in the summary file.

One major difference between biallelic SNPs and multiallelic SNPs is that the data transformation for multiallelic SNPs is not the same as for biallelic SNPs: it is base 2 logarithmic transformation for each allele. If a multiallelic SNP has alleles A, B, and C, then the transformation is $\log_2$(A) versus $\log_2$(B) versus $\log_2$(C). For a discussion of plotting multiallelic SNPs, see Chapter 8, “Executing Best Practices steps with command line software”.

The formatting for multiallelic posteriors and priors files is very different from the biallelic posteriors and priors files. Biallelic and multiallelic posteriors and priors data are not combined together in one file like the calls and summary data. There is one row of data per cluster per probeset, not just one row per probeset. If a multiallelic probeset has alleles A, B, and C, then the multiallelic posteriors file will have one row for each of the clusters: AA, AB, BB, AC, BC, CC. Each row has a value for "copynumber" (usually 2) and for "nAlleles" which is the total number of alleles genotyped in the probeset. In this example, there are three alleles so the value of "nAlleles" is 3 for all rows of the probeset.

The means per cluster are the mean value of the cluster for each of the SNP's alleles. The value of "AA-meanB" is the mean location in $\log_2$(B) space of the AA cluster. The covariance values are presented in a similar fashion: the value of "AA-varB" is the variance in $\log_2$(B) space of the AA cluster and "AA-covarAB" is the covariance in $\log_2$(A) and $\log_2$(B) space of the AA cluster. See Figure 3 for a figure of what the means and covariances are for a multiallelic SNP.

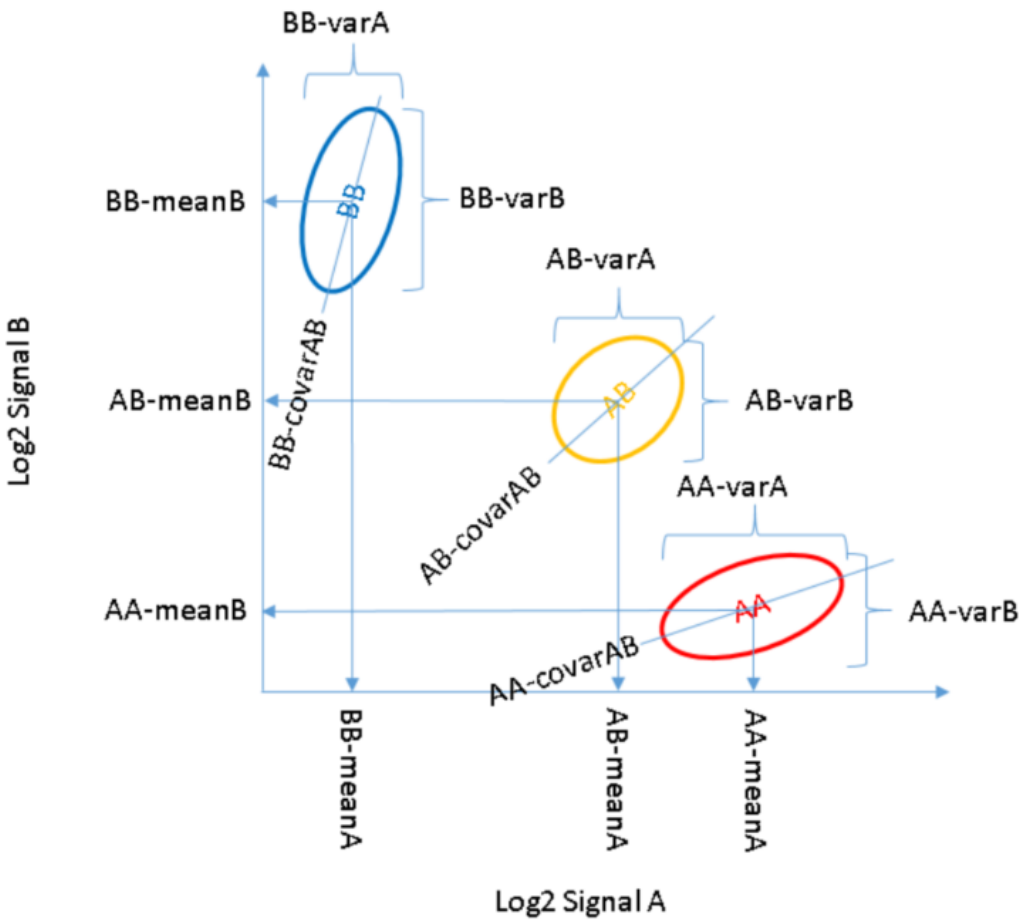


Figure 3 Visual interpretation of multiallelic means and covariances.

The mean and covariance values are formatted as numeric values pasted together with a comma, while each column in the multiallelic posteriors and priors files is separated with tabs. The number of means and covariances per cluster depend on the number of alleles per SNP. For a SNP with 3 alleles, each cluster has 3 means and 6 covariances. A SNP with 4 alleles has 4 means and 10 covariances per cluster. Because the number of means and covariances is different for SNPs with a different number of alleles, the order in which the means and covariances appear is given in the headers. Figure 4 shows the headers with the orders of the means and covariances for SNPs with 3, 4, and 5 alleles. The first SNP in the file is also shown. It has 4 alleles and has 4 means and 10 covariance values per cluster.

#%data-order-mean-nalleles-3=A,B,C
#%data-order-mean-nalleles-4=A,B,C,D
#%data-order-mean-nalleles-5=A,B,C,D,E
#%data-order-covariance-nalleles-3=varA,covAB,covAC,varB,covBC,varC
#%data-order-covariance-nalleles-4=varA,covAB,covAC,covAD,varB,covBC,covBD,varC,covCD,varD
#%data-order-covariance-nalleles-5=varA,covAB,covAC,covAD,covAE,varB,covBC,covBD,covBE,varC,covCD,covCE,varD,covDE,varE

probeset_id	copynumber	nAlleles	cluster	mean	nObsMean	covariance	nObsVar
AX-165679670	2	4	AA	11.971223,9.2	0.2	0.030000,0.000000,0	1
AX-165679670	2	4	AB	11.771223,10.	0.3	0.030000,0.000000,0	1
AX-165679670	2	4	AC	11.771223,9.2	0.3	0.030000,0.000000,0	1
AX-165679670	2	4	AD	11.771223,9.2	0.3	0.030000,0.000000,0	1
AX-165679670	2	4	BB	10.271197,10.	106	0.006743,0.004664,0	106
AX-165679670	2	4	BC	10.271223,10.	0.3	0.030000,0.000000,0	1
AX-165679670	2	4	BD	10.271223,10.	0.3	0.030000,0.000000,0	1
AX-165679670	2	4	CC	10.271223,9.2	0.2	0.030000,0.000000,0	1
AX-165679670	2	4	CD	10.271223,9.2	0.3	0.030000,0.000000,0	1
AX-165679670	2	4	DD	10.271223,9.2	0.2	0.030000,0.000000,0	1

Figure 4 The multiallelic posteriors for one SNP, with the mean and covariance orders in the header of the multiallelic posteriors file.

If the headers of a calls or posteriors file is missing, ps-metrics will use default values. For multiallelic priors or posteriors files, these are the standard values listed in Figure 4. The default order of the means and covariances is alphabetical by allele. If the headers have been removed from a multiallelic posteriors or priors file and the order of the means and covariances is not the default order, then any results from ps-metrics and ps-classification will be incorrect. Users must be careful not to delete the headers of the calls and multiallelic posteriors and priors files.

The multiallelic genotypes are produced from combining the results of multiple biallelic probesets that interrogate all of the possible alleles for a multiallelic SNP. The final Axiom™ genotyping calls, confidences, posteriors, and summary data are given for the final multiallelic data. Only ps-classification uses the biallelic breakdown of multiallelic SNPs when assigning the best and recommended probesets; no other function uses the intermediate biallelic data.

Special SNPs

Non-autosomal SNPs are referred to as special SNPs and are handled differently than the autosomal SNPs. There are seven genomic regions that have special SNPs: the non-psuedoautosomal region of the X chromosome (non-PAR X), the psuedoautosomal region of the X chromosome (PAR), the Y chromosome (Y), the Z chromosome (Z), the W chromosome (W), mitochondrial SNPs (MT), and chloroplast SNPs (CP). These regions can have a variable number of allele copies based on gender, or have fewer than two allele copies for all genders.

“PAR” indicates the pseudo-autosomal region on the X chromosome, where a matching SNP appears on the Y chromosome, and hence these SNPs act like autosomal probesets and not special SNPs with fewer than 2 copies in some or all samples.

Non-PAR X probesets interrogate the non-psuedo-autosomal region of the X chromosome, and do not have a matching SNP on the Y chromosome. Non-PAR X probesets have 2 copies for female samples and 1 copy for male samples. Y probesets have 1 copy for male samples and 0 copies for female samples. Mitochondria (MT) and chloroplast (CP) probesets have 1 copy for all genders. Z and

W probesets appear for avian species and are analogous to X and Y probesets, with the genders reversed. Z probesets have 2 copies for males and 1 copy for females. W probesets have 1 copy for females and 0 copies for males.

The special SNPs file lists all special SNPs for an array and is included as part of the library files for each array. The report file lists the genders for the samples that are genotyped and a new report file is generated when samples are genotyped.

Table 2 Special SNPs with gender and copy number information.

Region	Gender	Copy number	Notes
non-PAR X	Female	2	
non-PAR X	Male	1	
PAR	All	2	Treated as autosomal
Y	Female	0	
Y	Male	1	
Z	Female	1	Avian species only
Z	Male	2	Avian species only
W	Female	1	Avian species only
W	Male	0	Avian species only
MT	All	1	
CP	All	1	

In ps-metrics, samples are separated by gender for non-PAR X, Z, Y, and W probesets when the special SNPs and report files are supplied. MT and CP probesets are treated identically; non-PAR X and Z probesets are treated identically except that the male and female handling is switched; and Y and W probesets are treated identically except that the male and female handling is switched. The different treatment by gender is because of the difference in copy number by gender. The expected behavior of male and female samples is different due to the different number of alleles and ps-metrics and ps-classification take this into account when calculating metrics and assigning categories.

ps2snp files

Ps2snp files are part of the library files that come with each Axiom array. A ps2snp file contains a list of all probesets that are on an array, and the SNP that each probeset interrogates. This allows the user to see which combination of probesets is used to interrogate a SNP. Some SNPs only have one probeset, while other SNPs have multiple probesets on an array. The ps2snp file should have at least two columns, one of which is named “probeset_id” and one of which is named “snpid”.

Multiallelic SNPs have their own identification (multi_snp_id). If an array has multiple biallelic SNPs that are combined to produce the multiallelic genotype calls for a multiallelic SNP, then the ps2snp file will have two more columns: multi_snp_id and ordered_alleles. “multi_snp_id” holds the multiallelic SNP identification, and “ordered_alleles” holds all of the allele options for a multiallelic SNP in order of expected proportion.

Ps-classification uses the ps2np file when selecting the best performing probeset for a biallelic or multiallelic SNP. If the ps2snp file is not provided, ps-classification will not know which combination of probesets interrogate one marker, and a best performing probeset per SNP cannot be selected. If a 2-column ps2snp file is provided when there are multiallelic probesets on an array, then best probeset will be selected for each SNP id rather than for each multiallelic SNP id. Ps-classification will run without a ps2snp file; the best probeset selection step will simply be skipped.

What is a SNP cluster plot for AxiomGT1 genotypes?

A SNP cluster plot is generated for one probeset, which is designed to interrogate a given SNP. Each point corresponds to one sample whose allele array intensities have been transformed into the coordinate space that is used by the AxiomGT1 genotyping cluster algorithm. Functions for creating SNP cluster plots are provided by two Axiom™ software systems: (1) Axiom™ Analysis Suite, via the SNP Cluster Graph function (example shown in Figure 6), and (2) the SNPolisher™ package, via the Ps_Visualization function (example shown in Figure 5). Instructions for the Cluster Graph function usage are provided in Chapter 7, “Execute Best Practices steps with Axiom™ Analysis Suite”, and instructions for Ps_Visualization function usage are provided in Chapter 8, “Executing Best Practices steps with command line software”.

AxiomGT1 is a tuned version of the BRLMM-P (Affymetrix, 2007) clustering algorithm, which adapts pre-positioned genotype cluster locations (priors) to the sample data and calculates three posterior cluster locations. Genotype cluster locations for data with A and B alleles are defined by 2D means and variances for AA, AB, and BB genotype cluster distributions. Priors can be generic, meaning the same prepositioned location is provided for every SNP, or SNP specific, meaning the different pre-positioned locations are provided on a SNP by SNP basis.

Clustering is carried out in two dimensions. The X dimension is called “contrast” and the Y dimension is called “size”. They are log-linear combinations of the two allele signal intensities. For alleles A and B, contrast is log2(A/B) and size is (log2(A) +log2(B))/2. The genotype cluster plots produced by Axiom Analysis Suite and Ps_Visualization label the axes either with the size/contrast pair or with the formulas for size and contrast.

Genotype calls are made by identifying the genotype intensity cluster (AA, AB, or BB) each sample is most likely to belong to. The samples are colored and shaped by these genotype calls in the cluster plots. The AxAS SNP Cluster Graph and Ps_Visualization use the same defaults for genotype call assignments: BB calls are blue upside down triangles, AB calls are gold circles, and AA calls are red triangles. Note that is it possible to color the data according to other sample attributes, with different options in Ps_Visualization and the SNP Cluster Graph.

NoCall assignments are made for samples whose confidence scores are above the confidence score threshold (default = 0.15). The confidence score is essentially 1 minus the posterior probability of the point belonging to the assigned genotype cluster. Confidence scores range between zero and one, and lower confidence scores indicate more confident genotype calls. If the confidence score rises above the threshold, the genotype call for the sample is converted to a NoCall. The AxAS SNP Cluster Graph and Ps_Visualization use gray squares as the default plotting shape and color for NoCalls. The cluster variances are used to create ellipses around the cluster means in the cluster plots. Ellipses based on priors are dashed and ellipses based on posteriors are solid for all biallelic cluster plots.

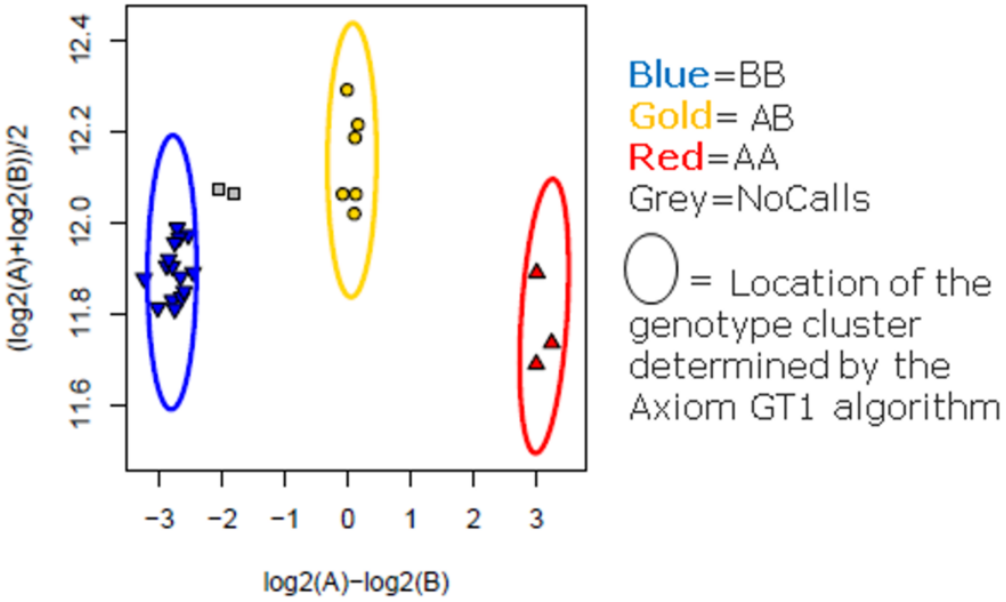


Figure 5 SNP cluster plot produced by the SNPolisher™ package, via the Ps_Visualization function.

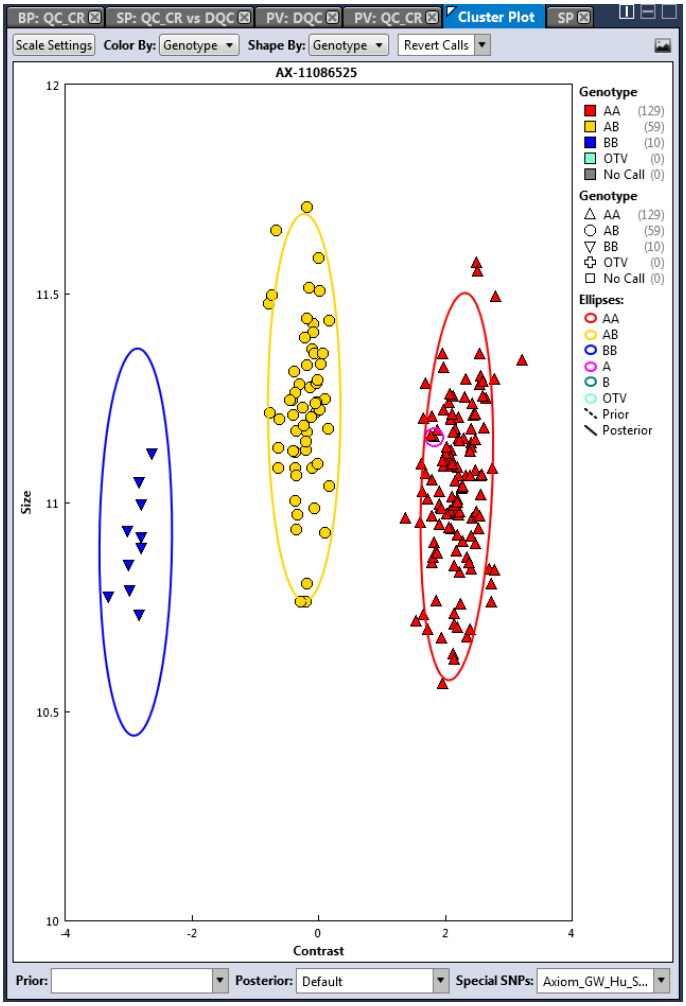


Figure 6 SNP cluster plot produced by Axiom™ Analysis Suite via *SNP Cluster Graph* function.

3 Best Practices Genotyping Analysis Workflow

Design the study to avoid experimental artifacts

Good experimental design practices (Pluzhnikov, et. al. 2008; Cardon & Palmer, 2003; Clayton, et. al. 2005; Zondervan & Cardon, 2007) include randomizing as many processing variables as possible. For a GWAS this means distributing the cases and controls across sample plates, not processing all samples of one type on one day, or having one individual or laboratory process the controls and another process the cases. For larger studies, it is suggested that the experimental design include at least one control sample (of known genotype) on each plate (for example, a HapMap sample) to serve as a processing control. The genotype calls obtained from the control sample can be compared to the expected genotype calls generating a concordance measurement. A low concordance score can indicate that there were either plate processing and/or analysis issues. Before beginning the laboratory work of processing the human samples, investigators should examine the ethnic backgrounds and pedigrees of the proposed samples to ensure that there is no population substructure present that could confound the analysis of data from cases and controls (for example, all of the controls are CEU, although the cases are YRI). For non-human samples the same principles apply, and samples should be randomized with regard to breeds, species, and subpopulations for the genome under study. In addition, researchers should ensure that their experiments are sufficiently powered to answer the question of interest. Again, it is best to examine all of these questions before the initiation of the project.

For a non-ideal study design where the cases and controls are not randomized, the function BatchAlleleConsistency is available in APT and is named ps-bac. BAC can be used to identify and remove SNPs with inconsistent genotypes due to shifts in intensity across samples that were processed in the separate case and control batches. See “Test for batch effects” on page 49 for more details on ps-bac.

Execute the required steps of the workflow

This section describes the eight steps that are required for the Best Practices Analysis Workflow and recommended for all Axiom™ Genotyping Arrays (Figure 7).

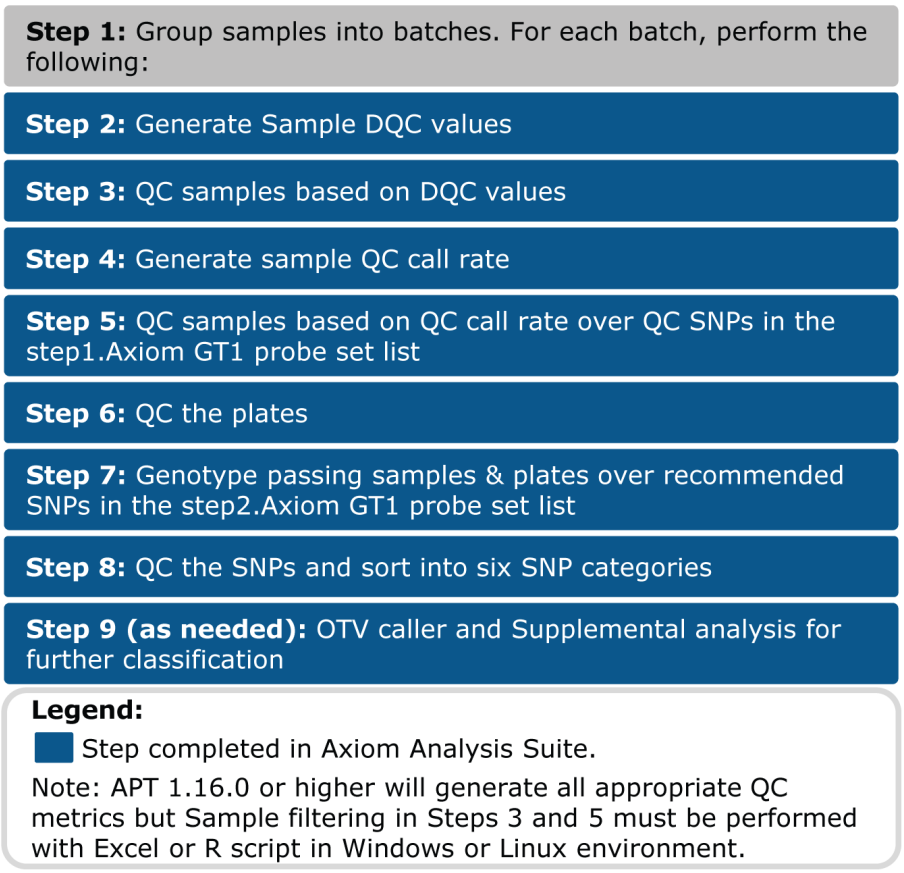


Figure 7 Steps for best practices genotyping analysis workflow.

The actual commands used to execute the steps differ between Axiom™ Analysis Suite and APT. Instructions for using Axiom™ Analysis Suite to execute the Best Practices Workflow are provided in Chapter 7, “Execute Best Practices steps with Axiom™ Analysis Suite”. Axiom™ Analysis Suite is the recommended software for most Axiom™ users. Instructions for using APT to execute Best Practices Workflow are provided in Chapter 8, “Executing Best Practices steps with command line software”.

Step 1: Group sample plates into batches

In general, group plates in as large a batch size as is computationally feasible, or up to 4,800 samples (fifty 96-array plates or twelve 384-array plates). Group plates in the order in which the plates were processed (for example, if using batches of 8 plates, it is preferable to group the first 8, the second 8, etc). The minimum batch size when using generic priors is 96 samples containing at least 90 unique individuals.

SNP-specific priors should be used when the total batch size is between 20 and 96 unique individuals. The specific genotyping option for large (≥96 samples) or small (<96 samples) batch sizes must be selected in all workflows. Each batch should contain either 15 or more distinct female samples or zero female samples. That is, if any female samples are going to be genotyped, at least 15 distinct female samples must be included in the batch.

The exceptions to these batching recommendations are:

- When plates have known significant differences; for example, when they have been processed at greatly different times (many months apart) or in different labs. In these cases, divide the plates into batches according to the date of processing and/or the lab where the samples were run. You can attempt to co-cluster plates with such differences, but plate QC guidelines (“Step 6: QC the plates” on page 26 and “Additional plate QC” on page 61) must be followed carefully.
- DNA samples that are extracted from different tissues or with different techniques should be grouped into separate batches. For example, blood-based, saliva-based, and semen-based samples should be grouped into separate batches.
- DNA that is amplified with an extra WGA step should be grouped into a separate batch.
- Polyploid samples with different genome ploidy levels should be grouped into separate batches. Polyploid samples should not be genotyped together with diploid samples in a single batch. Polyploid samples from different inbred lines may need to grouped into separate batches. Specifically DNAs for different elite inbred wheat lines have been observed to have different polyploid levels at the same genomic site.
- Samples with autopolyploid and allopolyploid genomes should be grouped into separate batches.
- Plant and animal samples from subpopulations that are greatly divergent from each other or from the array reference genome should be segregated and analyzed separately. What includes "greatly divergent" is a gray area and can require several rounds of Best Practices analysis to determine which subpopulations can be optimally batched together in a genotyping cluster run. Methods for genotyping divergent subpopulations require exploration by the user. One approach is to co-cluster the divergent populations and attempt to identify a subset of working SNPs for the population spectrum. Another approach is to co-cluster only samples from the separated divergent population, identify a subpopulation set of working SNPs.

Our guideline for maximum batch size is 4,800 samples or 50 Axiom™ 96-array plates per batch. This is based on internal Thermo Fisher analysis on the effects of batch size on genotyping quality, and achieving reasonable computation performance of the command-line analysis programs (APT and SNPolisher) with the system that will analyze the array plate batches (see Chapter 8, "Executing Best Practices steps with command line software"). As a reference point, a batch size of 55 Axiom™ 96-Array Plates, each with ~650K probe sets, requires about 16 hours to execute step 7 (Figure 7) using the apt-genotype-axiom command ("Best practices step 7: Genotype passing samples and plates using AxiomGT1.Step2" on page 101) on a Linux™ server with the following configuration: x86_64 architecture, 16 x 3 GHz Xeon™ core, and 128 GB of RAM. Note that this is without any computational parallelization.

Step 2: Generate sample DQC values

Before performing genotyping analysis on any samples, the quality of each individual sample should be determined. Steps 2 through 5 collectively identify poor quality samples using first a single-sample metric, Dish QC (DQC), followed by sample QC call rate test.

DQC is based on intensities of probe sequences for non-polymorphic genome locations (that is, sites that do not vary in sequence from one individual to the next). When subject to the two-color Axiom™ assay, probes expected to ligate an A or T base (referred to as AT non-polymorphic probes) produce specific signal in the AT channel and background signal in the GC channel. The converse is true for probes that are expected to ligate a G or C base (referred to as GC non-polymorphic probes). DQC is a measure of the resolution of the distributions of contrast values, where:

$$\text{Contrast} \sim = \frac{AT\ Signal - GC\ Signal}{AT\ Signal + GC\ Signal}$$

Distributions of contrast values are calculated separately for the AT non-polymorphic probes (which should produce positive contrast values) and GC non-polymorphic probes (which should produce negative contrast values). If sample quality is high, then signal will be high in the expected channel and low in background channel, and the two contrast distributions are well-resolved. A DQC value of zero indicates no resolution between the distributions of AT and GC probe contrast values, and the value of 1 indicates perfect resolution.

Step 3: QC the samples based on DQC

Samples with a DQC value less than the default DQC threshold should be excluded from "Step 4: Generate sample QC call rate using step1.AxiomGT1" on page 26. These samples should be either reprocessed in the laboratory or dropped from the study. The default DQC threshold value is 0.82 for most Axiom™ arrays.

Step 4: Generate sample QC call rate using step1.AxiomGT1

Not all problematic samples are detectable by the DQC metric prior to the first round of genotyping (see “Detect mixed (contaminated) DNA samples” on page 58). To achieve the highest genotyping performance, additional poor samples should be filtered post-genotyping so that these samples do not pull down the cluster quality of the other samples. The most basic post-genotyping filter is based on the sample QC call rate.

For this step, samples with passing DQC values are genotyped using a subset of probe sets (usually 20,000) that are autosomal, previously wet-lab tested, working probe sets with two array features per probe set. If no probe sets on the array have been wet-lab tested before array manufacturing (this is the case for many arrays with non-human SNPs), Thermo Fisher Scientific requests the user to provide at least a plate of Axiom™ data to identify probe sets that meet this criteria. Thermo Fisher Scientific will then provide the Axiom™ Analysis Library package (Table 1) for the array. Users should contact their local Field Application Scientist or **thermofisher.com/support** for more information.

Best Practices Step 4 is referred to as *Step1.AxiomGT1* genotyping in the instructions provided for genotyping with Axiom™ Analysis Suite (Chapter 7, “Execute Best Practices steps with Axiom™ Analysis Suite”), and APT (Chapter 8, “Executing Best Practices steps with command line software”). Genotypes produced by this step are only for the purpose of Sample QC and are not intended for downstream analysis.

Step 5: QC the samples based on QC call rate

Samples with a QC call rate value less than the default threshold (97% for most array types) should be excluded from step 7 genotyping. Such samples should be either reprocessed in the laboratory or dropped from the study.

Steps 3 and 5 are the sample QC tests developed for Axiom™ arrays, and are the minimum requirements of the Best Practices workflow. See “Additional sample QC” on page 57 for additional Axiom™ methods and general methods used in the field to detect outlier and problem samples.

Step 6: QC the plates

For Axiom™ genotyping projects, samples are processed together on a 24-, 96-, or 384-array plate. In step 6, basic plate QC metrics are calculated and all samples on plates with non-passing QC metrics should be excluded from the final genotyping run, which will be executed in step 7 of the workflow. The specification for a non-passing plate is when the average QC call rate of passing samples (passing steps 2-5) is less than 98.5%.

The reason for including a plate QC test in the Best Practices workflow is that plates whose sample intensities systematically differ from other plates for some probe sets, can contribute to mis-clustering events (described in “Evaluate SNP cluster plots” on page 40), whether processed separately or processed with all other plates in the batch. These differences can manifest themselves as putative differences in the

minor allele frequency (MAF) of SNPs over these samples relative to the remainder of the study set. If such a plate effect is also combined with a poor study design, where cases or controls are genotyped separately on different plates, this can greatly increase the false-positive rate in the GWA study. Even in a well-designed study, where cases and controls are randomized across plates, inclusion of such outlier plates decrease the power and/or increase false-positive rates.

The metrics and guidelines for plate performance are as follows:

Metrics:

- Plate pass rate

$$\text{Plate pass rate} = \frac{\textit{Samples passing DQC and 97\% QC call rate}}{\textit{Total samples on the plate}} \text{ X } 100$$

- Average QC call rate of passing samples on the plate = MEAN (QC call rates of samples passing DQC and 97% QC call rate thresholds)

Guideline for High-quality Plates

- Plate pass rate ≥95% for samples derived from tissue, blood or cell line, and ≥93% if sample source is saliva
- Average QC call rate of passing samples ≥99%

Guideline for Passing Plates

- Average QC call rate of passing samples ≥98.5%

The minimum guideline for passing plates is an average QC call rate of passing samples that is greater than or equal to 98.5% (gray and green zones in Figure 8). Ideally all plates in the batch will pass the guidelines for high-quality plates (green zone in Figure 8). Passing plates in the gray zone should be reviewed for plate processing problems. If there are no known plate processing problems, you can proceed with caution to include passing samples from such plates. Low sample pass rates can be caused by problematic sample sources for some but not all of the samples. As long as such samples are excluded by steps 2-5, the remaining samples can be included. All samples on non-passing plates (red zone in Figure 8) should be excluded from the Best Practices step 7 genotyping run, and samples on such plates should be reprocessed. The occurrence of non-passing plates should be rare (<5%). If the occurrence is higher, the lab is recommended to review the sample sources and/or plate processing practices with the local Thermo Fisher Scientific Field Application Scientist.

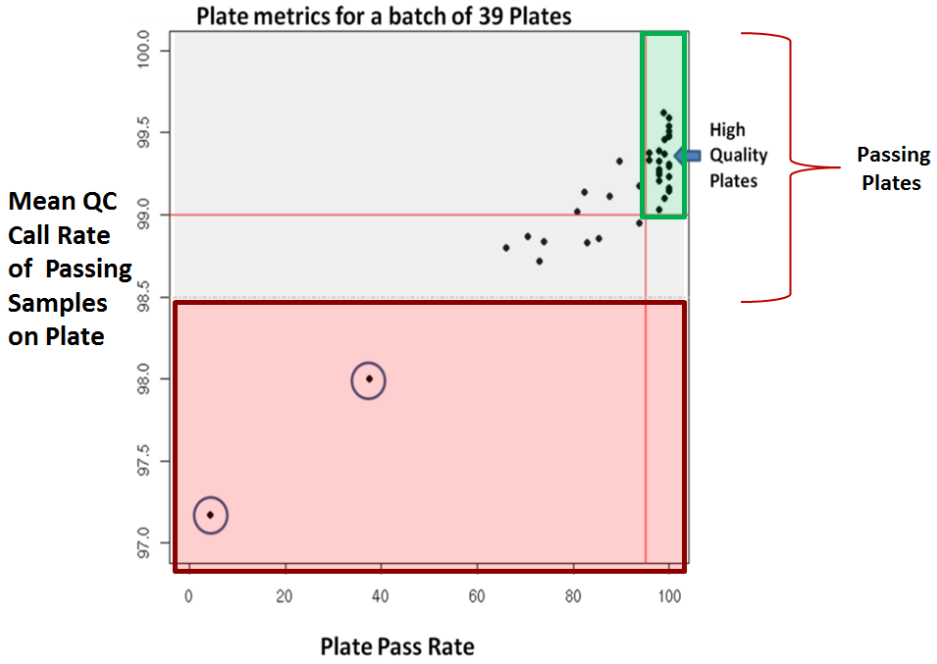


Figure 8 Graph of plate metrics for a batch of 39 plates of blood derived samples. Each plate is shown as a black dot. The graph is divided into three quality zones. The gray and green zones (with Mean QC call rates of passing samples ≥98.5%) are the zones for passing plates. The green zone flags high-quality plates with ≥95% sample pass rate for the plate (vertical red line on the right side of the graph) and the mean sample QC call rate of passing samples >99% per plate (horizontal red line). The gray zone flags marginal plates that should be subject to further review. The red zone flags non-passing plates that should be excluded from step 7 genotyping (enclosed in circles).

This section describes minimum required Plate QC step. See "Additional plate QC" on page 61 for additional Axiom™ specific methods and general methods that are used in the field to detect outlier plates and batches.

Step 7: Genotype passing samples and plates over step2.AxiomGT1 SNPs

For this step, all samples in the batch that passed sample QC and Plate QC (Steps 3, 5 and 6) are co-clustered and genotype calls are produced by the AxiomGT1 algorithm. Best Practices Step 7 is referred to as *Step2.AxiomGT1* genotyping in the instructions provided for genotyping with Axiom™ Analysis Suite (Chapter 7, "Execute Best Practices steps with Axiom™ Analysis Suite") and APT (Chapter 8, "Executing Best Practices steps with command line software").

Depending on the array, Step2.AxiomGT1 genotyping produces calls for all probesets on the array, or only a subset. Probesets excluded by Step2.AxiomGT1 genotyping are usually those probesets with repeatable performance problems and/or genetic complications.

As discussed in "What is a SNP cluster plot for AxiomGT1 genotypes?" on page 19, the AxiomGT1 algorithm can be executed with generic priors or SNP-specific priors. The best practice recommendation is to use SNP-specific priors for small

batches (≤96 samples). Use of generic priors is recommended for large batches (> 96 samples) when the study objective is a GWAS for a diploid genome. Use of generic priors for large batches allows the genotyping algorithm to dynamically adapt to observed cluster locations, and tends to maximize the number of well-clustered SNPs in a given batch. For small sample sets, SNP-specific priors are used to help the genotyping algorithm accurately call genotypes in the absence of observed intensities for the minor allele. All Axiom™ arrays are provided with analysis files (Table 1) for genotyping large batches and some arrays are provided with analysis files for genotyping small batches.

Specific arrays can benefit from usage of SNP-specific priors, even when the sample size is large. These can include arrays for genomes with large SNP-specific variation in cluster locations such as allopolyploid genomes (discussed below), arrays with a large fraction of SNPs that are monomorphic in the population, and arrays whose intended usage is genomic selection. Advanced Biobank pipelines can benefit from using SNP-specific priors as these priors function to anchor the genotype calls to improve the reproducibility of calls in separate batches and increase the number of SNPs recommended across all batches. Creation and testing for the appropriate SNP-specific priors requires study-specific development.

Note: Best Practices Step 4 Sample QC call rates (Step1.AxiomGT1) are often higher than the Sample call rates produced in Best Practices Step 7 (Step2.AxiomGT1). The call rates are higher because only tested, working SNPs are used for Step 4 QC call rates; whereas Step 7 call rates are often calculated over untested probesets with unpredictable performance.

Step 8: Execute SNP QC

The purpose of Step 8 is to identify probe sets that produce well-clustered intensities (see “Evaluate SNP cluster plots” on page 40) and whose genotypes are recommended for statistical tests in the downstream analysis. When more than one probe set has been designed to interrogate a SNP, the “best” probe set is identified. The overall approach is to sort the best probe set per SNP into categories that are based on a set of SNP QC metrics and then create a recommended probe set list for the downstream analysis. The options for categorizing SNPs are based on thresholds for the SNP QC metrics, where some thresholds have been adjusted for specific types of genomes.

Step 8 uses the ps-metrics and ps-classification functions. These functions are available in the APT software and fully integrated into Axiom™ Analysis Suite. Instructions for Axiom™ Analysis Suite usage are available in “Setup step 2, 3, 5, 6 and 8A, B: Set sample ps-metrics, plate ps-metrics, and SNP ps-metrics” on page 82, and for APT usage see Best Practices Step 8A and 8B “Best practices step 8A: Run ps-metrics” on page 104.

Step 8A: Create SNP QC metrics

The ps-metrics function calculates 17 basic SNP QC metrics for each probeset (probeset_id) that was genotyped in Step 7: Call Rate (CR), Fisher's Linear Discriminant (FLD), Homozygous FLD (HomFLD), Heterozygous Strength Offset (HetSO), Homozygous Ratio Offset (HomRO), minor allele count (nMinorAllele), minor allele frequency (MAF), number of clusters (Nclus), number of AA calls (n_AA), number of AB calls (n_AB), number of BB calls (n_BB), number of No Calls (n_NC), a hemizygous indicator (hemizygous), the minimum Mahalanobis distance (MMD), p-values from the genotype frequency test, the Hardy-Weinberg chi-squared test statistic, and the Hardy-Weinberg p-value. Values for five of these metrics (CR, FLD, HetSO, HomRO, and nMinorAllele) form the basis of the SNP classifications (briefly discussed below). Details of the CR, FLD, HetSO, and HomRO SNP QC metrics are described in Chapter 6, "SNP QC metrics and classification". SNP QC metrics are used in the ps-classification step (Step 8B).

Additional SNP QC tests used in the field are discussed in "Additional SNP metrics that can be used for SNP filtering" on page 79.

Ps-metrics workflows

Ps-metrics additionally calculates metrics for multiallelic probesets, metrics for probesets with copy-number aware genotype calls, metrics for the supplemental workflow, and metrics for SNP-specific priors calculations. Ps-metrics has four workflows: base metrics, supplemental metrics, SSP metrics, and multiallelic metrics. The base metrics workflow is the "standard" metrics workflow from previous versions (and includes copy-number aware metrics if needed). The supplemental workflow is the base metrics workflow plus the supplemental metrics. The SSP workflow is the base metrics plus the supplemental metrics plus the metrics needed for calculating SNP-specific priors. The multiallelic workflow is completely separate from all other workflows and is only run when a multiallelic posteriors file is provided. The supplemental workflow is not run on hemizygous SNPs, non-PAR X SNPs, Z SNPs, copy-number aware genotyped SNPs, or multiallelic SNPs.

Ps-metrics calculates additional metrics for genotype calls that include copy number information (whether a call is diploid, haploid, or copy number zero). When a report file and special SNPs file are provided, ps-metrics calculates gender-separated metrics for any probesets on chromosomes that have a different number of alleles by gender.

The number and combination of metrics produced depends on which workflows are run. There is a total of 47 biallelic metrics and 27 multiallelic metrics plus a numeric count metric for each allele. While it is possible to produce all metrics if the supplemental and SSP workflows are run with copy-number aware genotype calls and the multiallelic workflow is run, this situation is somewhat unusual. It is far more likely that a subset of all possible metrics will be produced by ps-metrics. See Chapter 6, "SNP QC metrics and classification" for more details on the different workflows in ps-metrics and for a list of the complete set of QC metrics produced by ps-metrics.

Step 8B: Classify SNPs using QC metrics

The ps-classification function is used to sort each probeset into eight classes based on the SNP QC metrics generated by ps-metrics. The classes are described in Table 3 and a visual representation of each class is displayed in Figure 9. The classifications are based on default QC thresholds that are shown in Table 4 for different genome types. While the user can change the thresholds if desired, we recommended using the default values. Other and OtherMA are not the same categories. Other indicates that there is more than one problematic issue with a SNP. OtherMA indicates that the probeset was used in the best probeset selection process for a multiallelic SNP (multi_snp_id), and there is a problematic issue with that multi_snp_id. Probesets can only be categorized as OtherMA when a 4-column ps2snp file with multi_snp_id's is used for best probeset selection in ps-classification.

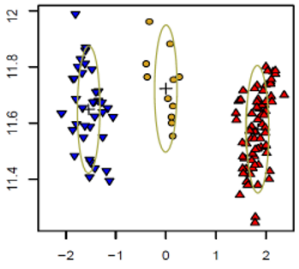
Table 3 Classification categories that probesets are sorted into by ps-classification.

Classification category	Description
PolyHighResolution (PHR)	SNPs with well-separated genotype clusters and two or more alleles in the genotype calls
NoMinorHom (NMH)	SNPs with well-separated genotype clusters; one cluster is homozygous and one is heterozygous for biallelic SNPs, only one homozygous cluster and one or more heterozygous clusters appear for multiallelic SNPs
MonoHighResolution (MHR)	SNPs with one well-formed genotype cluster; must be homozygous
UnexpectedGenotypeFreq	SNPs with a much larger than expected proportion of samples in one or more clusters (arrays with a provided genotype frequency file only)
Off-Target Variant (OTV)	SNPs with a possible OTV cluster (biallelic SNPs only)
CallRateBelowThreshold (CRBT)	SNPs with a very low call rate
Other	SNPs with more than one problematic issue

Table 3 Classification categories that probesets are sorted into by ps-classification. *(continued)*

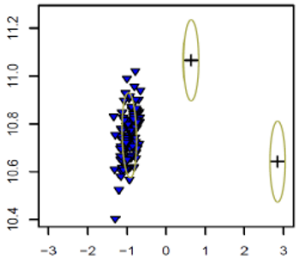
Classification category	Description
OtherMA	SNPs where the original classification was changed due to multiallelic best probeset selection
Hemizygous (without a special SNPs file)	SNPs that are hemizygous and cannot be identified as Y, W, MT, or CP because a special SNPs file was not supplied in ps-metrics

PolyHighResolution



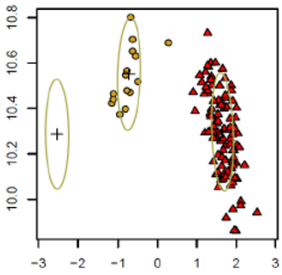
Good cluster resolution and at least two occurrences of the minor allele.

MonoHighResolution



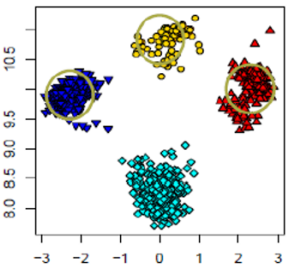
All genotyped samples are monomorphic.

NoMinorHom



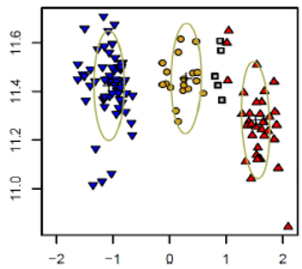
Good cluster resolution with no minor homozygous genotypes.

OTV



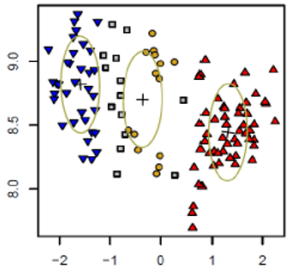
Off-target variants (OTVs) can be assayed by OTV genotyping.

CallRateBelowThreshold



SNP call rate is below the threshold, but other the QC metrics are good.

Other

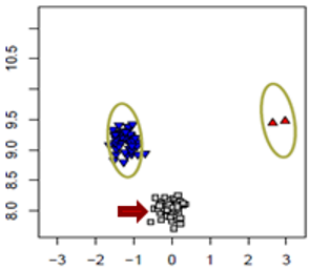


One or more QT metrics are below the thresholds. Expect lower quality genotypes.

Figure 9 Cluster plot examples and descriptions of the base SNP classification categories.

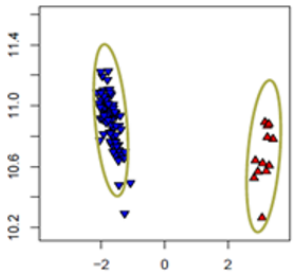
OTV SNPs are discussed further in “Adjust genotype calls for OTV SNPs” on page 45.

Y chromosome



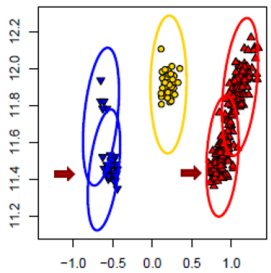
All female samples are set to "No Call" (red arrow).

Mitochondrial and Chloroplast SNPs



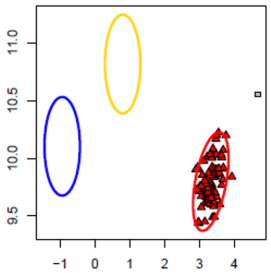
Only two clusters should be present.

Non-PAR X SNP



Male and female samples are visibly separated (red arrows).

Unexpected genotype frequency



All samples are genotyped AA but the AA frequency supplied is 0.01 and the BB frequency is 0.99, indicating that the samples are flipped from the expected distribution.

Table 4 Default QC thresholds for metrics.

Metric values must be greater than or equal to the threshold in order to be considered passing. HetSO.OTV is the HetSO threshold for OTV detection (see “Adjust genotype calls for OTV SNPs” on page 45). HomRO1, HomRO2 and HomRO3 are the HomRO thresholds for SNPs with 1, 2, or 3 genotypes, respectively. nMinorAllele is the threshold for the minimum number of minor alleles in order for a SNP to be classified as PolyHighResolution. Hom-ro indicates if the HomRO thresholds should be used. Polyploid genotypes do not use any of the HomRO thresholds so hom-ro is set to False and cannot be set to True. If genotype frequency p-values were calculated by ps-metrics, then any SNP with one or more p-values that are smaller than genotype-p-value-cutoff is classified as UnexpectedGenotypeFreq.

Metric	Human	Diploid	Polyploid
cr-cutoff	95	97	97
fld-cutoff	3.6	3.6	3.6
het-so-cutoff	–0.1	–0.1	–0.1
het-so-otv-cutoff	–0.3	–0.3	–0.3

Table 4 Default QC thresholds for metrics. *(continued)*

Metric	Human	Diploid	Polyploid
hom-ro-1-cutoff	0.6	0.6	N/A
hom-ro-2-cutoff	0.3	0.3	N/A
hom-ro-3-cutoff	–0.9	–0.9	N/A
num-minor-allele-cutoff	2	2	2
hom-ro	True	True	False
genotype-p-value-cutoff	1e-06	1e-06	1e-06

The best probeset per SNP is determined by the classification priority order: PolyHighResolution, NoMinorHom, MonoHighResolution, OTV, UnexpectedGenotypeFreq, CallRateBelowThreshold, Other, and OtherMA. For a SNP with two probesets, where one probeset is NoMinorHom and one probeset is MonoHighResolution, the probeset that has been classified as NoMinorHom will be selected as the best probeset. The file <axiom_array>.r<#>.ps2snp_map.ps in the Analysis Library File package (Table 1) contains the list of matched probesets and SNPs.

If the special SNPs file and the report file were provided to ps-metrics, ps-classification sorts hemizygous probesets into the eight categories; if these files are missing, ps-classification categorizes all hemizygous probesets as Hemizygous without any further analysis. Visual inspection of hemizygous SNPs is advised (see below). SNPs that are classified as OTV should be analyzed with the otv-caller function. Once the genotype calls have been relabeled, ps-classification can be re-run on the new genotype calls (see “Adjust genotype calls for OTV SNPs” on page 45).

The ps-classification function outputs the *Ps.performance.txt* file, which contains the probeset_id’s, QC metrics, hemizygous status, and an indicator if this probeset is the best for the SNP (BestProbeset), and which category the probeset belongs to (ConversionType) for each probeset. If all SNPs have one probeset, then every probeset is the best probeset by default. Any probeset that is in a recommended category and is the best probeset for a SNP will also be the BestandRecommended probeset for that SNP (see Step 8C for more details on recommended SNPs). If a ps2snp file has been provided, the snp_id column is included. Note that some versions of the performance file may have “affy_snp_id” as a column name instead of “snpid”. Column names and examples are shown in the following table (Table 5).

Table 5 Ps.performance column names and examples.

Column name	Example
probeset_id	AX-11481545
snpid/affy_snp_id	Affx-27771153
CR	99.2320
FLD	8.1936
HomFLD	17.9734
HetSO	0.4500
HomRO	2.5716
nMinorAllele	5123
Nclus	3
n_AA	3112
n_AB	3383
n_BB	870
n_NC	57
Hemizygous	0
ConversionType	PolyHighResolution
BestProbeset	1
BestandRecommended	1

In addition to the performance file, ps-classification outputs eight category files which contain the best probesets per SNP for each category. If a ps2snp file was not supplied, every probeset is the best probeset and therefore every probeset will be output into one of the eight category files. If a ps2snp file was supplied, then only the best probesets will appear in the category files. The eight files are named PolyHighResolution.ps, NoMinorHom.ps, Hemizygous.ps, MonoHighResolution.ps, CallRateBelowThreshold.ps, Other.ps, UnexpectedGenotypeFreq.ps, and OffTargetVariant.ps. Each category file is a one-column, tab-delimited text file with probeset_ids for the category. Each file has a column header called probeset_id. Note that ".ps" extension is a Thermo Fisher Scientific convention to indicate that the file contains a list of probeset IDs.

Ps-classification has four workflows that match the workflows in ps-metrics: base, supplemental, SSP, and multiallelic. The multiallelic workflow is run when there is a multiallelic metrics file, and it can either be run by itself or with one of the other three biallelic workflows. When a biallelic workflow and a multiallelic workflow are run together, the results are combined in the performance file. There is only one

performance file created by ps-classification. See Chapter 6, “SNP QC metrics and classification” for more information on workflows in ps-classification.

Step 8C: Create a recommended SNP list

SNPs that are categorized into recommended categories are of high quality and can be used in downstream analysis. SNPs that are not sorted into recommended classes for the genome type should be excluded from further downstream analysis. Table 6 shows when PolyHighResolution, NoMinorHom, and MonoHighResolution are recommended by genome type. CallRateBelowThreshold, UnexpectedGenoFreq, and Other/OtherMA are not recommended for any genome type. If a special SNPs file and report file were not used when creating the metrics file, hemizygous probesets are classified as Hemizygous. The Hemizygous category is recommended after a visual inspection for Human and Diploid genomes. If a genotype frequency file was not provided, the UnexpectedGenoFreq category will not appear. Probesets that are classified as OTV may re-classified into a recommended category after the otv-caller function is used for genotyping. Note that *recommended* may also be referred to as *converted* in user documents.

Table 6 Recommended SNP categories based on genome type.

Genome type	PolyHigh Resolution	NoMinorHom	MonoHigh Resolution	Hemizygous	OTV
Human	Recommended ✓	Recommended ✓	Recommended ✓	Recommended, visual inspection is advised ✓	Recommended after genotyping with *OTV_Caller* function
Diploid-inbred only	Recommended ✓	Not recommended	Recommended ✓	Recommended, visual inspection is advised ✓	Recommended after genotyping with *OTV_Caller* function
Diploid-outbred or mixture of inbred and outbred	Recommended ✓	Recommended ✓	Recommended ✓	Recommended, visual inspection is advised ✓	Recommended after genotyping with *OTV_Caller* function
Polyploid	Recommended ✓	Requires additional genetic knowledge	Not recommended	N/A	Recommended after genotyping with *OTV_Caller* function
Multiallelic	Recommended ✓	Recommended ✓	Recommended ✓	N/A	N/A

MonoHighResolution SNPs are recommended with caution, especially if the best probeset for the SNP site has never been tested. An extra test for recommending MonoHighResolution SNPs is to require that both probesets (if available on the array) for the SNP site are classified as MonoHighResolution and that the genotypes agree.

Hemizygous SNPs are recommended by default when all hemizygous metrics are missing, but visual inspection is advised (see below). OTV SNPs should be analyzed with the otv-caller function and SNPs may be classified into the recommended categories (see “Adjust genotype calls for OTV SNPs” on page 45). The recalled genotype calls should be visually inspected through plotting.

A total list of unique probesets for recommended SNPs is output by default by ps-classification. This ps file contains all probesets, one per SNP, that are recommended for downstream analysis. If the recommended.ps category file is missing, the list of recommended SNPs can be created manually by combining the category files for the recommended classes. The default value for Human is PolyHighResolution, MonoHighResolution, NoMinorHom (and legacy data includes Hemizygous); for Diploid the default is PolyHighResolution, MonoHighResolution, NoMinorHom; and for Polyploid the default is PolyHighResolution.

Special SNP classification

If the special SNPs and report files were provided for the SNP QC procedure (ps-metrics), then special SNP-specific metrics are calculated and used to classify all probesets that appear in the special SNP file except for PAR probesets on the X chromosome: non-PAR X, Y, MT, CP, Z, and W.

When the special SNPs file is provided, ps-metrics identifies the probesets by their chromosome type and calculates either gender-separated metrics (X, Y, Z, W) or haploid metrics (MT, CP). Ps-classification uses the special metrics to identify if the special SNPs meet the expected patterns for probesets with different ploidy by gender or for haploid-only probesets. The genders of the samples must be known to calculate the gender-separated metrics. If the gender information is not provided in the report file, then ps-metrics does not provide the special metrics but will provide the chromosome if the special SNPs file is provided. ps-classification can use the special SNPs file to categorize MT, CP, Y, and W probesets as Hemizygous. X and Z probesets are treated as though they are standard autosomal probesets.

If both the special SNPs file and the report file are missing, ps-metrics and ps-classification cannot identify the special SNPs. All probesets will be treated as autosomal probesets. In this case, the metrics and classification procedure will not accurately reflect the non-autosomal aspects of the SNPs. We strongly suggest that users provide the special SNPs file and report file when using performing SNP QC.

Visual SNP analysis for hemizygous SNPs

Probesets that are classified as Hemizygous should be inspected visually to determine if there are any performance issues. Hemizygous SNPs and genomes produce only two genotype clusters (A and B). These two clusters should be clearly resolved from one another.

Sub-figures A through F of Figure 10 show the expected pattern of homozygous genotype clusters for mitochondrial SNPs, and sub-figures G through L show the expected pattern of homozygous genotype clusters that are produced by the Y chromosome SNPs for male samples. In Figure 10 sub-figures G-L, a cluster of No Call data are visible in addition to the one or two expected homozygous genotype clusters. This No Call cluster is due to the presence of female samples within the

data set. Since female samples lack a Y chromosome, these samples produce data points with a signal equivalent to background signal that are automatically set to No Call in female samples. Note that W SNPs produce the same patterns as Y SNPs with reversed genders.

For Y, W, and MT SNPs, well resolved clusters are ideal. For Y SNPs, the called male samples should be separated from the NoCall female samples. For W SNPs, the called female samples should be separated from the NoCall male samples. If the male and female samples are combined, then these SNPs should be excluded. Figure 10 shows examples of Y SNPs to include and exclude for different number of calls of the minor allele. Similarly for MT SNPs, the two clusters should be separated. The clusters should not be sitting at 0 in contrast space for Y, W, and MT SNPs. Figure 12 shows examples of MT SNPs to include and exclude for cases with and without two alleles.

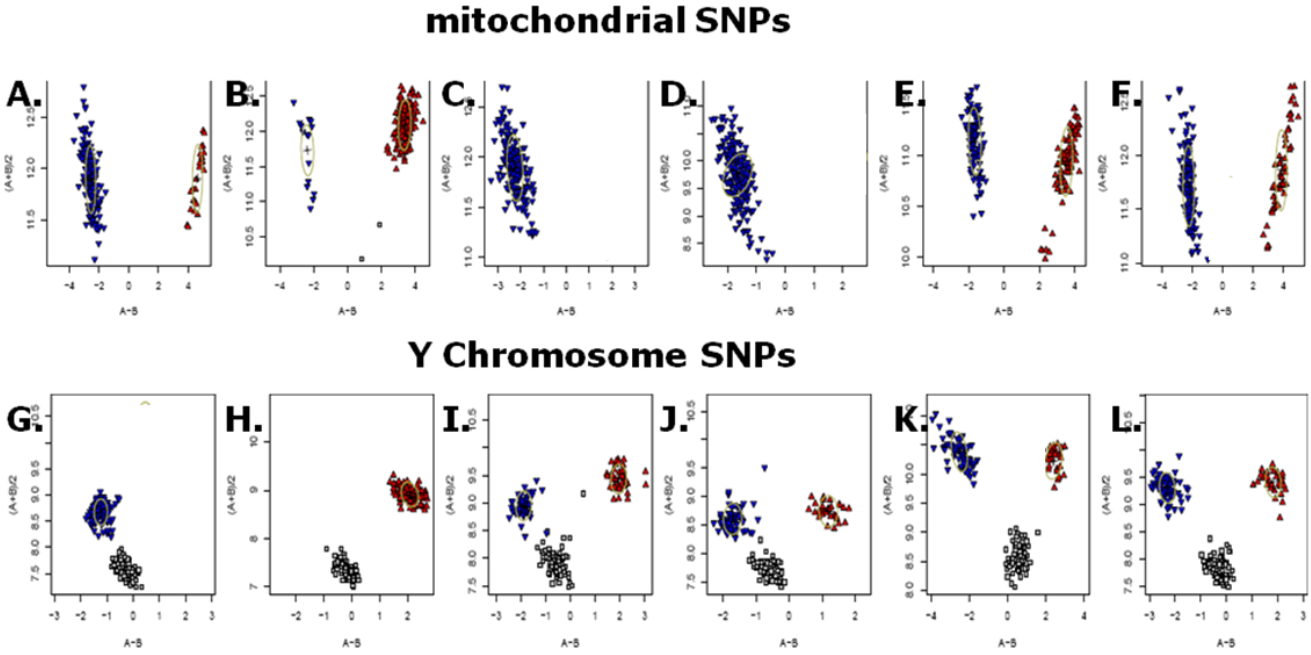


Figure 10 Cluster plots of mitochondrial and Y chromosome SNPs.

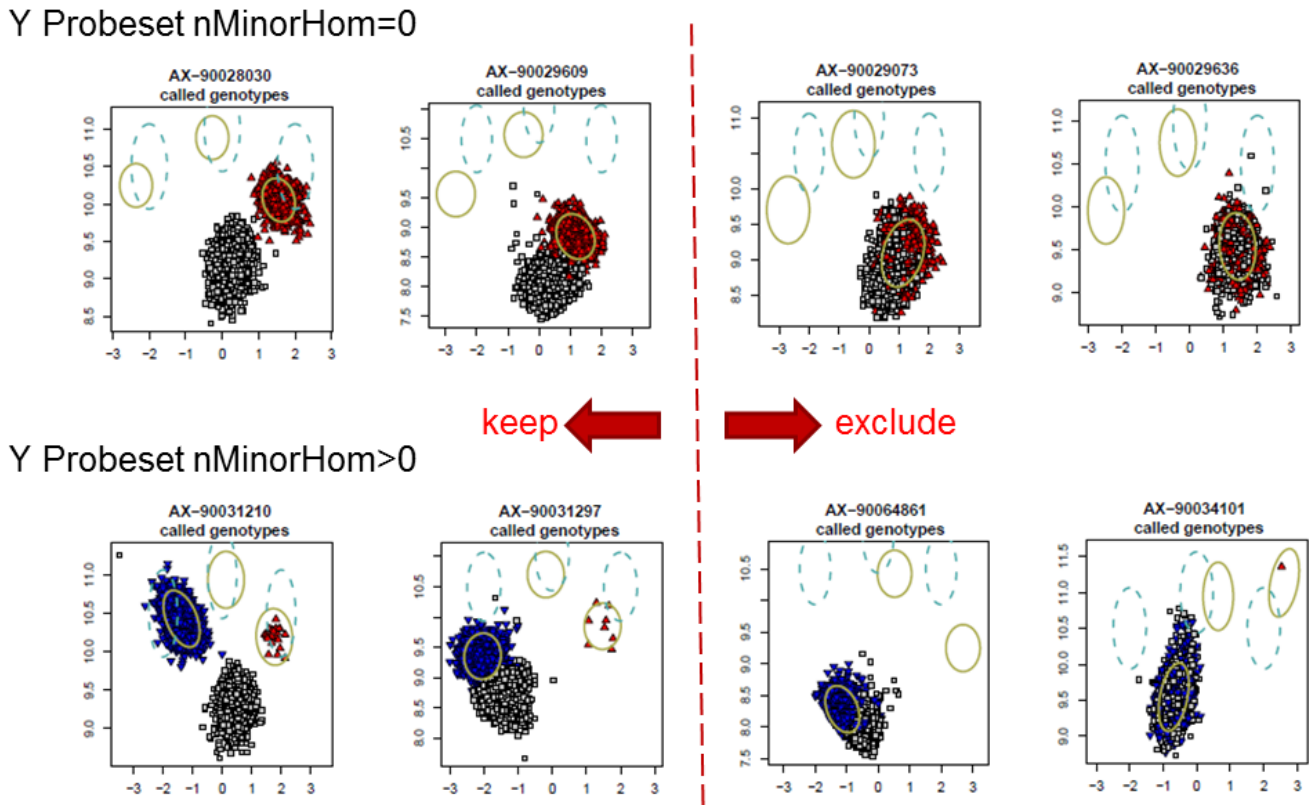


Figure 11 Y SNPs.

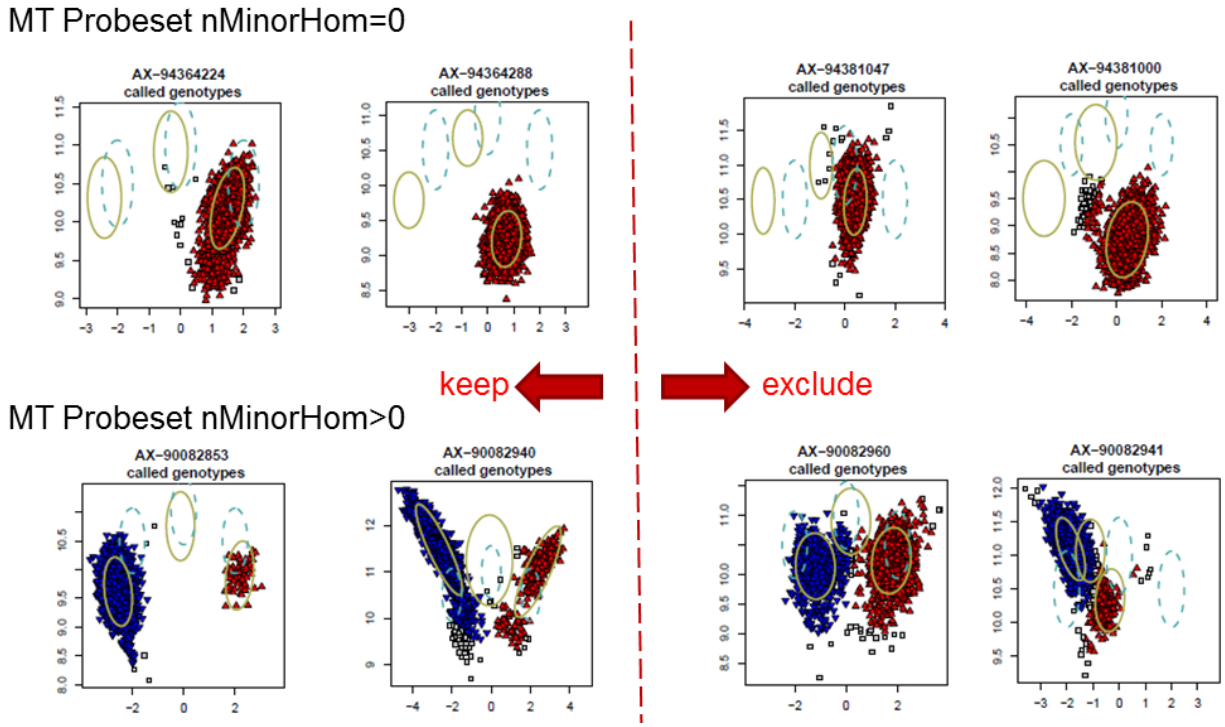


Figure 12 MT SNPs.

Evaluate SNP cluster plots

Visualization and understanding of SNP cluster plots (introduced in "What is a SNP cluster plot for AxiomGT1 genotypes?" on page 19) is a key component the Best Practices workflow. Users should view a small number (~200) of cluster plots of randomly selected SNPs from each of the ps-classification function categories (Figure 9) in order to check that SNPs have the expected cluster plot patterns for the category. SNPs with mis-clustered, multi-clustered, and/or poorly resolved clusters plots should be sorted into CallRateBelowThreshold or Other classes. SNPs in the default recommended categories (Table 6) should have clusters that are reasonably well separated from one another, have no visible batch effects or other cluster anomalies, and should not appear to be of the OTV type. SNPs in the OTV category should have a four cluster OTV pattern.

Functions for creating SNP cluster plots are provided by two Axiom™ software systems: (1) the SNPolisher™ package, via the Ps_Visualization function, and (2) Axiom™ Analysis Suite via the SNP Cluster Graph function. Instructions for the Cluster Graph and Ps_Visualization function usages are provided in Chapter 7, "Execute Best Practices steps with Axiom™ Analysis Suite", and Chapter 8, "Executing Best Practices steps with command line software"; respectively. Cluster plots in this section were produced by the Ps_Visualization function.

Well-clustered vs mis-clustered SNP cluster plot patterns

Figure 13 shows an example of a probe set for a SNP in a diploid genome with well-clustered intensities (left) and an example of a probe set with mis-clustered intensities (right). A well-clustered diploid genome SNP should have one to three approximately elliptical clusters, with the center of each cluster reasonably separated from the centers of the other clusters, and the position of the heterozygous cluster equal to or higher than the position of the homozygous clusters. The mis-clustered SNP example (right) is an example of "cluster split" where the correct BB genotype cluster has been incorrectly split into two clusters (BB and AB), and some of the BB samples are incorrectly called AB (gold). In addition the correct AB cluster has been mislabeled as an incorrect AA cluster (red). The miscalled AB cluster is lower on the Y axis than the BB cluster. This mis-clustering event is easily detected by the SNP QC metrics (CallRate, HetSO and FLD) and should be classified into the Other category. Genotype calls for such SNPs may be manually recalled using the *SNP Cluster Graph* function in Axiom™ Analysis Suite (Chapter 7, "Execute Best Practices steps with Axiom™ Analysis Suite").

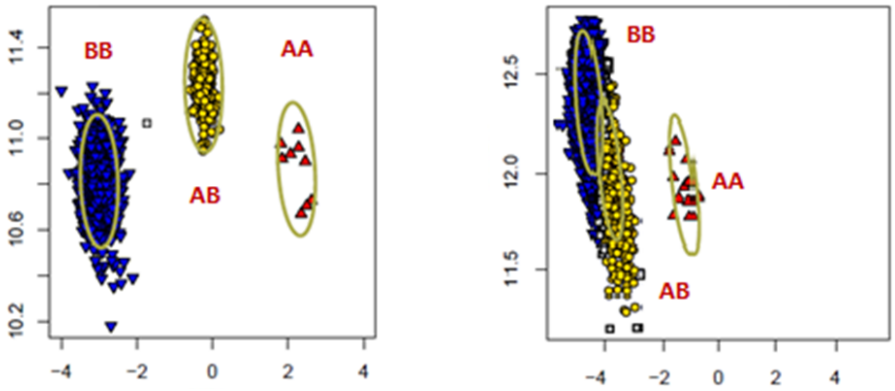


Figure 13 Examples of a well-clustered SNP (left) and a misclustered SNP (right).

Multi-cluster SNP cluster plot patterns

When a subset of samples in the batch co-cluster in their own intensity space, more than three intensity clusters can be produced (Figure 14). This multi-cluster pattern can be due to genuine genetic differences in the clustered samples, especially when genotyping plant and animal genomes; or the pattern can be an artifact due to extreme batch effects. Batch effects variables include sample collection source, plate ID, instrument, operator, sample type, processing date, and more.

Possible genetic differences can be due to inclusion of subpopulations with copy number variations at the given SNP site, or inclusion of subpopulations whose genomes have diverged from the reference population whose genome sequence was used to design the probes for the array. Genomes of divergent subpopulations can have interfering SNPs and indels relative to the array probe sequences that decrease the genotype intensities. OTV SNP sites (Didion, et. al. 2012) are extreme cases where genomes have diverged to the point where only background intensities are produced, and a fourth intensity cluster is formed at the het cluster position. An example is shown in Figure 14. The AxiomGT1 genotyping algorithm assumes a maximum of three genotype clusters for just two alleles and therefore combines additional intensity clusters into three genotype states, resulting in unpredictable mis-calling of the true, complex genetic states.

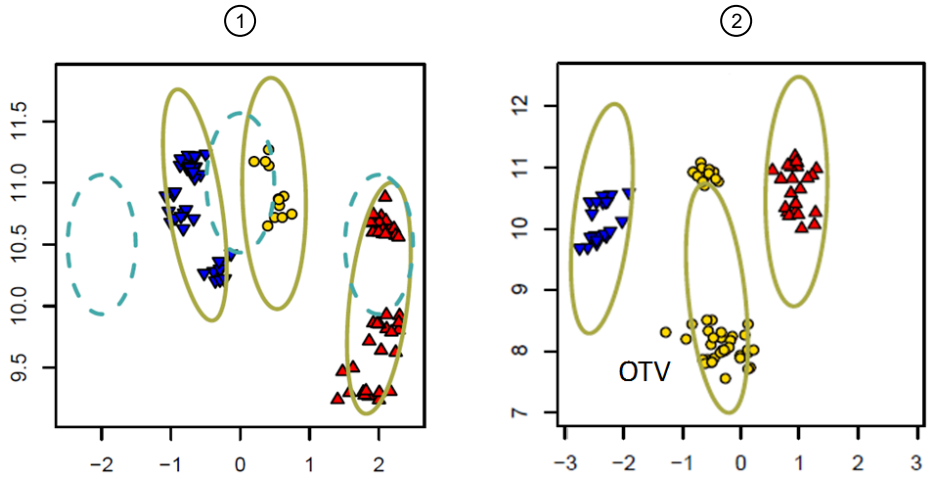


Figure 14 Examples of multi-cluster SNPs.

① Diploid plant SNP with 8 intensity clusters.
② Diploid plant SNP with 5 intensity clusters; one is an OTV.

SNP classification should classify multi-cluster SNPs as Other, CallRateBelowThreshold or OTV. Sometimes, these SNPs have complex patterns that escape the standard SNP QC filters for these classes. If visual examination identifies that multi-cluster SNPs are being included in any of the default recommended classes (Table 6), supplemental filters can be applied. SNPs in the OTV class can be correctly relabeled with four genotype states including OTV (see “Adjust genotype calls for OTV SNPs” on page 45). Both the supplemental options in ps-metrics and ps-classification are available for use in Axiom™ Analysis Suite.

The cluster graphs of the multi-cluster SNPs can be examined for possible causes of extra clusters by coloring samples according to different batch variables and/or known sample subpopulation structure (different breeds, lines, varieties, subspecies, etc). The *by-sample* coloring option is available in Axiom™ Analysis Suite and Ps_Visualization. If samples in outlier intensity clusters can be colored based on a common variable (for example a common Plate ID or a common subspecies) the potential root cause can be identified. It is suggested to repeat Best Practices Step 7 genotyping, excluding the samples that form outlier intensity clusters.

Allopolyploid SNP cluster plot pattern

Allopolyploid genomes contain more than two paired sets of chromosomes, where each set is referred to as a subgenome, and the subgenomes are derived from different species. The alleles of allopolyploid SNP sites usually segregate in just one subgenome, while remaining fixed in the homologous sites in the other subgenomes. Allopolyploid genomes occur in some plant and fish species and produce expected differences in SNP cluster patterns (Figure 15), relative to diploid genomes (Figure 13 left). The intensity contributions of fixed subgenomes do not create additional clusters but they shift and compress the clusters that are formed by the subgenome with the segregating alleles to the right (when A is the fixed allele) or left (when B is the fixed allele). The heterozygous genotype cluster is located between the homozygous genotype cluster along the Y (Size) axis. The AxiomGT1 genotyping algorithm dynamically adapts to the shifted cluster locations and allopolyploid SNPs with the expected pattern are classified as *PolyHighResolution* when the *Polyploid* option is selected in ps-classification (see Chapter 6, "SNP QC metrics and classification" for more information) or in the *Threshold Settings* in the Axiom™ Analysis Suite (see *Axiom™ Analysis Suite User Guide* Pub. No. 703307 for more information)."

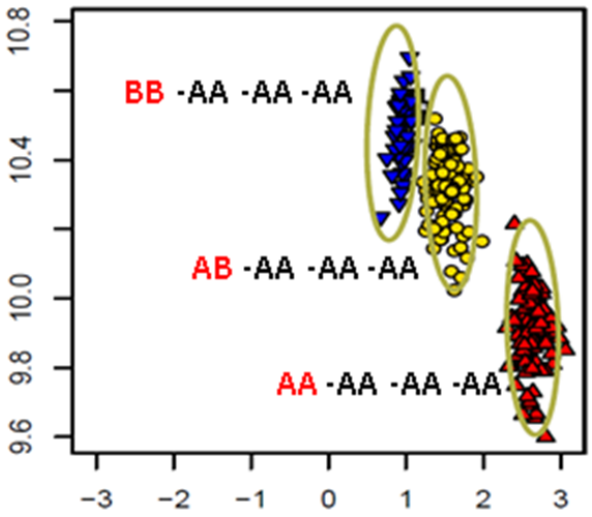


Figure 15 Cluster plot for an allo-octoploid plant. Each sample is colored by the AxiomGT1 genotype call (blue, gold, red) for the subgenome with the segregating allele. Each genotype cluster is labeled by the likely allo-octoploid genotype using the following notation: the genotypes of 4 subgenomes are separated by dashes, the genotype of the subgenome with the segregating allele is noted first (red), followed by the genotypes of the subgenomes whose alleles are fixed (black). It is likely that the genotypes of the fixed subgenomes are AA because clusters are shifted to the right in contrast space, which occurs when the A genotype dosage is higher than the B dosage.

SNP cluster plot patterns for inbred populations

Inbred populations produce few or no heterozygous genotypes and there is often a high frequency of both of the homozygous genotypes (Figure 16). These cases are classified as PolyHighResolution as long as the Polyploid or Diploid option is selected in ps-classification or *Threshold Settings*. However, a SNP with no heterozygous genotypes, such as that shown in Figure 16-3 is classified as Other if the Human option is selected in ps-classification or *Threshold Settings*. AxiomGT1 analysis options should be set to include the inbred penalty when genotyping inbred populations. For additional information on use the inbred penalty, see “Genotyping inbred samples” on page 53 in Chapter 4, “Additional functions and genotyping methods”.

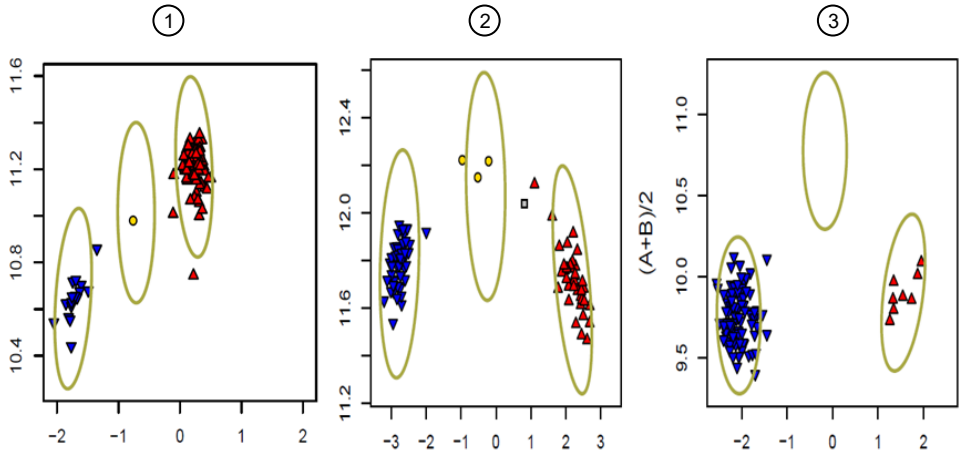


Figure 16 Cluster plots for inbred populations.

1. Allopolyploid plant
2. Diploid plant with heterzygous genotypes
3. Diploid plant without heterozygous genotypes

4 Additional functions and genotyping methods

Manually change genotypes

Sometimes, SNPs called incorrectly due to problematic cluster patterns can be corrected with expert manual intervention. Such cases include SNPs with cluster splits or multi-cluster SNPs, such as OTV cluster patterns that escape the OTV classification. Instructions are provided in “Visualize SNPs and change calls through Axiom™ Analysis Suite cluster graphs” on page 87 for Axiom™ Analysis Suite.

Adjust genotype calls for OTV SNPs

One of the SNP categories produced by ps-classification is OTV. The term "off-target variant" (OTV) refers to SNP sites (Didion, et. al. 2012), whose sequences are significantly different from the sequences of the hybridization probes, for some or all of the samples in the batch. OTV sites have reproducible and previously uncharacterized variation that interferes with genotyping of the targeted SNP. Interference can be caused by double deletions, sequence non-homology, or DNA secondary structures.

OTV SNPs display an OTV cluster with substantially lower hybridization intensities that is centered at zero in the Contrast (X) dimension and which falls below the true AB cluster in the Size (Y) dimension. OTV clusters are often miscalled as AB (Figure 17).

The otv-caller function performs post-processing analysis to identify miscalled AB clusters and identify which samples should be in the OTV cluster and which samples should remain in the AA, AB, or BB clusters. Samples in the OTV cluster are relabeled as OTV (Figure 17).

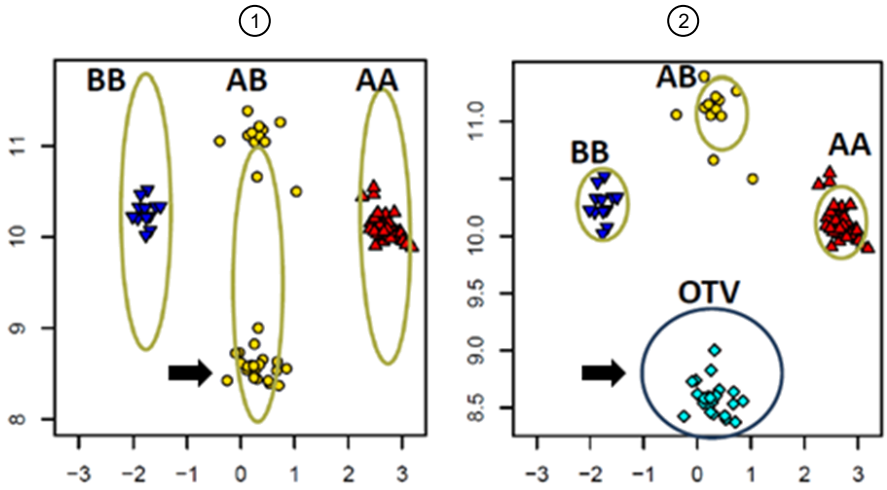


Figure 17 Effect of OTV calling on OTV cluster (arrow) genotypes.

(1) Before OTV genotyping the OTV cluster is mis-called as AB (gold).

(2) After OTV genotyping the OTV cluster has been identified and relabeled as a fourth OTV genotype cluster (cyan)

The intended usage for otv-caller is for SNPs that have been classified into the OTV class by ps-classification ("Step 8B: Classify SNPs using QC metrics" on page 31). otv-caller should be run on the probeset list in the file OTV.ps generated by ps-classification. This ensures that only the best probeset for a SNP has genotype calls changed to OTV if appropriate. If ps-classification was run with the supplemental filters, otv-caller should be run on a new list of probesets made by combining the probesets in the OTV.ps, AAvarY.ps, ABvarY.ps, and BBvarY.ps files.

otv-caller is fully integrated in Axiom™ Analysis Suite; see the *Axiom™ Analysis Suite User Guide* (Pub. No. 703307) for more details on running otv-caller. Instructions for generating SNP cluster plots for the recalled OTV genotypes and plotting the OTV cluster are provided in "Visualize SNP cluster plots with Ps_Visualization" on page 107.

Genotyping autotetraploids

Autopolyploids (occurring in some plant and fish species) are polyploids whose chromosome complement consists of more than two complete copies of the genome of a single ancestral species. SNP sites have a maximum of 6 possible genotypes (AA-AA, AA-AB, AA-BB, BB-AB, BB-BB, AB-AB) and 5 intensity clusters (AA-BB cannot be distinguished from AB-AB). Because AxiomGT1 genotypes a maximum of three genotype clusters, the workflow for assigning genotype calls for autotetraploid genomes is different from the workflow for allopolyploid and diploid genomes.

The R package fitTetra (**cran.r-project.org/web/packages/fitTetra/index.html**) produces genotypes for autotetraploids and is recommended for Axiom™ arrays that are designed to interrogate such genomes. FitTetra was developed by Dr. RE Voorips at Wageningen University's Plant Breeding section. For more information on the fitTetra algorithm, see (Voorrips, Gort & Vosman, 2011).

The SNPolisher™ package provides a workflow to:

1. reformat Axiom™ data for fitTetra input.
2. use the fitTetra R package for assigning genotype calls.
3. reformat fitTetra output to Axiom™ data formatting on the produced calls.

See the *SNPolisher™ Package User Guide* (Pub. No. MAN0017790) for detailed descriptions of the fitTetra input and output functions, as well as more information on the fitTetra package. The *SNPolisher™ Package User Guide* contains a detailed example of how to run the functions to produce Axiom-compatible calls, confidences, and posteriors files for autotetraploid data.

Increase the stringency for making a genotype call

Ps-calladjust is a post-processing function for rewriting less reliable genotype calls to "NoCall" by decreasing Confidence Score thresholds. Confidence Scores are discussed in “What is a SNP cluster plot for AxiomGT1 genotypes?” on page 19. Calls are set to "NoCall" when the Confidence Score is greater than the threshold. Note that decreasing the threshold results in increasing the Confidence Score stringency, which will result in more calls being rewritten as "NoCall" for the samples between clusters. Ps-calladjust can be used on all types of SNPs, including hemizygous SNPs and multiallelic SNPs.

Figure 18 shows the effect for one SNP, displaying how the set of "NoCall" samples between the AA and AB clusters increases as the Confidence Score value decreases. The user can specify which probsets should be checked by providing a probset list. See Chapter 8, “Executing Best Practices steps with command line software” for more information on ps-calladjust.

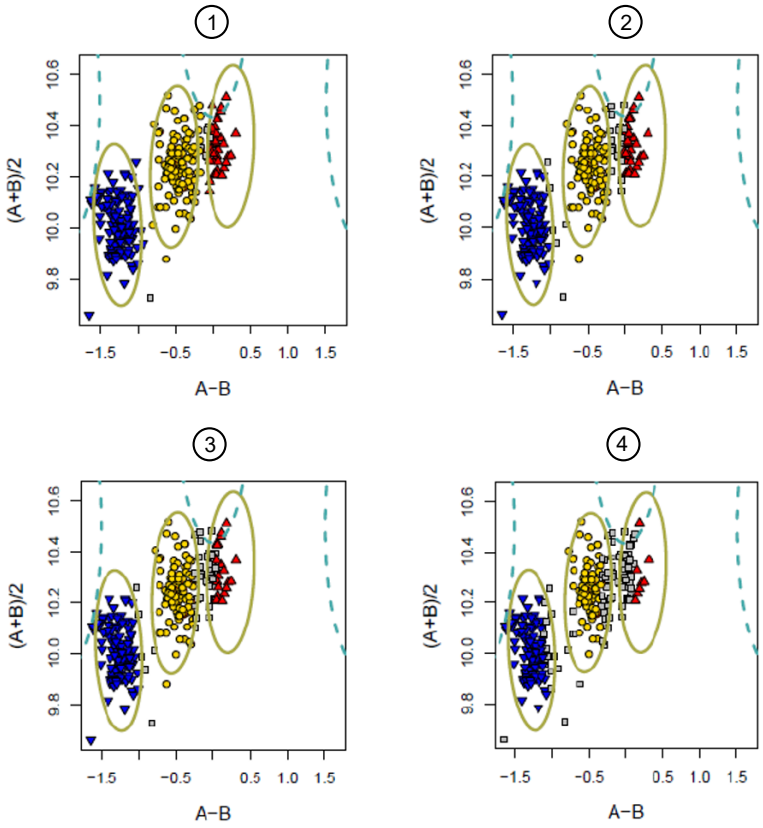


Figure 18 Cluster plots for one SNP, where genotypes are called with lower values of the Confidence Score (CS).

1. CS = 0.15 (default)
2. CS = 0.05
3. CS = 0.01
4. CS = 0.001

Genotyping probesets in variable copy-number regions

The Axiom™ Precision Medicine Diversity Research Array (PMD Array) and some derivative designs contains probesets that are in variable copy number regions. These probesets use the Copy Number Aware Genotyping (CNAG) algorithm for their correct genotyping. The CNAG algorithm uses copy number calls to guide genotyping and reports double deletion, haploid and diploid calls.

The CNAG algorithm requires two additional steps as part of the Best Practices Workflow, which should be performed between steps 6 and 7. In the first extra step, intensity summary signals are produced for all probesets. In the second extra step, Fixed-Region copy number analysis is performed on the summary signals. When this analysis is completed, the Best Practices Workflow continues with Step2 genotyping.

See Chapter 6, “SNP QC metrics and classification” for more information on how to run the CNAG algorithm in Axiom™ Analysis Suite, and Chapter 7, “Execute Best Practices steps with Axiom™ Analysis Suite” for more information on how to run the CNAG algorithm in APT.

Test for batch effects

Ps-bac is a post-processing function to tests for batch allele consistency and batch effects when samples were genotyped in batches (e.g., cases and controls were genotyped separately and together). Batch effects can cause shifts in intensity space, resulting in incorrect cluster fusion when different batch genotype calls are combined. Ps-bac tests for a difference in the B-allele freqency in samples that were genotyped separately and then in a larger group with other samples. If the processed samples are then genotyped together, these shifts may cause confusion about which cluster a sample is assigned.

Case-control studies are more likely to be affected by this type of shift. If the cases and controls are processed separately instead of being randomized across plates and batches, intensity shifts may create a bias between the case and control genotype calls. Control samples can also be used to determine if there are intensity shifts across batches by analyzing the B-allele frequencies of the control samples that are in different batches.

Figure 19 shows the effects of this shift when combining batches: the AA and AB clusters are shifted to the left in the cases (top left figure) compared to the AA and AB cluster for the controls (middle left figure), so the AA and AB clusters from the cases are incorrectly grouped with only the AB cluster from the controls, producing an enlarged AB cluster in the case-control combination (bottom left figure). In comparison, the right figure shows how the case-control combination is correctly genotyped when there is no shift of the AA and AB clusters.

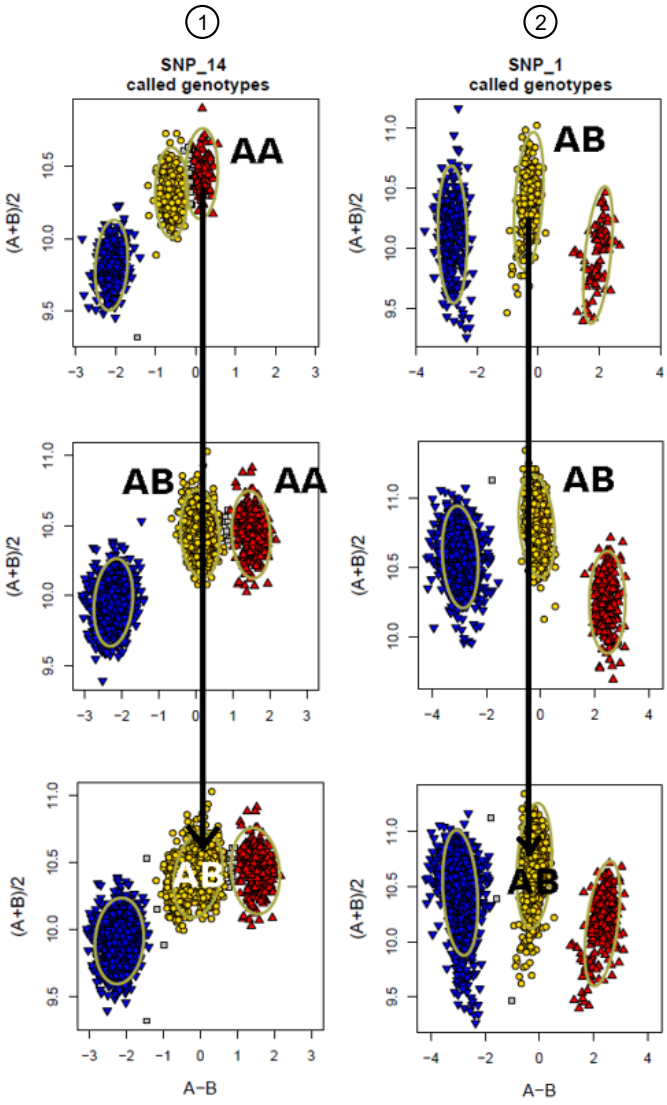


Figure 19

① Left: The AA and AB clusters are not in the same positions for the cases (top) and controls (middle), resulting in incorrect genotyping when they are combined (bottom). When the batches with B-allele shifts are combined, the clusters that are shifted to the left are incorrectly combined with only the heterozygous cluster from the other batches.

② Right: The AA and AB clusters are in approximately the same positions in the cases (top) and controls (middle), so that there are no problems when they are combined (bottom).

B-allele frequency (BAF) can be used to investigate if there has been an intensity shift by testing if the B-allele frequency for a batch is very different when a batch is genotyped by itself and when a batch is genotyped with all other batches. BAF is the ratio between the number of B alleles present in the samples to the total number of alleles:

$$BAF = \frac{\#\ \text{B alleles}}{\text{total}\ \#\ \text{of alleles}}$$

In the left-hand plot of Figure 19, the BAF for the cases is 0.55. When the cases and controls are genotyped together and the AA cluster of the cases has been combined with the AB cluster of the controls, the BAF is 0.65.

Genotyping all samples from all batches together should reduce batch effects for most probesets, but this cannot solve the batch combining problem demonstrated in Figure 19. This problem can be detected by using the BAF. When samples in one batch are miscalled and assigned to the wrong cluster, the number of B alleles will increase or decrease compared to the calls when the batch is genotyped by itself. Comparing the BAF between the two situations will indicate if the BAF is grossly different, indicating that one batch is shifted in intensity.

Ps-bac calculates the BAF for each probeset for each of the data sets. The values in the control data set and the cases data set can then be plotted against each other outside of ps-bac. If the BAF is similar, as it should be, most of the data points should fall along the diagonal x=y line. The more a data point deviates from that line, the greater the change in BAF (ΔBAF) between the cases and controls.

ΔBAF can be used to determine if plates should be genotyped separately or together.

If there is no intensity shift, then the BAF for each probeset should be very similar between the first batch and the other batches. We can look at ΔBAF to investigate: plot the BAF's of the smaller data set against the BAF's of the larger set of batches (where all samples are genotyped). This should create a straight line with a slope of 1 and an intercept of 0 (y = x) if there are no intensity shifts. ΔBAF can be observed by plotting the BAF's and observing which probesets have a large change of BAF when all samples are genotyped.

The default ΔBAF test value between the larger batch and smaller batch for any probeset is ≥ 0.02. Figure 20 shows the BAFs for 6 plates genotyped in the larger batch (x axis) versus the BAFs when genotyped by themselves (y axis). The red lines indicate the ΔBAF values that are within 0.02 of the line x=y. Any ΔBAF values outside of the red lines indicate probesets that are failing.

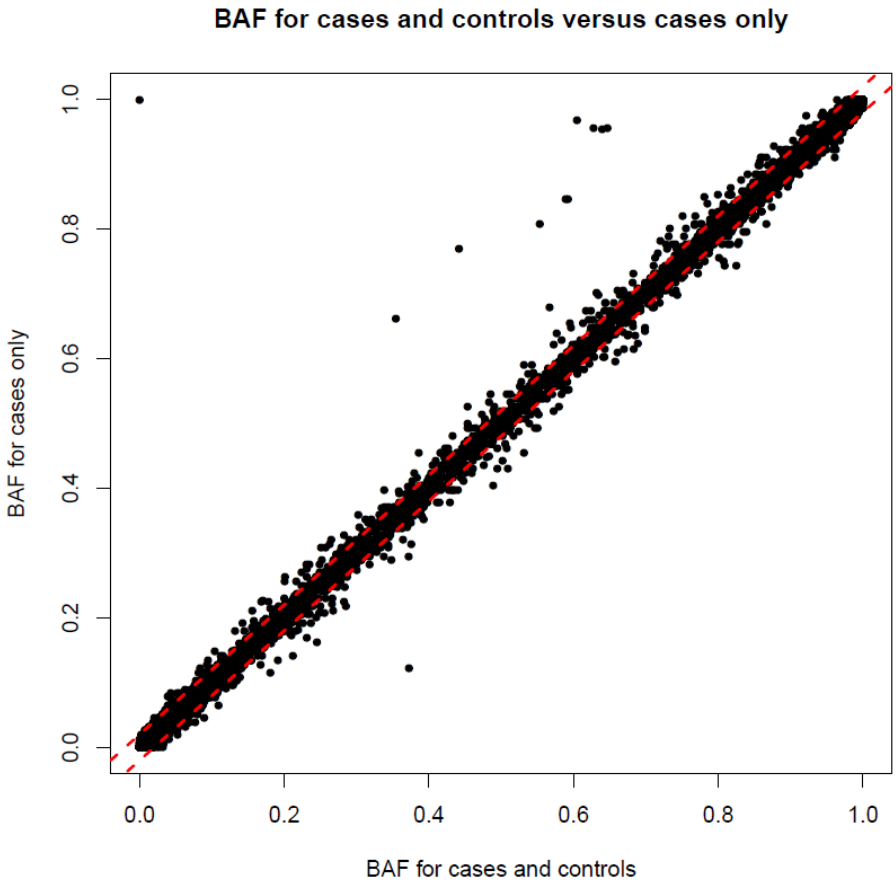


Figure 20 Case-control BAF vs cases BAF. Most probesets' BAF's only change slightly and are within the bounds. x: case-control BAF. y: cases BAF.

After probesets with problematic ΔBAF values have been identified, each batch should be genotyped separately and all QC analysis should be performed on each batch instead of combining batches. For example, if a case-control study has been genotyped on 10 plates, where the cases are all on plate 1 and the controls are on plates 2–10, the BAC test should be run to determine if the samples on plate 1 have been shifted away from plates 2–10 in size versus control space. Due to the AxiomGT1 algorithm genotypes all samples together, a plate that is shifted away from the other plates can possibly end up with genotype assignments based on the positions of the samples on plates 2–10 and not on plate 1.

In this example, plate 1 should be genotyped twice: once by itself, and once grouped together with plates 2–10. Then the BAC test can be used to compare the genotype call assignments on plate 1 when genotyped by itself versus when it was genotyped with plates 2–10 to see if there is a change in the B-allele frequencies of the genotypes (i.e. different genotypes have been assigned to the same samples in the two genotyping runs). If a significant difference is detected and the probesets on plate 1 fail the BAC test, the assigned genotypes for plate 1 from the larger genotyping run should be discarded in favor of the genotypes for plate 1 from the genotyping run with plate 1 only. All further analyses should be performed using the genotypes from the smaller genotyping run for plate 1.

Multiallelic SNPs and SNPs with copy-number aware genotypes should not be used in ps-bac because the formula for calculating BAF assumes that all SNPs are diploid, biallelic SNPs. Ps-bac will output an error if copy-number aware or multiallelic calls appear, but will continue running. Ps_Visualization can be used to plot SNPs and look for potential problems with intensity shifts. Samples from different batches can be highlighted to investigate potential batch effects.

Creating smaller genotyping output files

Ps-extract extracts data from calls, confidences, summary, references, posteriors (biallelic and multiallelic), and priors (biallelic and multiallelic) files for a supplied set of probesets and/or samples. Every input file has a matching output file. Ps-extract is intended as a helper function for other functions that only need some of the probesets or samples from larger files, such as Ps_Visualization. Ps-extract is available in APT and the SNPolisher™ package.

Genotyping inbred samples

The AxiomGT1 algorithm is flexible and with the proper settings, will successfully genotype inbred samples. To do so, an inbred penalty is applied. The inbred penalty biases the genotyping algorithm to call two clusters as homozygous AA and BB, instead of a homozygous and heterozygous cluster. By applying a penalty to heterozygous calls, the inbred penalty increases the accuracy of the genotyping calls on inbred samples.

Identify if an inbred penalty is needed

If the samples are known to be inbred, the inbred penalty should be used. If unsure whether the samples are inbred or not, proceed with genotyping without the inbred penalty and review the cluster plots for the PolyHighResolution SNPs. If a very low number of heterozygous calls are present, then the analysis should be redone using the inbred penalty. Figure 21 shows an example of a data set where the inbred penalty should be applied.

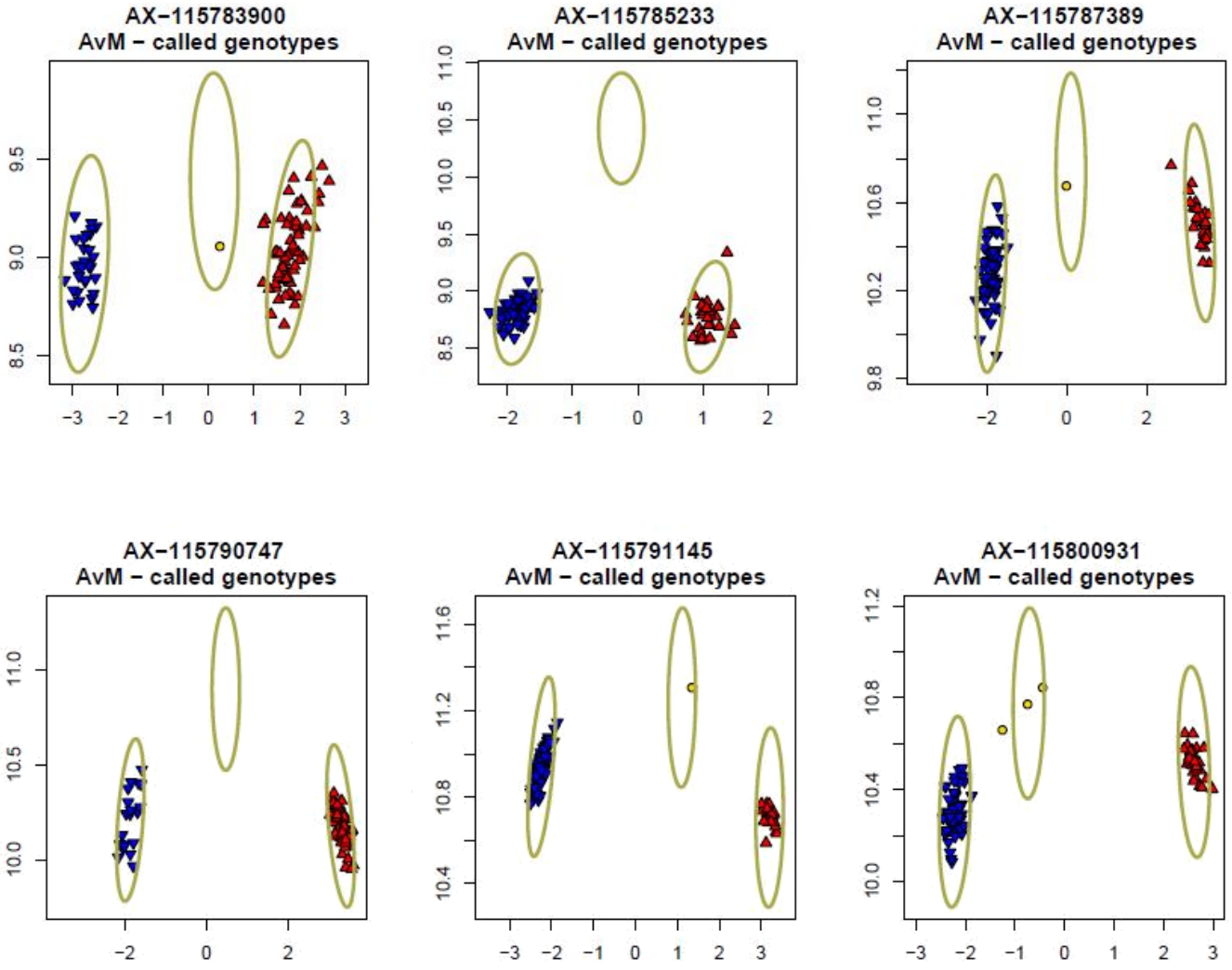


Figure 21 The inbred penalty is recommended for use with samples where a low number of het calls are expected or observed.

How to use the inbred penalty setting

In order to use the inbred penalty setting in any Applied Biosystems™ software, an inbred penalty text file must be created. This is a two column file with the headers of “cel_files” and “inbred_het_penalty”. The rest of the lines should be the CEL file names and a value for the het penalty (see Table 7). The inbred penalty can range from 0 (no penalty) to 16 (max penalty). It is recommended to provide an inbred penalty of 4 to all samples. This is a medium penalty, and it works very well when applied to all samples, including those expected to have some level of heterozygosity. If higher levels of heterozygosity than expected are observed in the resulting data, the penalty value can be increased. Conversely, if lower levels of heterozygosity than expected are observed, the penalty value can be decreased. Alternatively, certain samples can be given higher (or lower) penalty values, based on expected heterozygosity. However, it has been observed that providing a medium penalty to all samples can lead to successful inbred genotyping. Only samples in this file will have the inbred penalty applied to them.

Table 7 Example inbred file with penalty set to 4 for all samples.

cel_files	inbred_het_penalty
GT0011_001_a.CEL	4
GT0011_002_a.CEL	4
GT0011_003_a.CEL	4
GT0011_004_a.CEL	4

Axiom™ Analysis Suite

In Axiom™ Analysis Suite, the inbred penalty file is provided for both Sample QC and Genotyping in the Analysis Settings pane of the New Analysis Tab (Figure 22). Click both "Inbred" radio buttons and provide the inbred file by clicking the "…" button. The suite automatically uses the inbred file that is provided during analysis. See Chapter 7, “Execute Best Practices steps with Axiom™ Analysis Suite” for more information on running best practices with Axiom™ Analysis Suite.

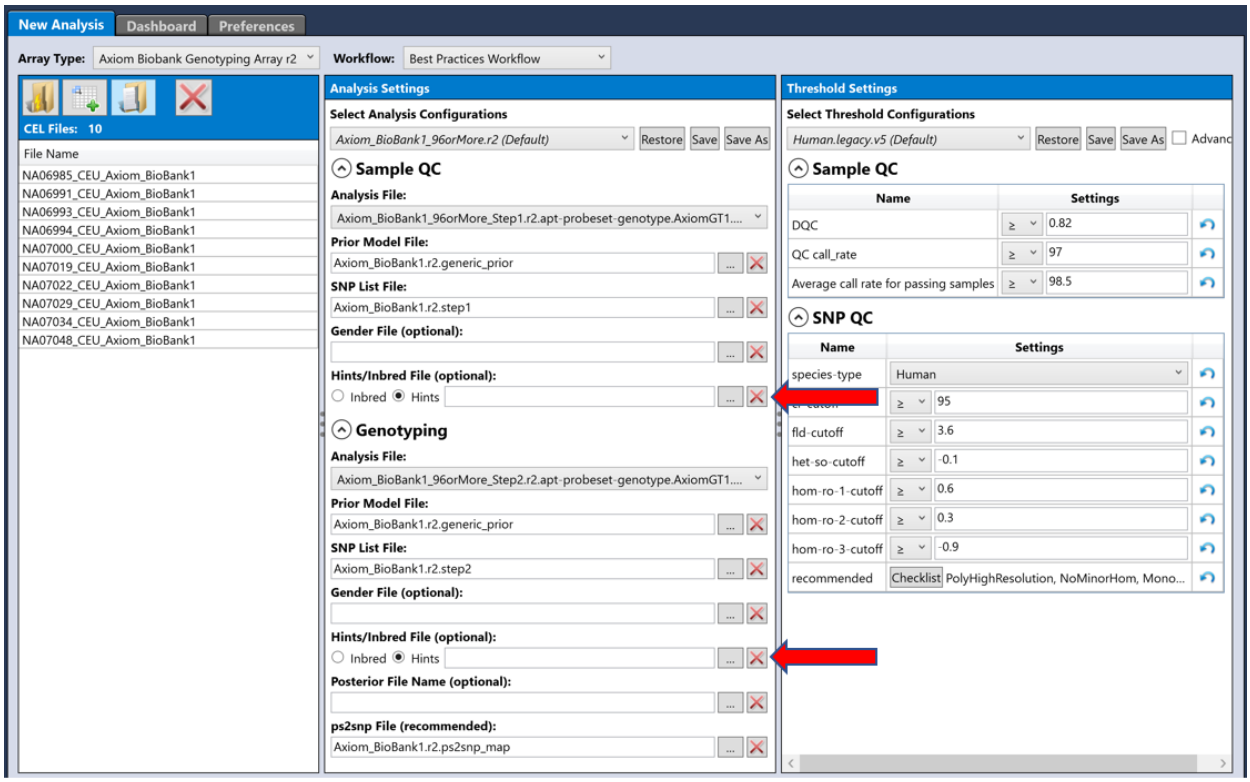


Figure 22 Axiom™ Analysis Suite analysis window. Selecting the inbred penalty file for genotyping inbred samples.

Applied Biosystems™ Array Power Tools

To use the inbred het penalty in Applied Biosystems™ Array Power Tools (APT), pass the --read-inbred <inbred_file> command when running both Step1 and Step2 genotyping. See Chapter 8, “Executing Best Practices steps with command line software” for more information on running best practices with APT.

5 Additional sample and plate QC

Additional sample QC

Detect sample mix-ups

A critical component to a successful GWAS and other studies is that the identities of the samples in the study set are not confused during the sample and array processing. For human samples, Axiom™ arrays contain a set of "Signature SNPs" whose genotypes uniquely identify the individual, and the software conveniently produces a signature SNP report in the pre-genotyping QC process. Thermo Fisher recommends checking that the number of unique signatures in the genotyping samples match the count that is expected in the study set, and that the signatures of expected replicates are the same and are found in the expected plate positions. In addition, a check that the called genders match the expected genders for each sample is recommended.

Unusual or incorrect gender calls

Samples with either unusual or incorrect gender calls (as determined by comparing the reported gender for each sample with the actual gender and/or by comparing the genders of repeated samples) should be carefully examined before they are included in analyses. Methods for checking gender and detecting sex chromosome aneuploidy are presented in *Quality control and quality assurance in genotypic data for genome-wide association studies* (Laurie, et. al. 2010).

Genotyping gender call process: cn-probe-chrXY-ratio_gender

In Axiom™ Analysis Suite, the gender calling algorithm that is used to populate the "Computed Gender" column in the report.txt file is called cn-probe-chrXYratio_gender method. The cn-probe-chrXY-ratio_gender method is more robust when dealing with lower quality samples. Optimal genotyping of sex chromosome SNPs requires use of the correct model type, haploid, or diploid. Haploid models are used for X and Y chromosome SNPs, when the gender call is "male", while diploid models are used for X chromosome SNPs, when the gender call is "female". A "No Call" is made for Y chromosome SNPs when the gender call is female.

The cn-probe-chrXY-ratio_gender method determines gender based on the ratio (cn-probe-chrXYratio_gender_ratio) of the average probe intensity of nonpolymorphic probes on the Y chromosome (cnprobe-chrXY-ratio_gender_meanY) to the average probe intensity of nonpolymorphic probes on the X chromosome (cn-probe-chrXY-ratio_gender_meanX). The probe intensities are raw and untransformed for these calculations, and copy number probes within the pseudoautosomal regions (PAR region) of the X and Y chromosomes are excluded. For human samples, if the ratio is

less than 0.54, the gender call is female, and if it is greater than 1.0, the gender call is male. If the ratio is between these values, the gender call is unknown. For non-human samples, the gender thresholds may vary.

Detect mixed (contaminated) DNA samples

This section discusses patterns produced by mixing of genomes from multiple individuals. The more of these patterns that occur for a sample, the more likely it is that contamination is the causal factor. However, since contamination is not the only cause of these patterns, ultimately the investigator's judgment is required to determine whether these samples should be included in further analyses.

Samples have relatively high DQC and low QC call rate (QCCR) values

In general, higher DQC values correlate with higher sample call rates (see Figure 23); one exception is when samples are contaminated. DQC values are produced by non-polymorphic probes and so are not sensitive to the mixing of DNA from different individuals. However contamination causes QC call rates to decrease. Figure 23 shows the effect of deliberately mixing 4 samples (enclosed in box). Figure 23 includes one plate (green points) where some samples were accidentally contaminated during pipetting. In both plots, the contaminated and deliberately mixed samples fall obviously below the curve that is formed by the uncontaminated samples.

If the analysis of the DQC and QC call rate correlation pattern of a plate reveals a significant number of samples with high DQC values and low sample QC call rates, it may be an indication of sample contamination that is associated with these samples. If the source of sample contamination is understood, it is possible to proceed with the study after eliminating just those samples that obviously fall into the contamination zone.

Note: The contamination produces the pattern in Figure 23, but it has also been observed that large image artifacts on the array surface can produce this pattern as well.

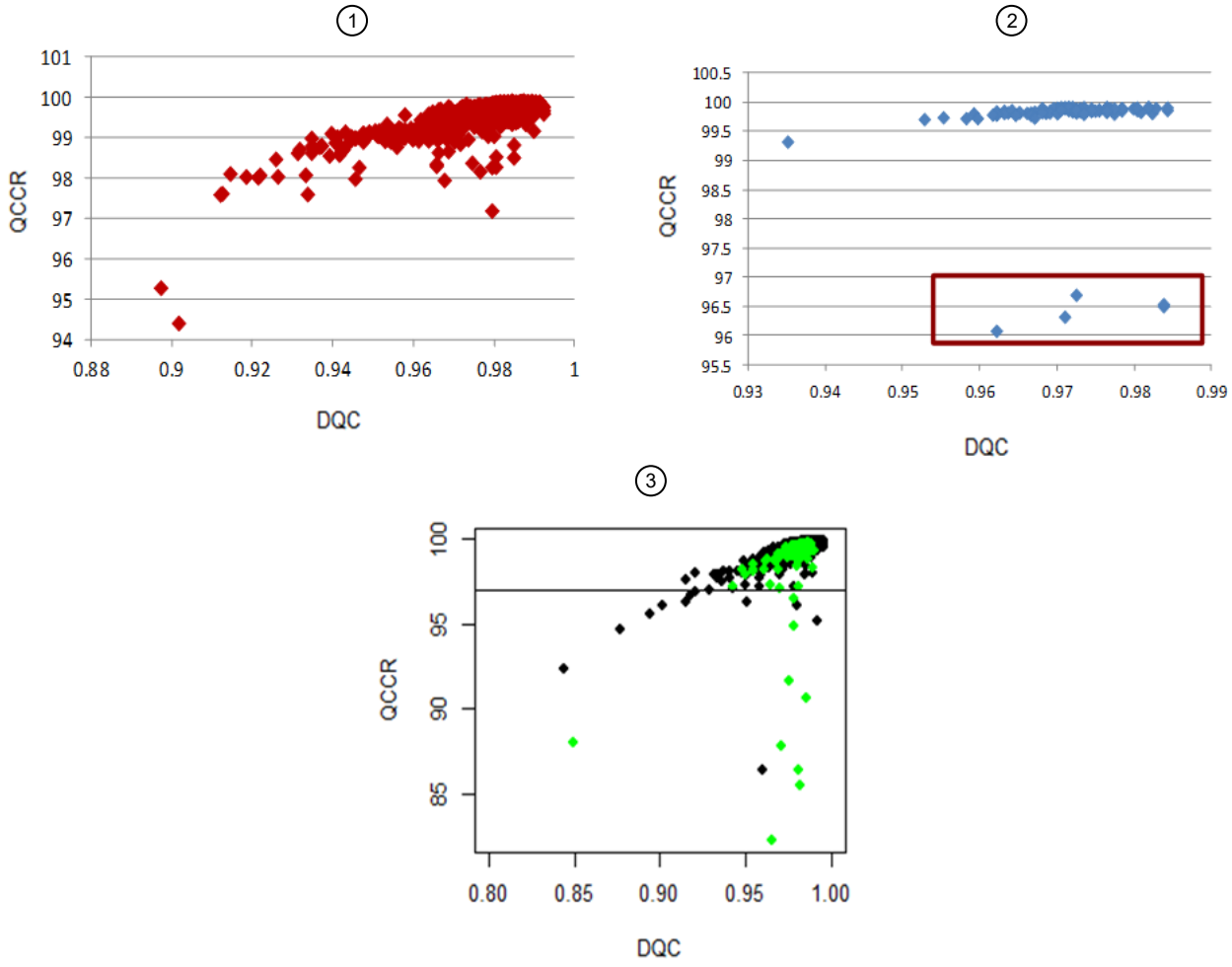


Figure 23 DCQ vs QC call rate (QCCR) plots.

① Representative data set of 10 plates with no obvious contamination problems.
② One plate including 4 samples (enclosed in box) where DNAs were deliberately mixed.
③ Five plates, one plate (green) contains samples that were accidentally contaminated during pipetting.

Samples have a high percentage of unknown gender calls

If male and female DNA are mixed in high enough proportions, the Axiom™ gender calling algorithm will set the call to unknown. Note that individuals with unusual genders (for example, XXY) will also tend to have gender unknown calls.

Samples tend to fall between the genotype clusters formed by the uncontaminated samples

The cluster plots in Figure 24 include deliberately mixed samples (red) and these points fall between the cluster locations for pure BB, AB, and AA genotypes. Both SNP Cluster Graph and Ps_Visualization can color specific samples in a cluster plot.

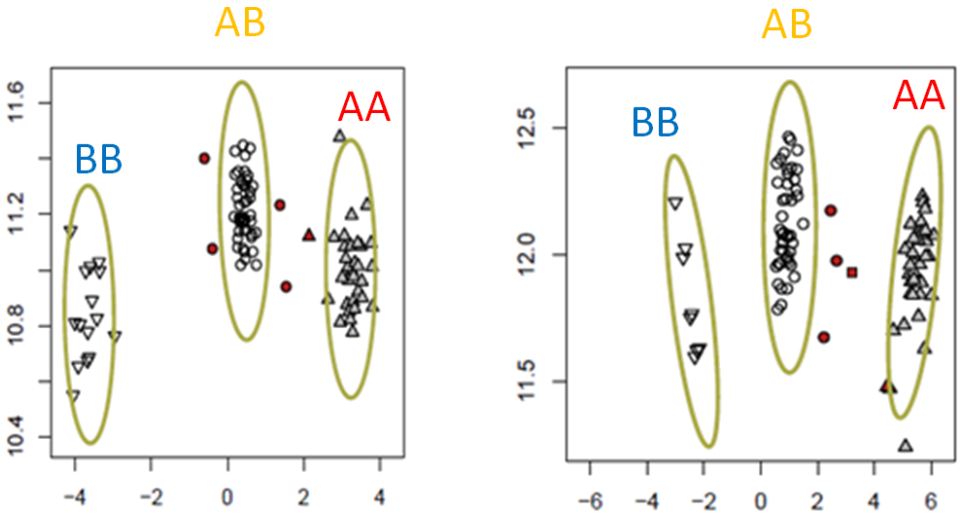


Figure 24 Cluster plots for two SNPs. Deliberately mixed samples are colored red. Uncontaminated samples are colored gray.

Unusual patterns of relatedness

Cross-contamination of samples can cause samples to appear to be related to each other when examining their genotypes. Depending on the extent of the cross-contamination, it can be just a pair of samples or entire sections of the plate that show increased relatedness. Relatedness can be examined using the method that is described in the "Relatedness" section of *Quality control and quality assurance in genotypic data for genome-wide association studies* (Laurie, et. al. 2010).

Increased calculated heterozygosity

Cross-contamination of samples increases the calculated heterozygosity relative to pure samples in the data set, due to mixing of homozygous genotypes with heterozygous or opposite homozygous genotypes. Note that poor quality, pure samples also exhibit increased calculated heterozygosity.

The heterozygosity of a sample is the percentage of non-missing genotype calls that are heterozygous (AB). The Sample Table in Axiom™ Analysis Suite provides this information under the "het_rate" column.

Additional plate QC

This section discusses general methods used in the field to detect outlier plates and batches. It is not feasible to give absolute thresholds on most of these methods for outlier detection, but careful consideration should be applied prior to including samples from flagged outlier plates in further analyses.

Evaluate pre-genotyping performance with DQC box plots

Monitoring DQC plate box plots (Figure 25) is an effective method for early flagging of problematic plates and detecting trends in plate performance, because DQC is a single sample metric that is calculated early and quickly for every sample on every plate (“Step 2: Generate sample DQC values” on page 25).

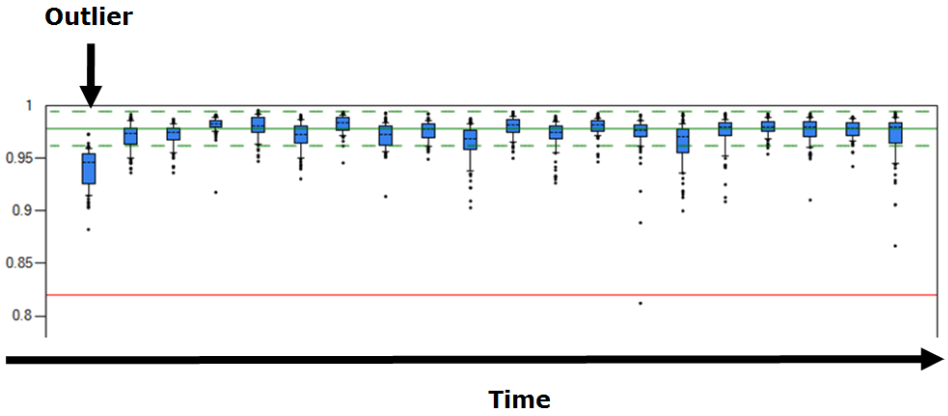


Figure 25 Box plots of DQC values per plate.

The solid green line near the top of the graph represents the median of the medians across all plates. The dashed lines represent ±2 standard deviations from the median of medians. The red line near the bottom of the graph at 0.82 indicates the recommended DQC threshold. In this example, the first plate (identified by the arrow) is an outlier because the upper bound of the box (that is, the 25th% of the DQC mean of this sample plate) is lower than 2 standard deviations below the median of medians.

A suggested approach is for each plate of samples, create a box plot of the DQC values, arrange them in chronological order, and identify the median DQC value for each plate. Next, identify the median of the DQC medians and the standard deviation for each array plate. Finally, identify any plates whose 25th percentile (upper bound of the box) is lower than 2 standard deviations below the median of medians. Such outlier plates should be flagged for further consideration especially if the box plot is visually much lower than the remainder of the plates. We note that being an outlier by this "2 standard deviation definition" does not necessarily mean that the performance is poor. The most important metric for determining which plates should be included in the Best Practices Step 7 cluster set is the average QC call rate of passing samples (“Step 6: QC the plates” on page 26). Axiom™ Analysis Suite contains features to create box plots of any metric (see the *Axiom™ Analysis Suite User Guide* (Pub. No. 703307) for more information).

Monitor plate controls

As part of routine processing for large genotyping studies, it is good practice to include at least one control sample with known genotypes on each plate (for example, a HapMap sample). The calls that are obtained on the plate can be compared to the expected calls (to obtain a measurement of genotyping concordance between the genotypes of the control samples and the genotypes of the known sample) to help indicate whether there were plate processing or analysis problems. A less robust but acceptable indicator of performance is to measure reproducibility by genotyping duplicate samples (the genotypes of which cannot be conclusively known, as they are with HapMap samples) and then comparing the genotype reproducibility measurement between the duplicated samples. In addition, the gender call for each replicate of the sample should be the same. As with the DQC plots, the concordance value of the controls at the plate level should be tracked over time to detect trends and/or outlier plates.

Check for platewise MAF differences

Assuming a randomized study design, the SNP minor allele frequency (MAF) values on a given plate should not systematically differ from the MAF values for the same SNPs on the remainder of the plates. Such a shift in MAFs can reflect mis-clustering events over the samples on such plates. A chi-squared analysis is a simple method for automatically detecting this type of effect (Pluzhnikov, et. al. 2008). A description of this method as described in *Quality control and quality assurance in genotypic data for genome-wide association studies* (Laurie, et. al. 2010) and summarized here.

To detect batch effects on allelic frequencies, we use a homogeneity test as suggested by N. J. Cox (Pluzhnikov, et. al. 2008). The MAF for a SNP on a given plate is compared to the MAF for the same SNP on the remaining plates in the batch.

If there are a total of n samples distributed across the plates, where the ith plate has n_i samples and a sample MAF p_i for a SNP, p_(i) is the sample MAF for that SNP across all plates except the ith plate, and p(bar) is the average MAF for that SNP across all plates, then the chi-squared test statistic with one degree of freedom for the difference between the ith plate's MAF for a SNP and all other plates MAF for that SNP is:

$$n_{(i)}\frac{(p_i - p_{(i)})}{[n\bar{p}(1-\bar{p})]}$$

These statistics are averaged across SNPs to measure how different the plates are from each other. Batches that appear to be outliers must be examined carefully to determine whether their deviation can be accounted for by biological characteristics of the samples, which can be difficult in projects with multiple sources of ethnic variation and/or relatedness among samples.

6 SNP QC metrics and classification

Steps 8A and 8B of the Best Practices Workflow involve producing a set of SNP QC metrics and using them to classify SNPs. While ps-metrics has several options that can produce a large set of metrics and ps-classification has different workflows for performing more complex SNP classification, there are four basic metrics that are used in all classification workflows: SNP call rate (CR), Fisher's linear discriminant (FLD), heterozygous strength offset (HetSO), and homozygote ratio offset (HomRO). CR measures the proportion of samples that received a genotype call, FLD measures cluster quality, HetSO measures the expected position of the heterozygous cluster, and HomRO measures the expected positions of the homozygous clusters.

There are four matching workflows that can be run in ps-metrics and ps-classification: base, supplemental, SSP, and multiallelic. The multiallelic workflow is run whenever there are multiallelic probesets, and is run independently of the other three workflows that are for biallelic probesets. See Appendix A, "Complete Set of SNP QC Metrics Produced by ps-metrics" for more details on the cluster means and variances and the metrics calculated by ps-metrics. See Appendix B, "Complete set of classification thresholds used by ps-classification" for more details on the classification thresholds.

SNP metrics produced by ps-metrics (Step 8A)

SNP call rate (CR)

SNP Call Rate (CR) is the ratio of the number of samples assigned a genotype call (the number of samples that have something other than "NoCall") to the number of total samples over which a genotype call is attempted for the SNP. SNP call rate is a measure of both data completeness and genotype cluster quality (at low values). Very low SNP call rates are due to a failure to resolve genotype clusters (Figure 26). Poor cluster resolution can produce inaccurate genotypes in the samples that are called or a non-random distribution of samples with no-calls and can lead to false-positive associations in a GWA study.

CR is given as a percentage. 100% (no NoCall's) is the maximum value for CR and 0% (all NoCall's) is the minimum value for CR.

SNP Call Rate = # Samples Called/N

Samples Called = the number of samples assigned a genotype call

N = the total number of samples

Although CR is correlated with genotype quality, the performance of marginal SNPs falls along a continuum and there is no perfect threshold for filtering out problematic SNPs from a pool of SNPs providing optimal power for a study. We recommend

setting the filtering thresholds for CR based on the species under study and visually examining the cluster plots for SNPs with CR just above or below the threshold. This examination can result in the inclusion of some SNPs with CR just below the threshold and the removal of some SNPs with CR just above the threshold. See Table 4 for default CR thresholds used in ps-classification.

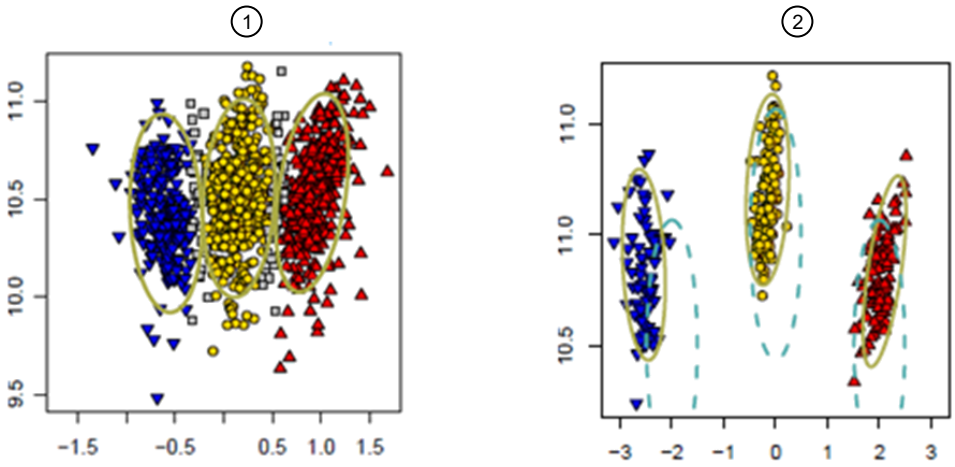


Figure 26 SNPs with different SNP call rates (CR).

1. SNP with low (93.0%) CR.
2. SNP with high (99.4%) CR.

Fisher's linear discriminant (FLD)

Fisher's linear discriminant (FLD) is a measurement of the cluster quality of a SNP. High-quality SNP clusters have well-separated centers, and the clusters are narrow. High-quality clusters can be identified by examining the shape and separation of the SNP posteriors that are produced during genotyping. FLD is essentially the smallest distance between the heterozygous (middle) cluster center and the two homozygous cluster centers in the X dimension. CR and FLD are correlated, but sometimes FLD detects problems that are not captured by CR.

$$\underline{\textit{Fisher's Linear Discriminant (FLD)}} = Min(i = aa, bb)\left\{\frac{|M_{ab} - M_i|}{sd}\right\}$$

where M_{aa}, M_{bb}, and M_{ab} are the centers of the homozygous (aa, bb) and heterozygous (ab) clusters in the contrast dimension (X axis), and sd is the square root of the pooled variance across the three distributions. FLD is undefined (NA) if either the heterozygous or one of the homozygous clusters are empty.

HomFLD is a version of FLD calculated for the homozygous genotype clusters. HomFLD is undefined for SNPs without two homozygous clusters.

Figure 27-1 shows an example of a SNP with low FLD. In this case, the clustering algorithm has found the location of the BB cluster to be too close to the AB cluster producing a FLD of 2.95. Conversely, the well-clustered SNP in Figure 27-2 has a high CR and separated cluster centers, producing a FLD of 7.97.

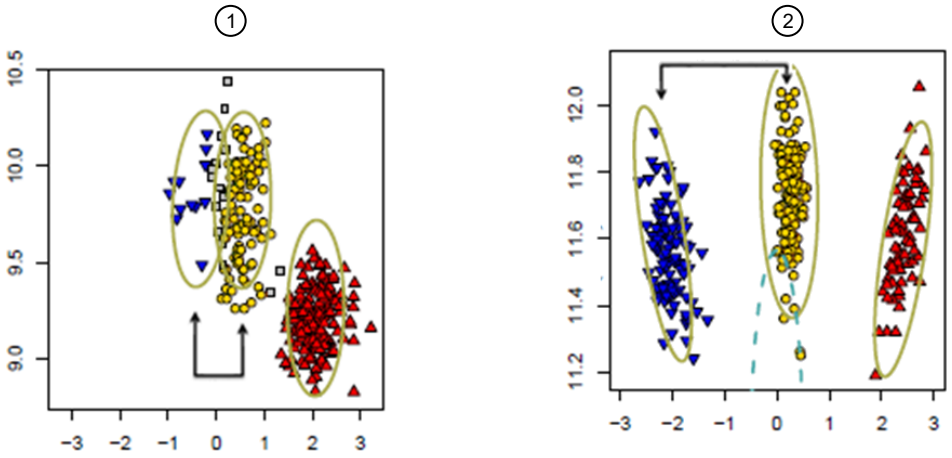


Figure 27 Examples of SNPs with low FLD (A) and high FLD (B).

(1) SNP with low FLD (2.95). FLD detects that the heterozygous cluster center and a homozygous cluster center are too close together in the X dimension.

(2) SNP with high FLD (7.97). The three cluster centers are far apart from each other in the X dimension.

Heterozygous strength offset (HetSO)

Heterozygous strength offset (HetSO) measures how far the heterozygous cluster center sits above or below the homozygous cluster centers in the size dimension (Y axis). Low HetSO values are produced either by misclustering events or by the inclusion of samples that contain variations from the reference genome. Most well-clustered diploid SNPs have positive HetSO values as shown in Figure 28-1 (HetSO of 0.39).

$$\text{HetSo} = A_{ab} - A_{bb} - (A_{aa} - A_{bb}) \times \left(\frac{M_{ab} - M_{bb}}{M_{aa} - M_{bb}} \right)$$

where M_{aa} is the center of the AA cluster on the X axis and A_{aa} is the center of the AA cluster on the Y axis. $A_{ab} - A_{bb}$ is the observed difference between Y centers for the AB and BB clusters and the remaining terms of the HetSO formula produce the difference between the Y center for the BB cluster and the predicted Y center for the AB cluster based on the straight line running between the Y centers for the AA and the BB clusters.

Visually, SNPs with low HetSO show average signal value in the Y dimension that is much lower for the heterozygous cluster than for the homozygous clusters. Figure 28-2 shows a SNP with a very low HetSO value (–0.82). This is an OTV SNP and should either be removed from the downstream genotyping analysis or be reanalyzed with otv-caller. Figure 28-3 shows a multicluster SNP with one very large homozygous cluster in blue (BB), divided into several subclusters. The heterozygous AB cluster sits very far below the BB cluster and has a negative HetSO value (–0.35). Figure 28-4 shows a larger homozygous cluster in blue (BB) and a large cluster that has been split between heterozygous AB calls (yellow) and homozygous AA calls (red). This cluster split has caused the true heterozygous cluster to be called as the homozygous cluster. This produces a HetSO value of –0.18.

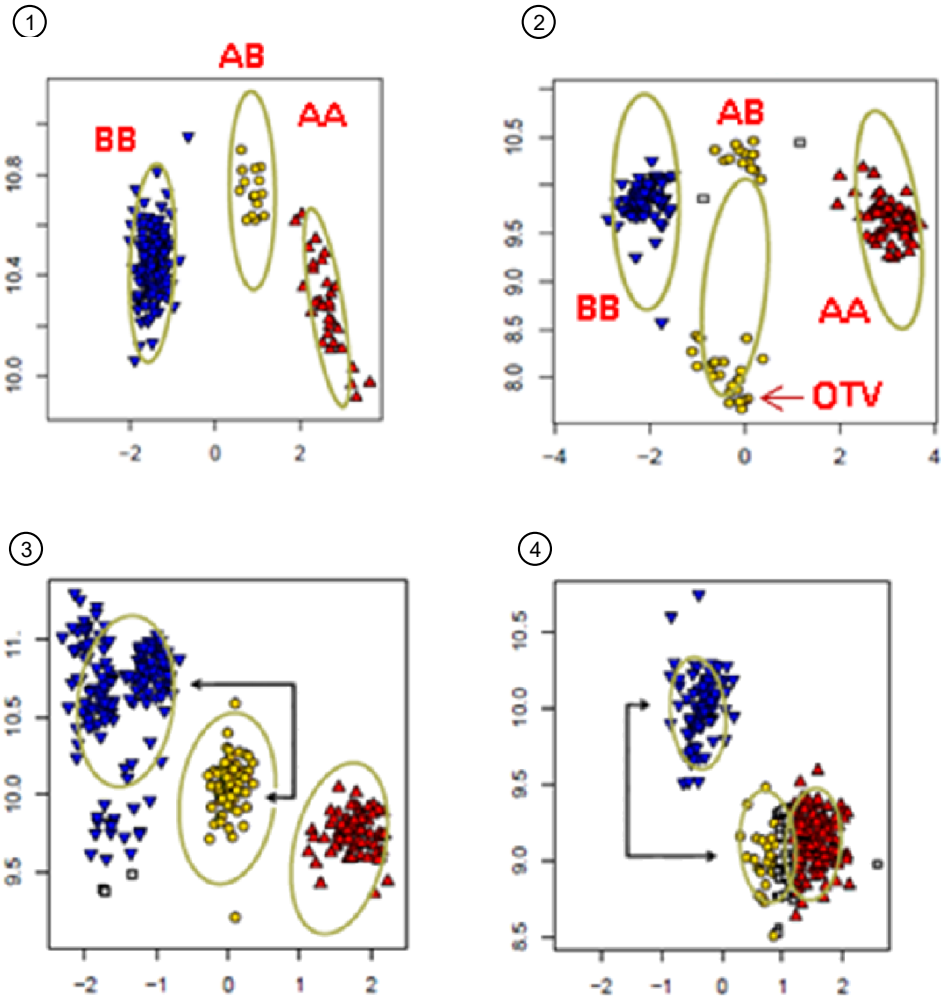


Figure 28 Examples of SNPs with different HetSO values.

1. HetSO = 0.39
2. HetSO = 0.82
3. HetSO = 0.35
4. HetSO = 0.18

Homozygote Ratio Offset (HomRO)

Homozygote Ratio Offset (HomRO) is the distance to zero in the contrast dimension (X axis) from the center of the homozygous cluster that is closest to zero. If there is only one homozygous cluster, HomRO is the distance from that cluster center to zero on the X axis.

$$\text{HomRO} = \begin{cases} min(M_{aa}, \|M_{bb}\|) & \text{If both homozygous clusters are on the correct side of 0} \\ -M_{bb} & \text{If both homozygous clusters are to the right of 0} \\ M_{aa} & \text{If both homozygous clusters are to the left of 0} \end{cases}$$

where *Maa* is the center of the AA cluster on the X axis, and *Mbb* is the center of BB cluster on the X axis.

The heterozygous cluster center should be found approximately at 0 on the X axis. If the clusters are shifted from their expected positions, then the heterozygous clusters are far away from zero. A negative or low value of HomRO generally indicates that

the algorithm has mislabeled the clusters. The AA cluster should be on the right side of zero (positive Contrast values) and the BB cluster should be on the left side of zero (negative Contrast values). A negative HomRO value implies that one of the homozygous clusters is on the wrong side of zero.

Figure 29-1 shows a misclustered SNP with a negative HomRO value (–0.51). The homozygous BB cluster (blue) is on the wrong (positive) side on the x-axis and the heterozygous AB cluster (yellow) is not over zero on the x-axis. Figure 29-2 shows a well clustered SNP with a positive HomRO value (2.21), where the AA (red) cluster is to the right of zero, the AB cluster (yellow) is over zero, and the BB cluster (blue) is to the left of zero, as expected

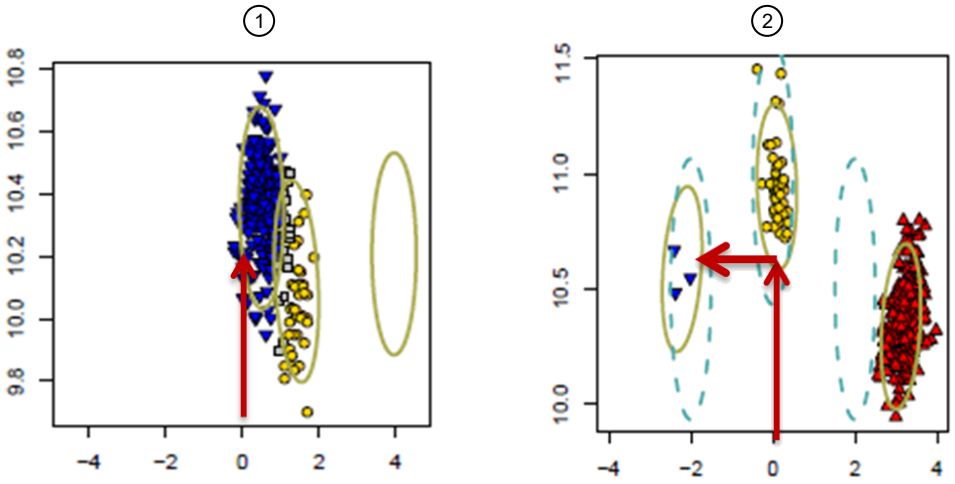


Figure 29 Examples of SNPs with low HomRO (1) and high HomRO (2).

(1) SNP with low HomRO (–0.51).
(2) SNP with high HomRO (2.21).

Base workflow

The base workflow in ps-metrics produces the 4 metrics listed above, the Hardy-Weinberg statistics and p-values, and 11 additional metrics. See below for more details on the Hardy-Weinberg test statistics and p-values. See Appendix A, "Complete Set of SNP QC Metrics Produced by ps-metrics" for details of the other 11 metrics. If a special SNPs and a report file are provided, then the metrics are calculated on the male and female samples and the mean locations for the clusters are recalculated from the summary data for each gender. If there are copy-number aware genotype calls in the data, additional copy-number aware metrics are calculated. If a genotype frequency file is provided, genotype frequency p-values are produced.

Table 8 Base workflow metrics.

Metric		
Call Rate (CR)	Number of minor alleles (nMinorAllele)	Hemizygous
Fisher's Linear Discriminant (FLD)	Minor allele frequency (MAF)	Minimum Mahalanobis Distance (MMD)
Homozygous FLD (HomFLD)	Number of AA calls (n_AA)	GenotypeFreqPvals
Heterozygous Strength Offset (HetSO)	Number of AB calls (n_AB)	HW test statistic
Homozygous Ratio Offset (HomRO)	Number of BB calls (n_BB)	HW p-value
Number of clusters (Nclus)	Number of NoCalls (n_NC)	

Supplemental workflow

The supplemental workflow in ps-metrics consists of the base workflow plus the recalculation of diploid cluster means and variances directly from the summary data. The supplemental workflow skips any hemizygous or non-PAR X or Z SNPs, any SNPs with copy-number aware genotypes, any multiallelic SNPs, and any SNPs with a reserved control marker name. The recalculated means and variances are displayed in Table 9. The z-scores are calculated as part of the matching supplemental workflow in ps-classification.

Table 9 Supplemental workflow recalculated metrics.

Metric		
AA.meanX	AB.meanX	BB.meanX
AA.meanY	AB.meanY	BB.meanY
AA.varX	AB.varX	BB.varX
AA.varY	AB.varY	BB.varY

SSP workflow

The SSP workflow should be run only under guidance from Thermo Fisher Scientific as part of the process for generating SNP-specific priors (SSPs). When the SSP workflow is run, all metrics from the supplemental workflow are calculated along with three metrics used in selecting the SSPs. The edge metrics (BB_dis_x_adj, AA_dis_x_adj) measure the closest point in the AA or BB cluster to any point in the AB cluster where the point in the AA or BB cluster must be within a certain vertical distance of points in the AB cluster. freq_diff measures the difference in the frequency of assigned calls between genders.

Multiallelic workflow

Ps-metrics has been updated to handle SNPs that are multiallelic: there is more than one alternate allele. A multiallelic posteriors file must be supplied to run the multiallelic workflow. Multiallelic SNPs and biallelic SNPs can be output together in the summary file (one row per allele per probeset) and the calls file (one row per probeset with dynamic numeric call assignments), but the formatting for the posteriors data is different for multiallelic and biallelic SNPs (see Figure 4 on page 17 (multiallelic posteriors file)). The multiallelic workflow is run automatically in ps-metrics for any multiallelic SNPs that appear in the input file. The biallelic metrics calculated in the base, supplemental, and SSP workflows are not appropriate for multiallelic SNPs. The multiallelic workflow calculates a different set of metrics that are useful with multiallelic SNPs.

The multiallelic workflow produces a large number of metrics, which can be split into three groups: counting metrics, minor allele frequency (MAF) metrics, and copy-number and background metrics. Genotype frequency p-values are also calculated for any multiallelic probesets that are listed in the genotype frequency file.

Counting metrics

The number of counting metrics is different for each dataset because each allele that appears in the dataset has the count of that allele calculated for each probeset. For example, if most of the SNPs in a dataset have 3 (A, B, C) or 4 (A, B, C, D) alleles but one SNPs has 5 alleles (A, B, C, D, E), then there will be 5 allele count metrics in the metrics file.

The metric *count_ma* is calculated for each allele on each SNP and is the number of times that allele appears in a genotype call for the probeset. The counts are named for each allele (e.g., *count_ma_A* for the A allele). SNPs where there is a *count_ma* metric for an allele that is not present in the genotype calls will have NA for that allele's count. In the example where one SNP has 5 alleles but most SNPs in the dataset have 3 or 4 alleles, the counts for *count_ma_E* will be NA for all SNPs except the one that has 5 alleles. Note that this is the count of the number of alleles that appear per SNP, not the number of assigned genotypes (for example, AA is 2 counts of allele A).

Call rate (*CR*), number of clusters (*Nclus*), number of assigned calls (*nCalls*), and number of NoCalls (*NC*) are self-explanatory. *nAllelesTested* and *nAllelesDetected* indicate the number of total possible alleles that a probeset can have and how

many of those alleles actually appeared in the assigned calls for that probeset. For example, a probeset may be designed to interrogate five alleles (A, B, C, D, E) but the samples that are genotyped only have three of the five alleles (A, B, D). In this case, *nAllelesTested* is 5 and *nAllelesDetected* is 3. *nAllelesTested* is always greater than or equal to *nAllelesDetected*.

NHetClus is the number of heterozygous clusters. *nMajorAlleles* is the maximum of all allele counts (*count_ma*). *maxMinorAllele* is the second largest allele count. *nMinorAlleles* is the sum of all allele counts without *nMajorAlleles*. *HomCount* is the number of samples in all homozygous clusters. *MajorHomCount* is the number of samples in the largest homozygous cluster. *MinorHomCount* is the number of samples in all homozygous clusters without *MajorHomCount*. *HetCount* is the number of samples in all heterozygous clusters. *nSamples* is the total number of samples (where non-diploid calls have been removed).

Minor allele frequency (MAF) metrics

Due to the presence of more than one minor allele, the conventional definition of minor allele frequency (MAF) is not appropriate for multiallelic SNPs. Ps-metrics calculates two different values for MAF using two different definitions. *MAFall* is the minor allele frequency across all non-major alleles (*nMinorAlleles*). *MAFmax* is the minor allele frequency using only the second largest allele count (*maxMinorAllele*). Both MAF's are calculated by taking the respective allele counts and dividing by the total number of alleles assigned in calls (2 times the number of assigned genotype calls that are not NoCall).

$$MAFall = \frac{nMinorAlleles}{2 * (nCalls - NC)}$$

$$MAFmax = \frac{maxMinorAllele}{2 * (nCalls - NC)}$$

Signal and background metrics

Several multiallelic metrics involve the comparison of the background signal values to the allele signal values. An allele can have different copy numbers per probeset, and the value depends on which samples are genotyped. Copy number values can be 0, 1, or 2. See Appendix A, “Complete Set of SNP QC Metrics Produced by ps-metrics” for a complete discussion of background and copy number.

HomMMA is the minimum of the means for each allele's homozygous cluster. The higher the *HomMMA* value, the stronger the allele signal is and the more defined the cluster.

Biallelic-derived multiallelic metrics

Several multiallelic metrics are analogous to bialellelic metrics and are created by transforming the multiallelic posterior cluster locations into size vs contrast space for each biallelic pair of alleles. For the FLD calculations, a weighted variance across all biallelic pairs is used to mimic the pooled variance found in the biallelic posteriors. The formulas for the transformed means and variances use means, variances, and covariances from the multiallelic posteriors file for each pair of alleles. These transformed means and weighted variances are used in calculating the biallelic-derived metrics for multiallelic probesets.

FLD_MA is the average FLD between all populated homozygous and heterozygous clusters across all biallelic pairs. Low *FLD_MA* values indicate low resolution clusters with little separation. For a cluster X, X_{hom} indicates all populated homozygous clusters, X_{het} indicates all populated heterozygous clusters, and $var_{weighted}(X)$ indicates the recalculated, weighted variances.

$$FLD_{MA} = average\left(\frac{|mean(X_{hom}) - mean(X_{het})|}{\sqrt{var_{weighted}(X)}}\right)$$

MinFLD_MA is the same calculation as *FLD_MA* except the minimum value is taken instead of the average value.

$$MinFLD_{MA} = minimum\left(\frac{|mean(X_{hom}) - mean(X_{het})|}{\sqrt{var_{weighted}(X)}}\right)$$

HomFLD_MA is similar to *FLD_MA* except it uses all pairwise combinations of populated homozygous clusters instead of homozygous-heterozygous pairs.

$$HomFLD_{MA} = minimum\left(\frac{|mean(X_{hom1}) - mean(X_{hom2})|}{\sqrt{var_{weighted}(X)}}\right)$$

Taken together, *FLD_MA*, *MinFLD_MA*, and *HomFLD_MA* can be used to determine if the clusters in a multiallelic probeset are well separated and tightly clustered.

HomRO_MA is the minimum of the absolute distance between the mean of a populated homozygous cluster and 0 on the X axis (contrast) across all the biallelic pairs. This is similar to taking the minimum of all biallelic *HomRO* values for every pairwise biallelic combination. *HomRO_MA* values should be much larger or smaller than 0, indicating that the locations of the homozygous clusters are far from the origin and have stronger non-zero signal.

Let M_{hom} be the mean of a homozygous cluster.

$$\text{HomRO\_MA} = minimum_{\forall i \neq j} \begin{cases} min(M_i, \|M_j\|) & \text{If both homozygous clusters are on the correct side of 0} \\ -M_j & \text{If both homozygous clusters are to the right of 0} \\ M_i & \text{If both homozygous clusters are to the left of 0} \end{cases}$$

As in biallelic genotyping, the heterozygous strength offset (*HetSO*) calculation uses the homozygous clusters and the heterozygous cluster per biallelic pair to determine where the heterozygous cluster is in relation to the homozygous clusters. It is expected to normally be situated above the homozygous clusters along the size dimension. *HetSO_MA* is the minimum *HetSO* across all the biallelic pairs. A high *HetSO_MA* value indicates that the heterozygous clusters are well separated from the homozygous clusters and their strong signal is visible by their placement above the homozygous clusters.

For all allele pairs i and j, let mean(Y_{ij}) be the mean in the Y dimension of the ij cluster and let mean(X_{ij}) be the mean in the X dimension of the ij cluster. Then *HetSO_MA* is:

$$\text{HetSO\_MA} = minimum_{\forall i \neq j} \left\{ mean(Y_{ij}) - \left(mean(Y_{jj}) + (mean(Y_{ii}) - mean(Y_{jj})) * \frac{mean(X_{ij}) - mean(X_{jj})}{mean(X_{ii}) - mean(X_{jj})} \right) \right\}$$

The full set of multiallelic metrics is displayed in Table 10

Table 10 Multiallelic workflow metrics.

Metric		
“count_ma” for each allele	nMinorAlleles	NC
CR	MAFall	HomMMA
nAllelesTested	HomCount	FLD_MA
nAllelesDetected	MajorHomCount	MinFLD_MA
Nclus	MinorHomCount	HomFLD_MA
NHetClus	HetCount	HetSO_MA
nMajorAlleles	nSamples	HomRO_MA
maxMinorAllele	nCalls	GenotypeFreqPvals

SNP categorization by ps-classification (Step 8B)

The four classification workflows can only be run if the appropriate metrics have been produced by ps-metrics. In APT, the user must select the workflow previously in ps-metrics. In Axiom™ Analysis Suite, ps-metrics and ps-classification are run together and the user only needs to select which complete workflow to run.

Base workflow

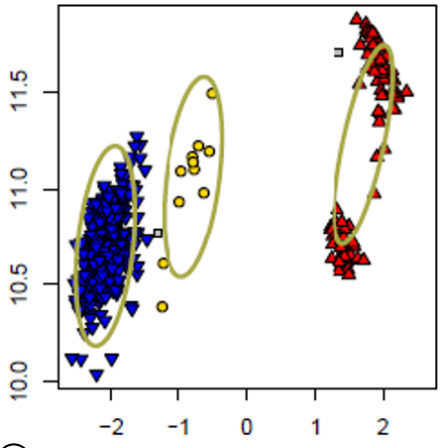
This workflow is used to classify biallelic SNPs using the base metrics file produced by ps-metrics. If the metrics file includes probesets that are special SNPs or that have copy-number aware genotyping, ps-classification uses the gender-separated metrics and the diploid/haploid/copy-number zero metrics in the more complex classification algorithm for biallelic probesets with more than the 3 basic genotype clusters. See Appendix B, “Complete set of classification thresholds used by ps-classification” for more information on the thresholds used with special SNPs classification and copy-number aware classification.

The base workflow classifies all probesets into the seven basic categories (PHR, NMH, MHR, CRBT, OTV, UnexpectedGenoFreq, Other), as well as the Hemizygous category if there are haploid probesets without haploid metrics.

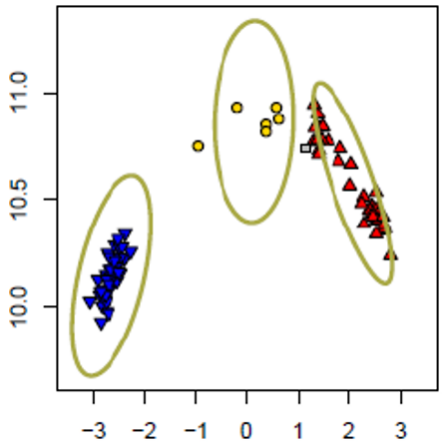
Supplemental workflow

The supplemental workflow uses additional filters that may be required for polyploidy organisms, those with highly repetitive or complex genomes, or when studying inbred populations. It is not run on hemizygous SNPs, non-PAR X SNPs, Z SNPs, copy-number aware genotyped SNPs, or multiallelic SNPs. This workflow provides numerous filtering options and classifies biallelic SNPs into seventeen potential categories, which consist of the original eight categories from the base workflow and nine additional categories. See Figure 30 and Figure 31 for a visual representation of these categories.

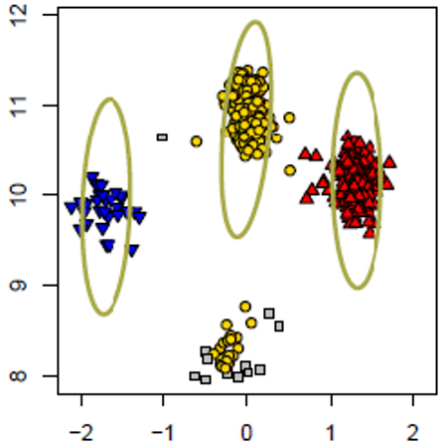
- nMinorHom: SNPs with a low minor homozygous count. The filter can be used to identify genotyped SNPs impacted by sample outliers.
- UnexpectedHeterozygosity: SNPs with heterozygosity higher than expected when compared to the minor allele frequency multiplied by a value.
- AAvarianceX: SNPs with high homozygous AA cluster variance in the X dimension.
- AAvarianceY: SNPs with high homozygous AA cluster variance in the Y dimension.
- ABvarianceX: SNPs with high heterozygous cluster variance in the X dimension.
- ABvarianceY: SNPs with high heterozygous cluster variance in the Y dimension.
- BBvarianceX: SNPs with high homozygous BB cluster variance in the X dimension.
- BBvarianceY: SNPs with high homozygous BB cluster variance in the Y dimension.
- HomHomResolution: SNPs with poor separation between homozygous clusters.



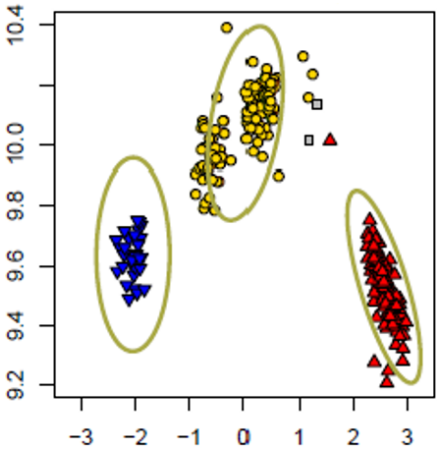
① AAvarianceY: High homoygous AA cluster variance in the Y dimension.



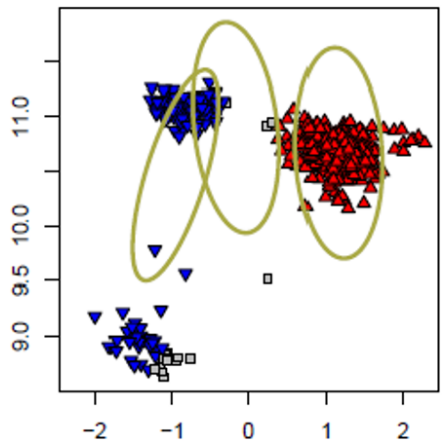
② AAvarianceX: High homozygous AA cluster variance in the X dimension.



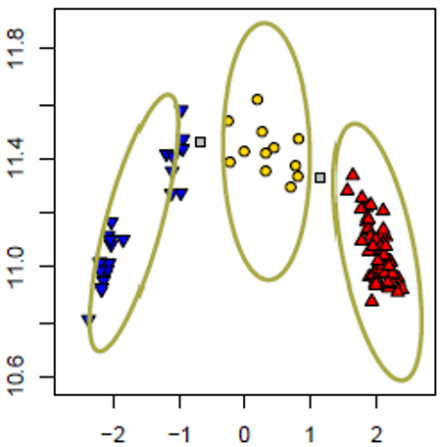
③ ABvarianceY: High heterozygous cluster variance in the Y dimension.



④ ABvarianceX: High heterozygous cluster variance in the X dimension.

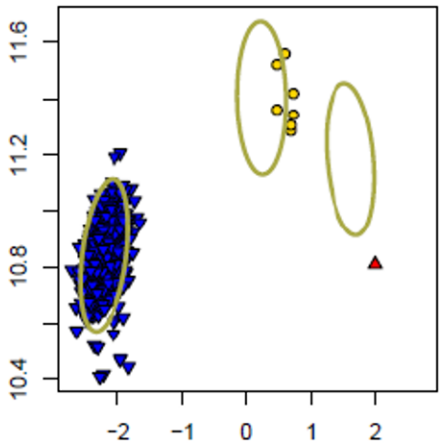


⑤ BBvarianceX: High homozygous BB cluster variance in the X dimension.

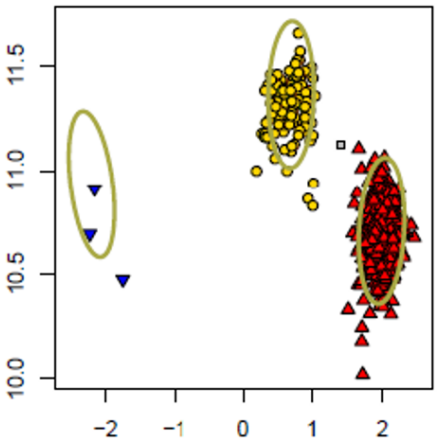


⑥ BBvarianceY: High homozygous BB cluster variance in the Y dimension.

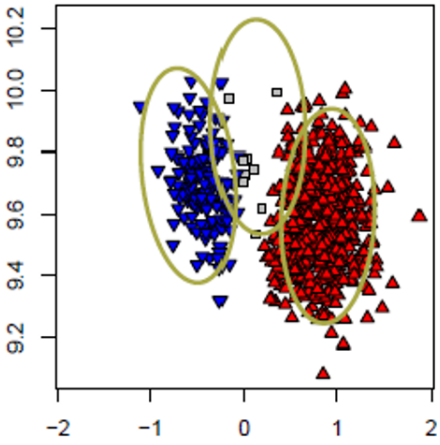
Figure 30 SNP cluster plots of categories from the supplemental workflow.



(1) nMinorHom: Low minor homozygous count. The filter can be used to identify genotyped SNPs impacted by sample outliers.



(2) UnexpectedHeterozygosity: Heterozygosity higher than expected when compared to the minor allele frequency multiplied by a value.



(3) HomHomResolution: Poor separation beteen homozygous clusters.

Figure 31 SNP cluster plots of categories from the supplemental workflow.

See Appendix B, “Complete set of classification thresholds used by ps-classification” for more information on the thresholds used in the supplemental workflow.

SSP workflow

The internal software to generate SNP-specific priors uses the supplemental categorizations, and so the supplemental and SSP workflows are identical in ps-classification with the exception that the output performance file contains the three additional SSP metrics when the SSP workflow is run in ps-metrics and ps-classification. This workflow only needs to be run under direction from Thermo Fisher Scientific as part of the SSP generation process.

Multiallelic workflow

The multiallelic workflow classifies multiallelic SNPs into seven categories:

- PolyHighResolution (PHR): SNPs with two or more alleles and well-separated genotype clusters
- NoMinorHom (NMH): SNPs with two or more alleles and well-separated clusters but only one homozygous cluster
- MonoHighResolution (MHR): SNPs with one allele detected and a well-formed genotype cluster
- UnexpectedGenotypeFreq: SNPs with a much larger than expected proportion of samples in one or more clusters
- CallRateBelowThreshold (CRBT): SNPs with a very low call rate
- Other: SNPs with more than one problematic issue
- OtherMA: SNPs where the original category was changed due to multiallelic best probeset selection

The biallelic category OTV is not appropriate for multiallelic SNPs.

See Appendix B, “Complete set of classification thresholds used by ps-classification” for the classification thresholds for multiallelic probesets.

Visualization of multiallelic SNPs is more complex than biallelic SNPs. See “Multiallelic plotting” on page 113 for more information on plotting multiallelic probesets.

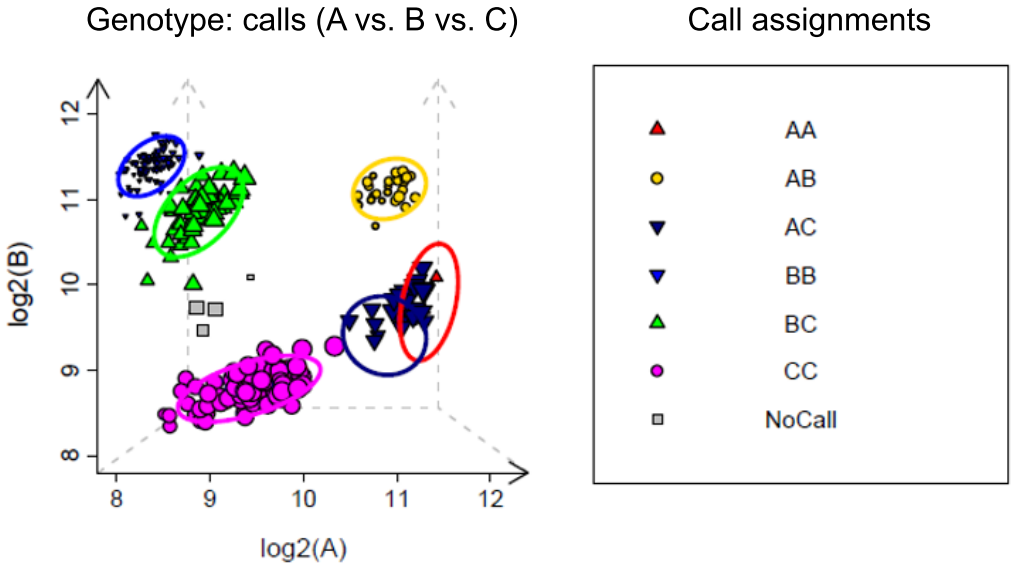


Figure 32 Multiallelic SNP that is classified as PolyHighResolution.

A multiallelic SNP that is classified as PolyHighResolution has multiple well-formed clusters that are separated in various biallelic combinations. The plot legend displays the shapes and colors assigned to each genotype call.

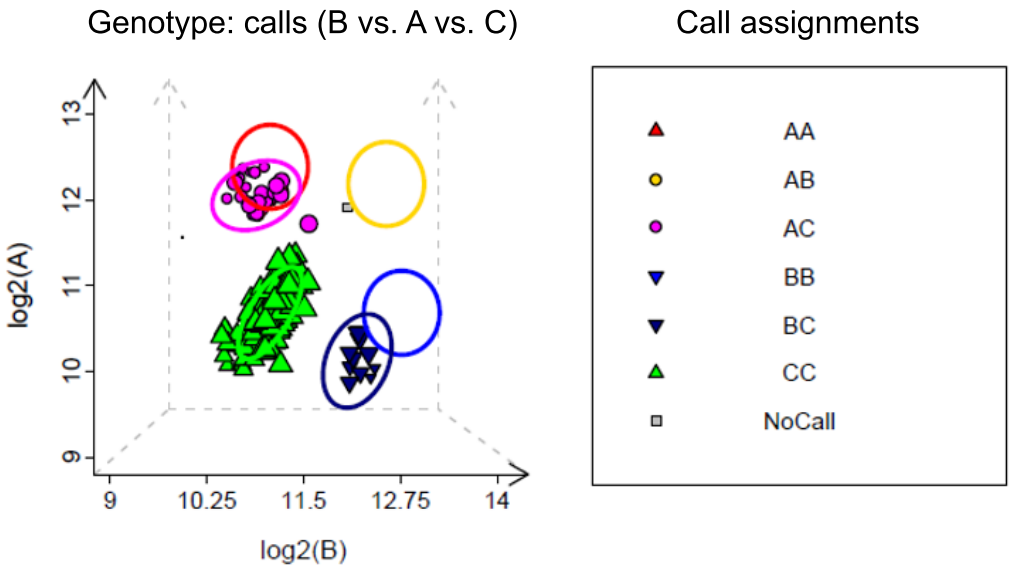


Figure 33 Multiallelic SNP that is classified as NoMinorHom.

A multiallelic SNP that is classified as NoMinorHom has two or more alleles detected but contains at most one homozygous cluster. Hemizygous multiallelic probesets can never be classified as NoMinorHom.

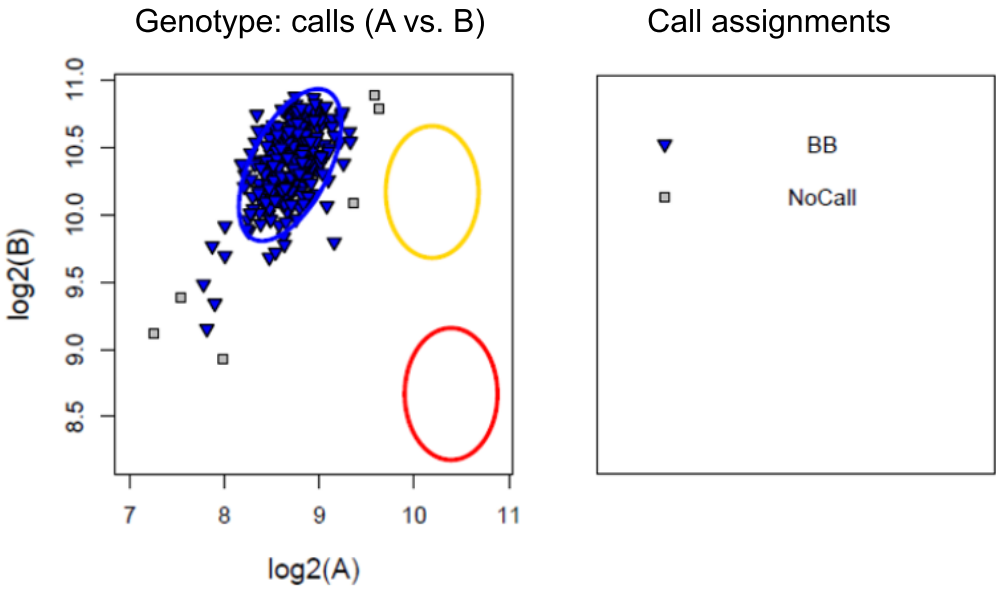


Figure 34 Multiallelic SNP that is classified as MonoHighResolution.

A multiallelic SNP that is classified as MonoHighResolution has only one homozygous cluster that is well-clustered and in the correct location.

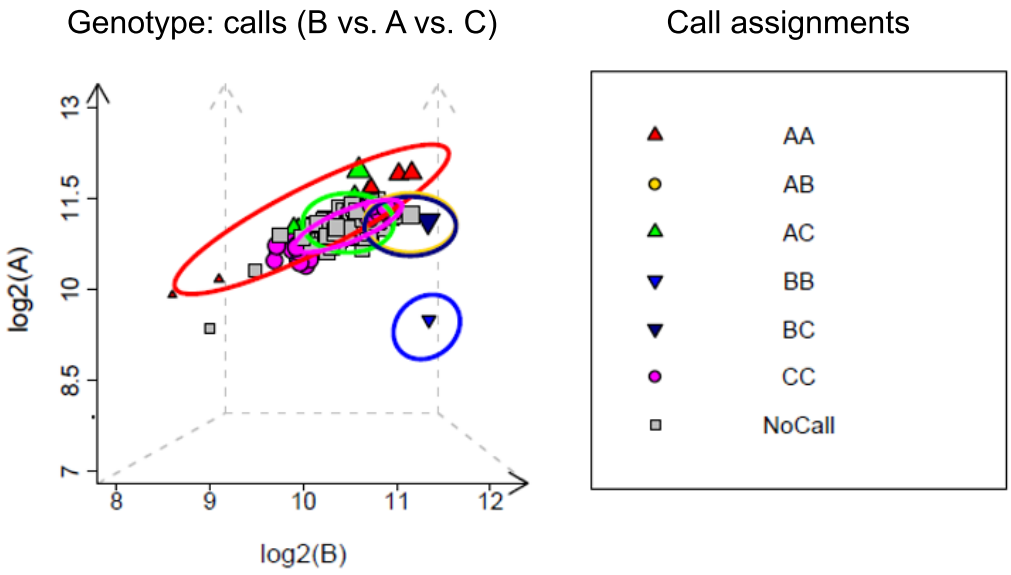


Figure 35 Multiallelic SNP that is classified as CallRateBelowThreshold.

A multiallelic SNP that is classified as CallRateBelowThreshold has tightly clustered and well separated clusters but a very high NoCall rate.

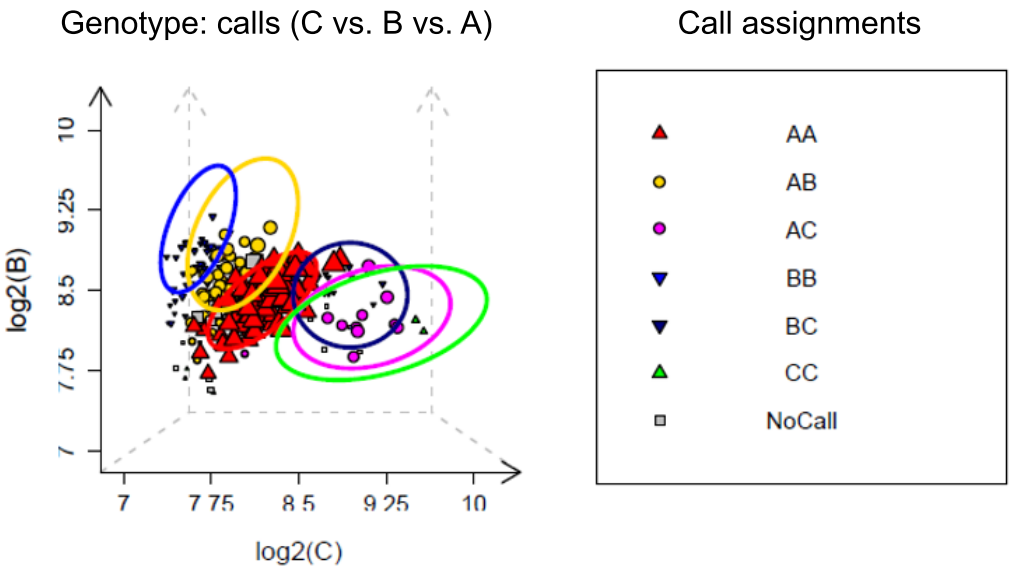


Figure 36 Multiallelic SNP that is classified as Other.

A multiallelic SNP that is classified as Other has clusters that are not well separated or clearly defined.

Additional SNP metrics that can be used for SNP filtering

This section describes additional SNP metrics (Hardy-Weinberg p-value, Mendelian trio error, and Genotyping Call Reproducibility) that can also be appropriate to examine as part of the SNP filtering process. Hardy-Weinberg p-values (H.W.pValue) are calculated by ps-metrics. Axiom™ Analysis Suite has features to calculate sample reproducibility. No Axiom™ software is provided for calculating Mendelian trio error.

For these additional metrics, absolute QC and pass/fail thresholds can only be set in the context of the study design. The general guideline is to examine the distribution of each metric, and then examine cluster plots for SNPs with outlier values and over a collection of randomly selected SNPs.

Thresholds can be set based on consideration of three properties:

- the absolute value of the metric,
- the deviation from the mean/median values, and
- the expectation (based on an examination of cluster plots) that SNPs below a threshold are likely to be misclustered.

Hardy-Weinberg p-value

The Hardy-Weinberg p-value (H.W.pValue) is a measure of the significance of the difference between the observed ratio of heterozygote calls in a population and the ratio that is expected if the population is in Hardy-Weinberg equilibrium (HWE). The test should be performed on unrelated individuals with relatively homogeneous ancestry.

P-values are calculated per probeset according to the test for deviation away from the expected allele frequencies under the Hardy-Weinberg equilibrium model. For two alleles A and a, the expected frequencies for n samples under population equilibrium are:

$$\begin{aligned} \mathrm{E(AA)} &= p^2 n \\ \mathrm{E(Aa)} &= 2p(1-p)n \\ \mathrm{E(aa)} &= (1-p)^2 n \end{aligned}$$

The p-values are calculated from a chi-squared test which compares the expected values with the observed values:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

The chi-squared test is used when there are 10 or more samples per cluster. When one or more clusters have fewer than 10 samples, the chi-squared test is not appropriate and Fisher's Exact Test is used to directly calculate the p-value. There is no test statistic and the value in the HW.chisquared.statistic column is NA when the Exact Test is performed.

Although genotyping artifacts can produce low H.W.pValue values, using this as a SNP QC metric can be tricky because a low p-value can be caused by true genotypic frequency deviation. Examination of cluster plots indicates that most extreme deviations (pvalue<1e-09) are due to poorly performing SNPs. In Axiom™ Analysis Suite, the SNP Summary Table provides the HW.chisquared.statistic and H.W.pValue metrics.

Mendelian trio error

Mendelian errors can be detected in parent-offspring trios. Mendelian trio error rate is calculated as the number of errors that are detected in a particular family divided by the number of families in which the offspring and parents have available genotypes. This method of error detection is less efficient than other methods because many genotyping errors are consistent with Mendelian inheritance (for example, the offspring of AB and BB parents can have a true BB genotype but is called as AB and this error will not influence the Mendelian trio error rate). SNPs that have high Mendelian error rates in the study should be examined in cluster plots for symptoms of mis-clustering.

Genotyping call reproducibility

SNP genotyping error rates can be estimated from the reproducibility of genotype calls (excluding No Calls) of replicated samples. One approach is to use duplicated pairs of samples and count the number of pairs with discordant calls. Given that mean error rates are low, a large number of duplicated pairs are required to provide enough precision to meaningfully detect SNPs with error rates significantly higher than the main body of the SNPs (the overall error rate is still low in absolute value). As discussed in Laurie et al., (2010) approximately 30 duplicated pairs of samples are needed to generate enough precision for this type of analysis. Discordance rates can also be calculated from the ~60 samples divided into replicate sets of greater than two. In this case, a slightly more complicated algorithm is required. For each replicated sample set, the approach is to first calculate a consensus genotype for the sample at the SNP. The number of discordant calls for the sample set equals the number of samples in the set whose genotype does not agree with the consensus genotype. The total number of discordant calls for the SNP equals the sum of discordant calls over the sample sets. DNA sample quality can vary considerably, and these differences in sample quality can influence the genotyping call error rates among samples. Therefore, the replicated sample sets should consist of at least five different study samples, and if any of the specific samples or plates are poorly performing outliers, they should be removed from use in the reproducibility test. If this quantity and variety of replicates are not available, reproducibility can still be used as a coarse filter for SNPs with obvious low values.

To calculate sample reproducibility over Best and Recommended probesets in Axiom™ Analysis Suite, first filter the **Probeset Summary Table** tab for **Best and Recommended** probesets. From the **Sample Table** tab, select **Compare all combinations and Compare SNPs within Probeset Summary Table**, then click **OK**.

7 Execute Best Practices steps with Axiom™ Analysis Suite

This chapter provides an overview of the QC and genotyping workflows to be used in Axiom™ Analysis Suite version 1.0 and higher

Execute steps 1-8 with Axiom™ Analysis Suite

Axiom™ Analysis Suite setup

Axiom™ Analysis Suite is designed to execute all parts of the Best Practices Workflow in one program. In this program all of the QC, library files and SNP QC settings are entered in the **New Analysis** tab of the software. This is recommended methodology of executing the Best Practices Workflow. Figure 37 shows the full workflow for Axiom™ Analysis Suite.

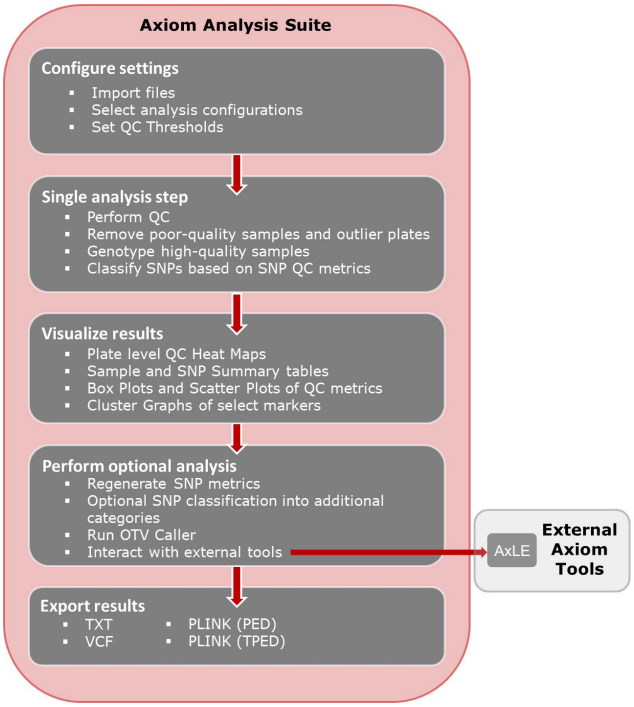


Figure 37 Full best practices workflow using Axiom™ Analysis Suite.

Analysis library files and annotation files can be directly downloaded from within the software, or they can be manually downloaded from **www.thermofisher.com** and unzipped into the current library folder.

For more detailed instructions on how to install Axiom™ Analysis Suite, obtain analysis library and annotation files, or set up a new analysis batch, consult the *Axiom™ Analysis Suite User Guide* (Pub. No. 703307), available at the support section of the Thermo Fisher website.

Step 1: Group samples into batches

Axiom™ Analysis Suite is designed for handing batches up to 50 plates of samples. Please see “Step 1: Group sample plates into batches” on page 24 Step 1 for information on batch recommendations.

To add samples to the analysis batch, click **Import CEL Files**, navigate to your CEL file location and highlight the samples you wish to add (Figure 38).

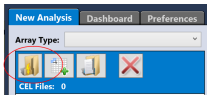


Figure 38 Import CEL Files button.

Setup step 2, 3, 5, 6 and 8A, B: Set sample ps-metrics, plate ps-metrics, and SNP ps-metrics

All of these steps are entered at the same time in Axiom™ Analysis Suite. The Threshold Settings window provides a single location to enter and edit all of the ps-metrics that are associated with the Best Practices Workflow (Figure 39). Three default configurations are available: human, diploid, and polyploid. Some array types may have array-specific configurations. To run the analysis, select the appropriate default configuration. See Chapter 3, “Best Practices Genotyping Analysis Workflow” and Chapter 6, “SNP QC metrics and classification” of this guide for more information on the thresholds.

Figure 39 Threshold settings window.

Step 4 and 7: Generate sample QC call rate using step1.AxiomGT1 and genotype passing samples and plates over step2.AxiomGT1 SNPs

The **Analysis Settings** window provides a single location for setting the appropriate library files for both step1 and step2 analysis (Figure 40). Typically two default configurations are available for an array: <96 samples or ≥96 samples, though some arrays can have more than two available. The default analysis configuration automatically selects the correct analysis .xml file. Additional optional settings are available for use, such as inbred penalty for inbred samples and hints files. See section (“Genotyping inbred samples” on page 53) for more information on inbred samples. Hints files are not recommended for most users.

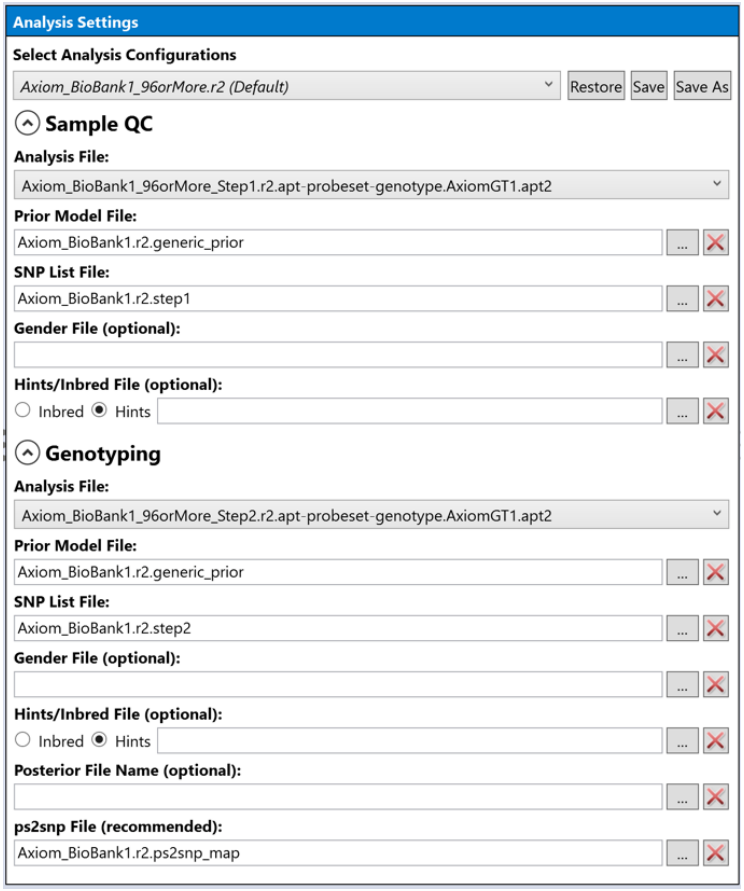


Figure 40 Analysis Settings window.

Select the analysis configuration appropriate for your study based on the number of samples and use any optional settings if desired, for example inbred penalty file if your samples are inbred. If using an inbred penalty, you should ensure to load it for both sample QC and genotyping.

Run analysis and review data

After setting up all three windows of the **New Analysis** tab, clicking **Run Analysis** executes all QC steps with the library files provided (Figure 41). Make sure to select an appropriate Output Folder if different from the default location and create a Batch Name for the analysis batch before starting the analysis run. After the analysis is finished, the review the results.

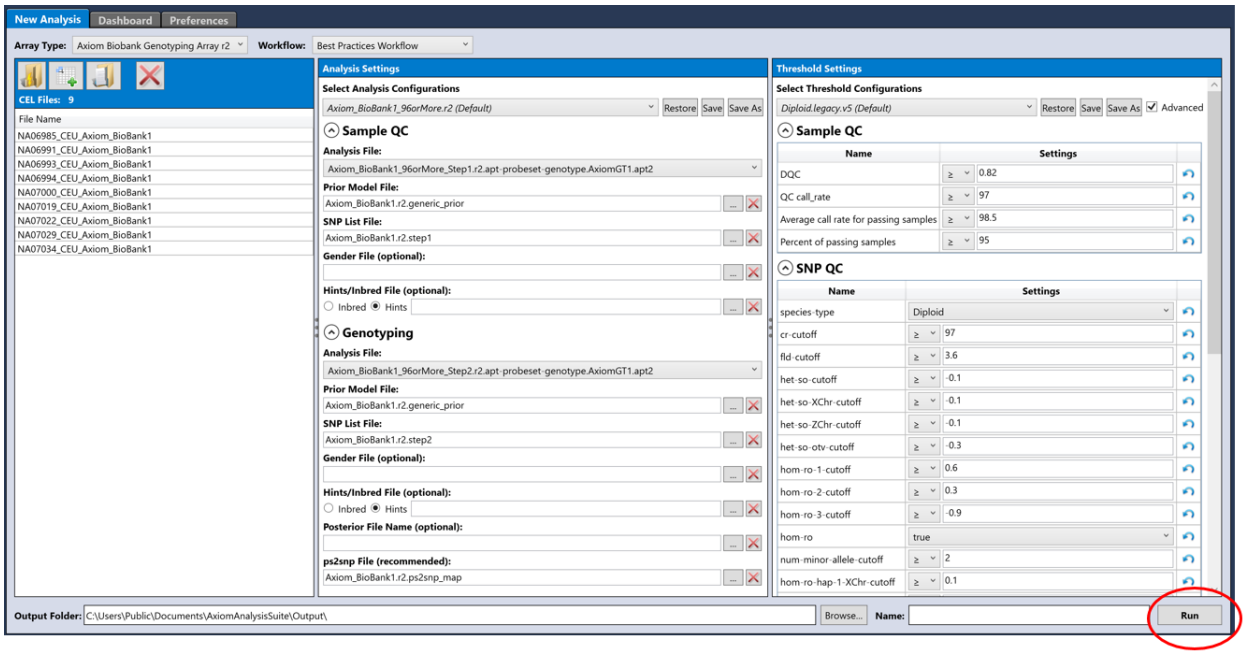


Figure 41 Run analysis button.

Axiom™ Analysis Suite creates a batch analysis folder with all of the data from the batch. Right-clicking the folder allows you to open the data in the Axiom™ Analysis Suite Viewer. Three tabs are available on the left half of the screen. Five plots are made automatically for QC purposes and the cluster plots are available. The **Summary** tab provides an overview of the analysis (Figure 42). The **Sample Table** tab provides sample level information (Figure 43). We recommend reviewing all of the QC plots created. The **Probeset Summary** tab provides SNP level information (Figure 44). See the *Axiom™ Analysis Suite User Guide* (Pub. No. 703307) for more information on these tabs and the default plots created.

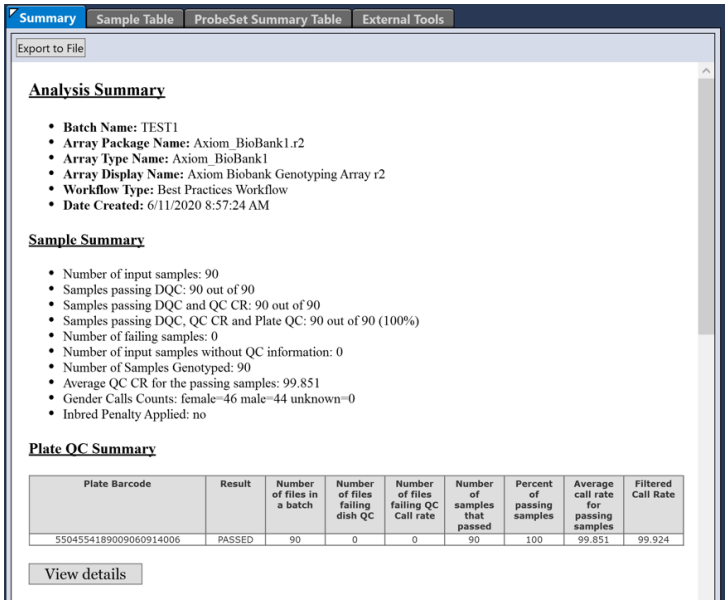


Figure 42 Summary tab.

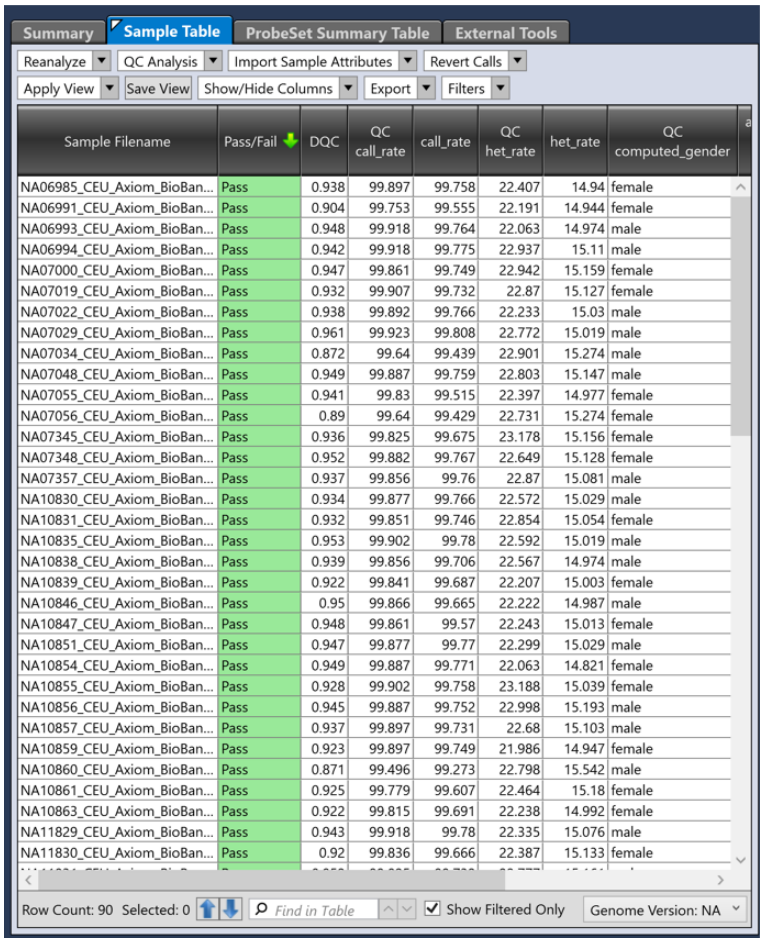


Figure 43 Sample Table tab.

probeset_id	affy_snp_id	ConversionType	BestandRecommended	CR	MinorAlleleFrequency	H.W.p-Value
AFFX-KIT-000001	Affx-3289...	PolyHighResol...	0	100	0.456	0.1
AFFX-KIT-000002	Affx-2381...	PolyHighResol...	1	100	0.167	0.7
AFFX-KIT-000003	Affx-2601...	PolyHighResol...	0	100	0.128	0.6
AFFX-KIT-000004	Affx-2641...	PolyHighResol...	0	100	0.5	0.8
AFFX-KIT-000005	Affx-1052...	NoMinorHom	1	98.889	0.0337	
AFFX-KIT-000008	Affx-1820...	CallRateBelowT...	0	94.444	0.447	0.
AFFX-KIT-000009	Affx-8147...	PolyHighResol...	1	100	0.0833	0.
AFFX-KIT-000012	Affx-3000...	PolyHighResol...	1	98.889	0.416	0.2
AFFX-KIT-000013	Affx-3127...	PolyHighResol...	1	98.889	0.152	
AFFX-KIT-000014	Affx-9152...	PolyHighResol...	1	100	0.183	0.7
AFFX-KIT-000015	Affx-1394...	PolyHighResol...	1	100	0.156	
AFFX-KIT-000016	Affx-2607...	PolyHighResol...	1	100	0.222	0.5
AFFX-KIT-000017	Affx-8225...	NoMinorHom	1	100	0.0778	
AFFX-KIT-000018	Affx-2735...	PolyHighResol...	1	100	0.311	0.2
AFFX-KIT-000019	Affx-8280...	PolyHighResol...	1	100	0.461	0.3
AFFX-KIT-000021	Affx-2431...	CallRateBelowT...	0	96.667	0.282	0.02
AFFX-KIT-000022	Affx-2739...	PolyHighResol...	1	100	0.444	0.4
AFFX-KIT-000023	Affx-8535...	NoMinorHom	1	100	0.0389	
AFFX-KIT-000025	Affx-8120...	PolyHighResol...	1	100	0.489	0.8
AFFX-KIT-000026	Affx-2157...	PolyHighResol...	1	100	0.467	0.7
AFFX-KIT-000027	Affx-3117...	PolyHighResol...	0	100	0.389	0.2
AFFX-KIT-000029	Affx-4803...	PolyHighResol...	1	98.889	0.129	0.03
AFFX-KIT-000031	Affx-6946...	PolyHighResol...	1	100	0.133	0.6
AFFX-KIT-000032	Affx-1142...	PolyHighResol...	1	100	0.428	0.
AFFX-KIT-000033	Affx-1616...	PolyHighResol...	1	100	0.489	0.8
AFFX-KIT-000035	Affx-1339...	PolyHighResol...	1	100	0.278	0.6
AFFX-KIT-000036	Affx-1995...	PolyHighResol...	1	100	0.278	
AFFX-KIT-000037	Affx-1328...	PolyHighResol...	1	100	0.461	0.1
AFFX-KIT-000038	Affx-2599...	PolyHighResol...	1	100	0.494	0.3
AFFX-KIT-000040	Affx-4525...	PolyHighResol...	1	100	0.394	0.
AFFX-KIT-000041	Affx-4366...	NoMinorHom	1	98.889	0.107	0.5
AFFX-KIT-000042	Affx-3333...	Other	0	100	0.0889	0.1
AFFX-KIT-000043	Affx-3435...	PolyHighResol...	1	100	0.283	0.

Figure 44 Probeset Summary Table tab.

Visualize SNPs and change calls through Axiom™ Analysis Suite cluster graphs

Axiom™ Analysis Suite contains functions for plotting SNP cluster graphs (“What is a SNP cluster plot for AxiomGT1 genotypes?” on page 19) and produces plots that are similar to the output from Ps_Visualization. Since Axiom™ Analysis Suite executes all steps of the Best Practices Workflow, the cluster graphs are made automatically. For a more detailed introduction to the SNP Cluster Graph function, see the *Axiom™ Analysis Suite User Guide* (Pub. No. 703307).

The SNP Cluster Graph allows you to adjust the shape and color of the samples. Figure 45 shows a cluster plot where the AA cluster is red, the AB cluster is yellow, the BB cluster is blue; female samples are plotted as triangles and male samples are plotted as circles. You can select and change the calls of samples through the plotted cluster graph. See “Evaluate SNP cluster plots” on page 40 for interpretation of cluster graphs. In the SNP Summary table, the Conversion Type column provides the category that ps-classification has classified a SNP to be in. It is recommended that each category of SNPs be visually reviewed.

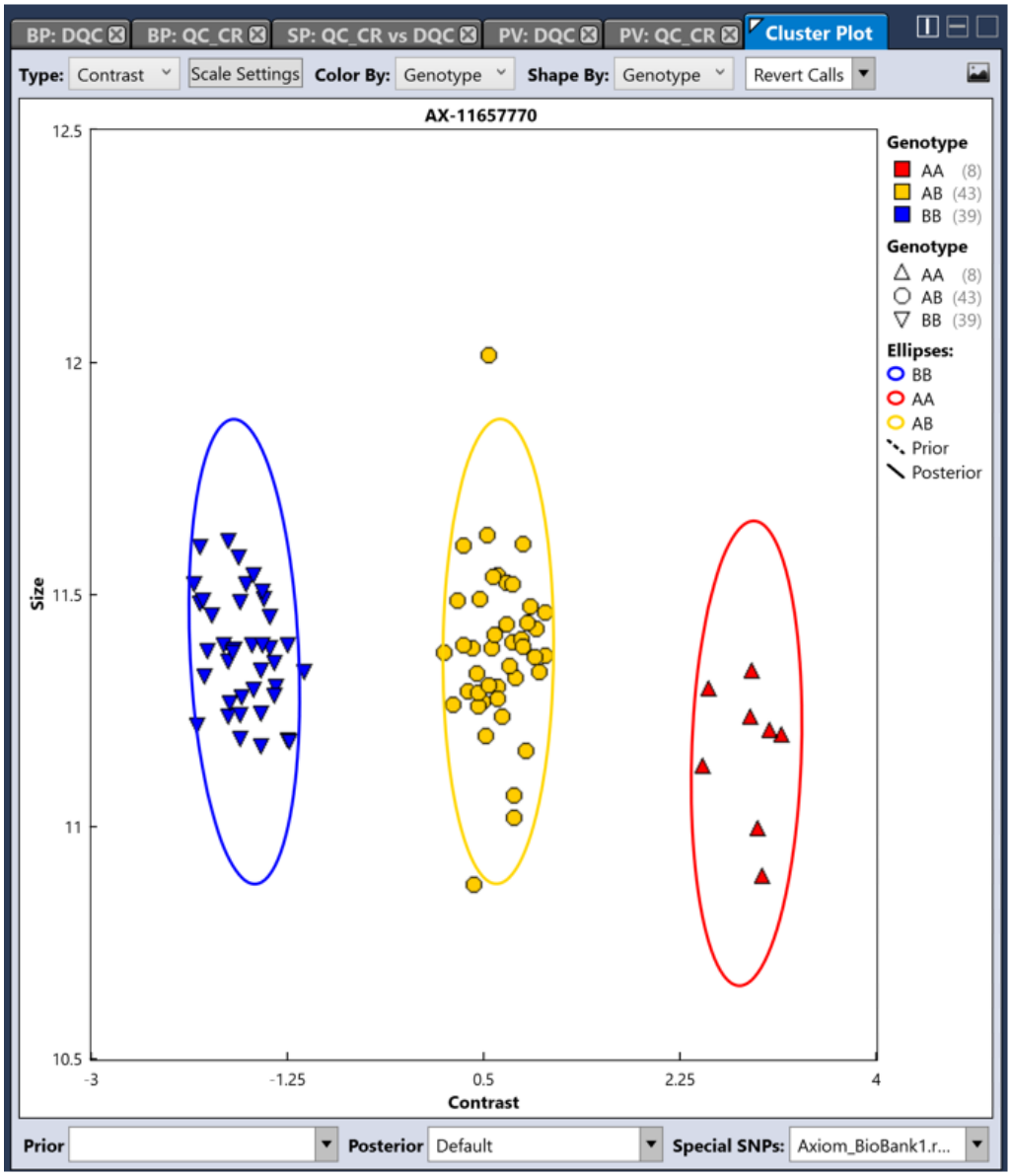


Figure 45 SNP cluster graph.

Display a particular SNP

To display a particular SNP, click the corresponding row in the SNP Summary Table. The cluster graph updates to display the data for the SNP.

Select a single sample

To select a single sample, click the data point in the SNP cluster graph. The selected sample is highlighted in the Sample Table (Figure 46).

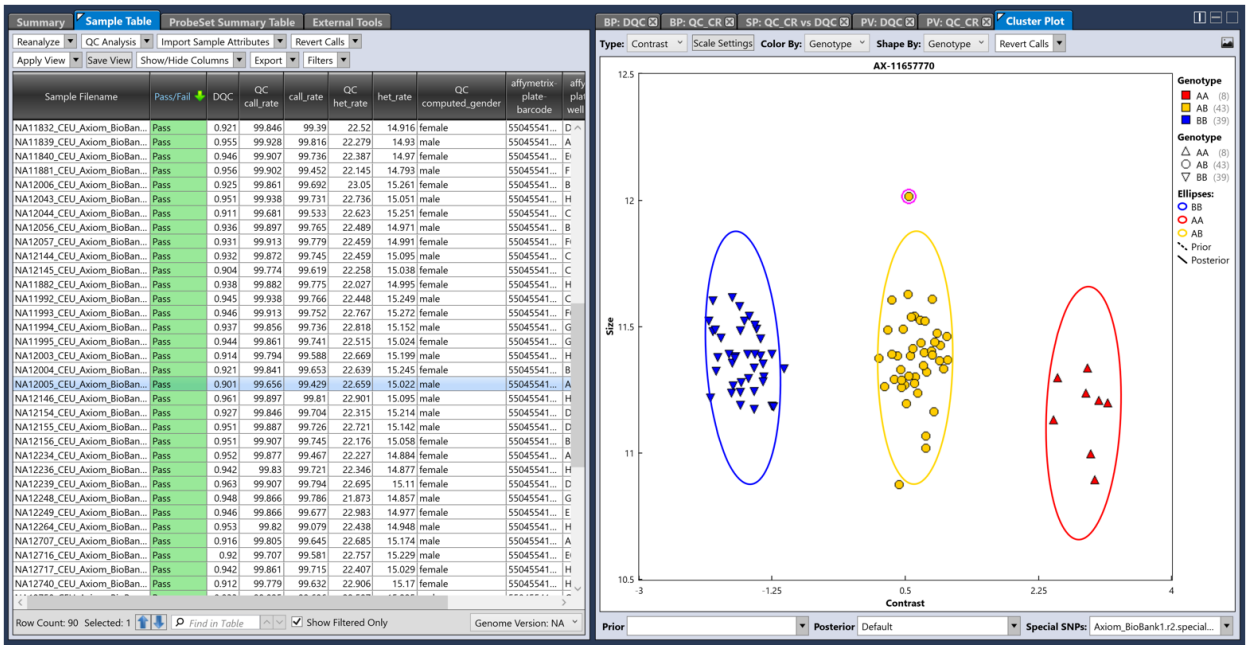


Figure 46 Selected sample is highlighted in sample table.

Select multiple samples

To select multiple samples, draw a closed shape around a group of samples by clicking the plot and circling the samples with the mouse before releasing the mouse button (Figure 47). The lasso function automatically draws a straight line to the starting point of the shape if the mouse button is released before the shape is closed. The samples in the group and the rows in the Sample Table are selected when the button is released.

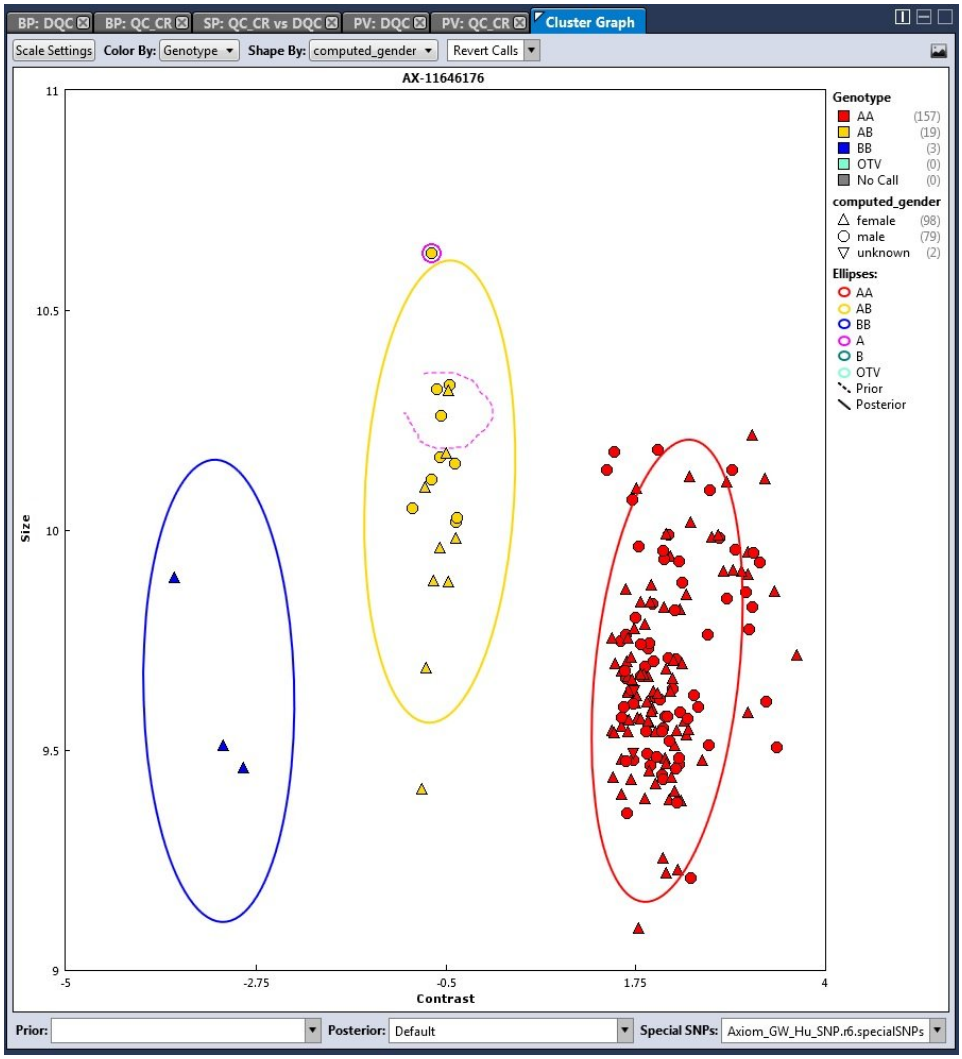


Figure 47 Use the lasso function to select multiple samples.

Manually change a sample's call

To manually change a sample's call, click the sample to select it, then right-click. The Change Call menu appears. Select the new call (Figure 48).

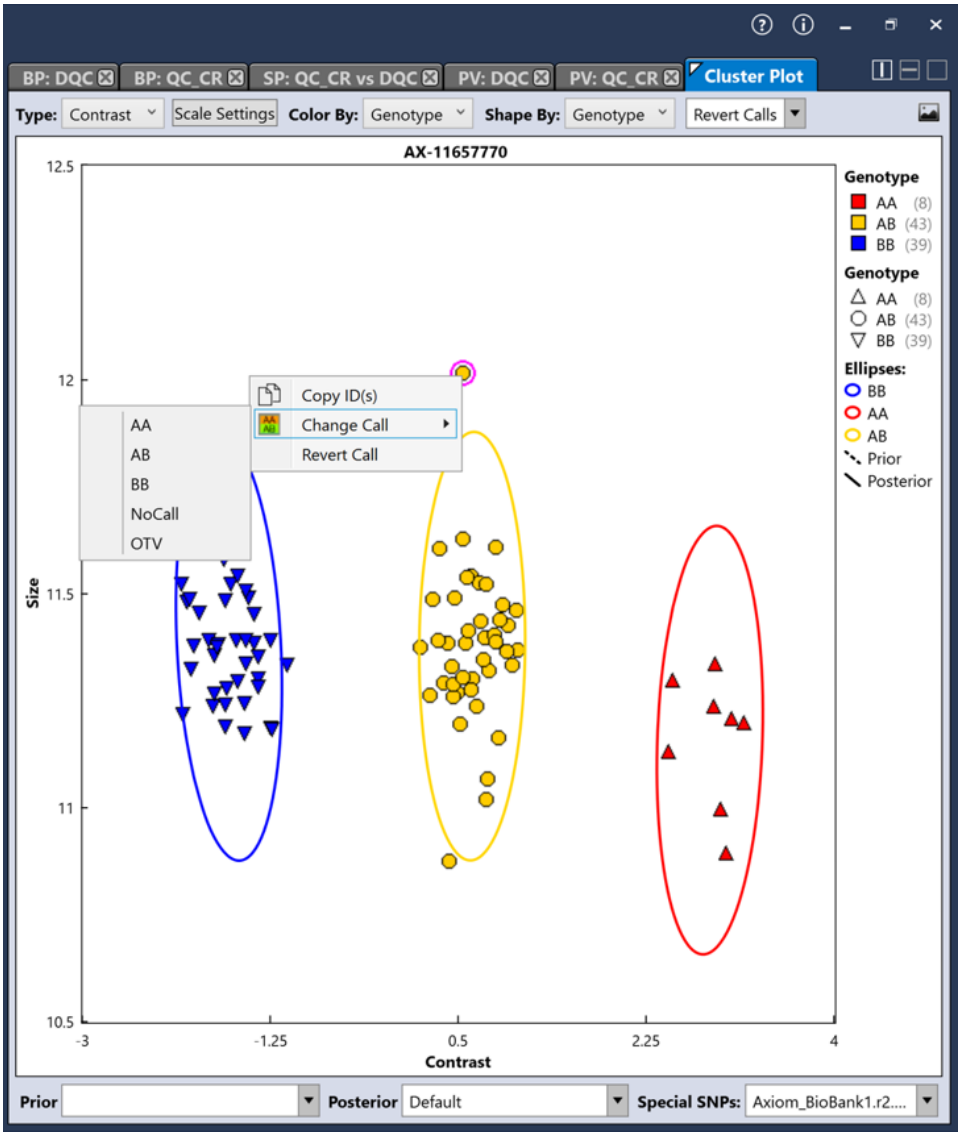


Figure 48 Manually change a call.

Lasso function

The lasso function can be used in several different cases including cluster splits. In Figure 49, the top half shows a cluster split, and the bottom image shows the graph after setting the samples correctly to AB.

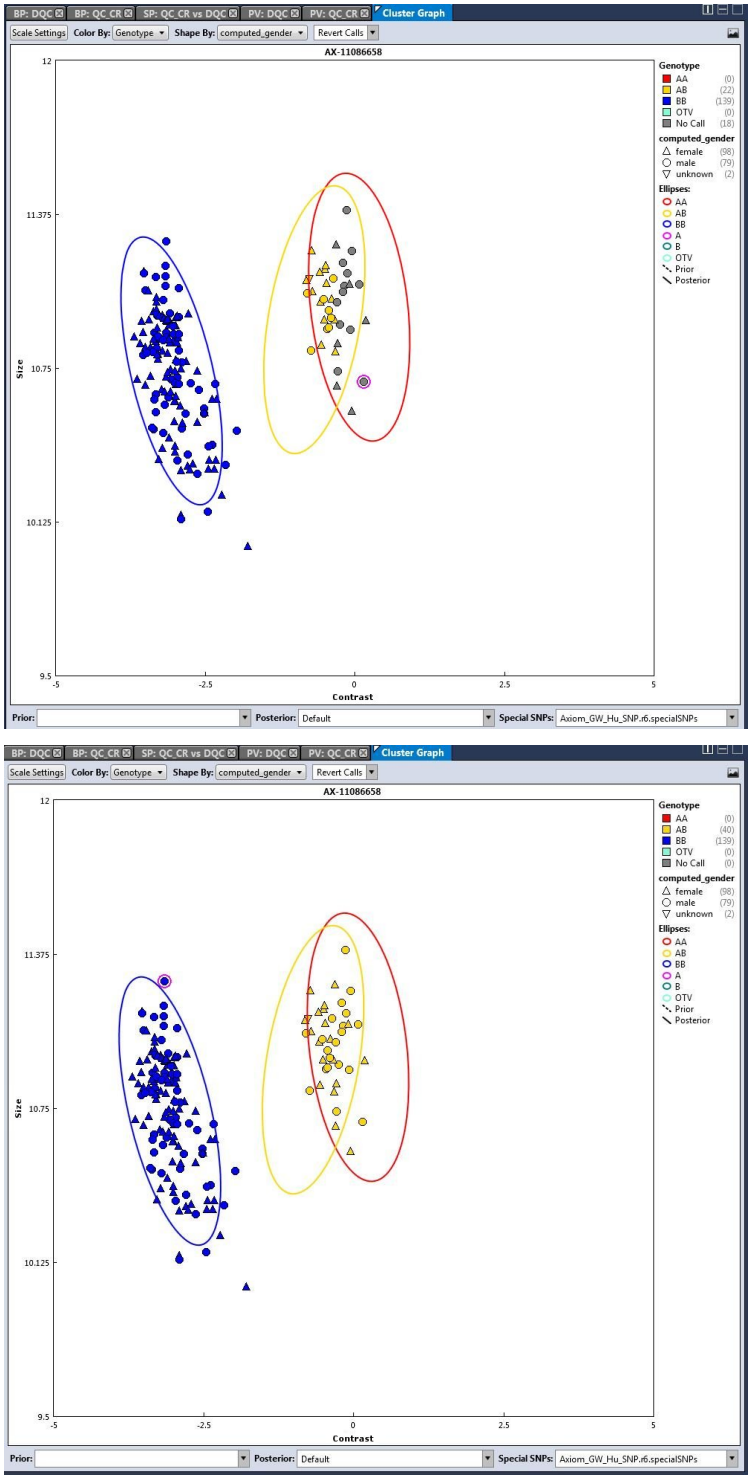


Figure 49 Use the lasso function to change calls in a cluster split.

Save a cluster plot

There are two different methods for saving a cluster plot:

- To save a single plot, the save image button is in the upper right corner of the cluster plot tab.

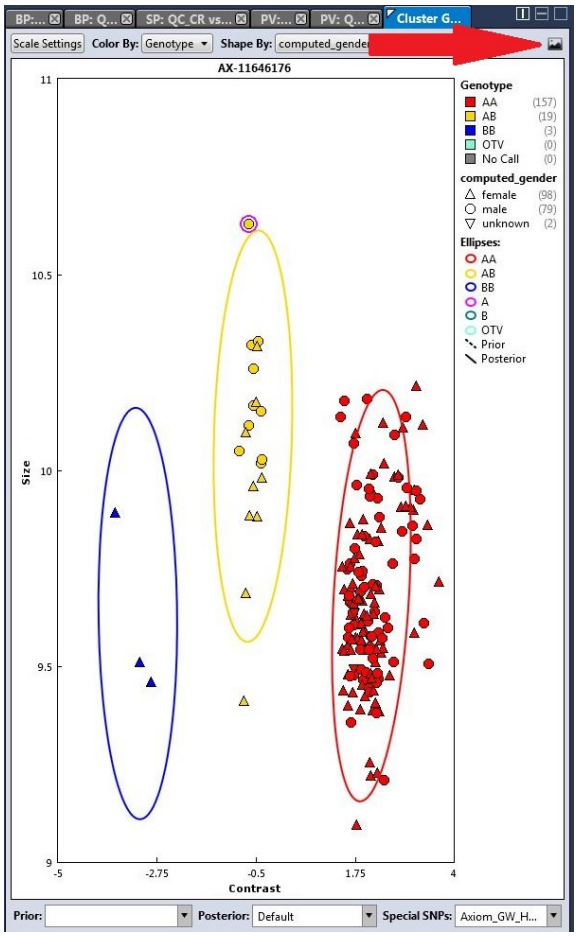


Figure 50 Cluster plot save image button.

- To save multiple cluster plots to a single PDF file do the following:

1. Click the **Export** dropdown in the **Probeset Summary Table**, then choose **Export cluster plots to PDF**.

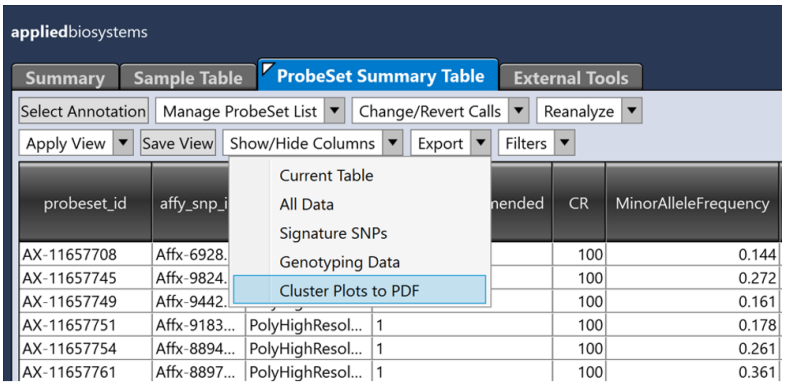


Figure 51 Export multiple cluster plots to PDF file format.

2. This opens the Report Settings window. Select to export all SNPs from Current Table or Random SNPs from Current Table.

Step 8C: Create a recommended SNP list

There are two ways to create a recommended probeset list: importing the automatically generated recommended probeset list or creating one manually. If you are doing no additional analysis (such as otv-caller), you can use the automatically generated probeset list. If you are doing additional analysis, you should create the recommended probest list manually. Axiom™ Analysis Suite creates a recommended SNP list based on the "Recommended" settings in the **Threshold Analysis** tab (Figure 52) and selecting the best probe set for a marker as indicated in the **BestProbeset** column of the **Probeset Summary** tab. By default this setting matches Table 6. To import this SNP list into Axiom™ Analysis Suite:

1. Click the **Manage SNP List** dropdown.
2. Click the **Import SNP List to Batch** option.
   The recommended.ps is under the "SNPolisher" folder of the batch folder.

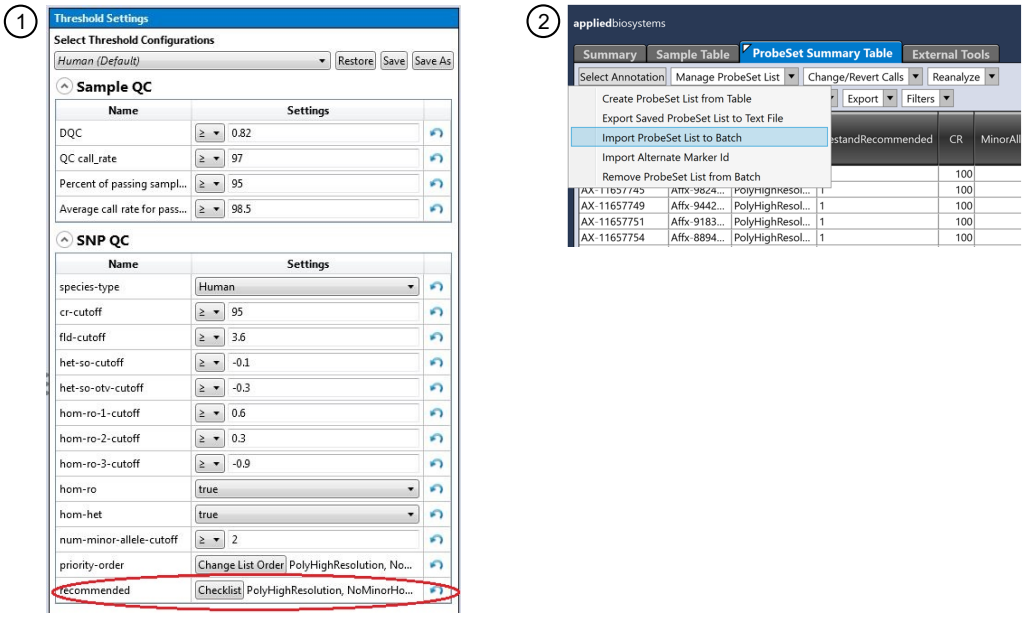


Figure 52 Recommended SNP lists.

(1) Recommended settings in threshold analysis tab.
(2) Import SNP list to batch.

If you want to create your own recommended probe set list, do the following:

- Filter the Probeset Summary Table on the **Conversion Type** by right-clicking the column header and applying a filter (Figure 53), also, SNPs should be filtered on the **Best Probeset** column.
- Click the **Manage SNP List** dropdown and click **Create SNP List from Table**.
- Name the SNP list and click **OK.**
- Under the **Manage SNP List** dropdown, click **Export Saved SNP List to Text File**.
- Select the previous saved SNP list from the dropdown and click **OK**.

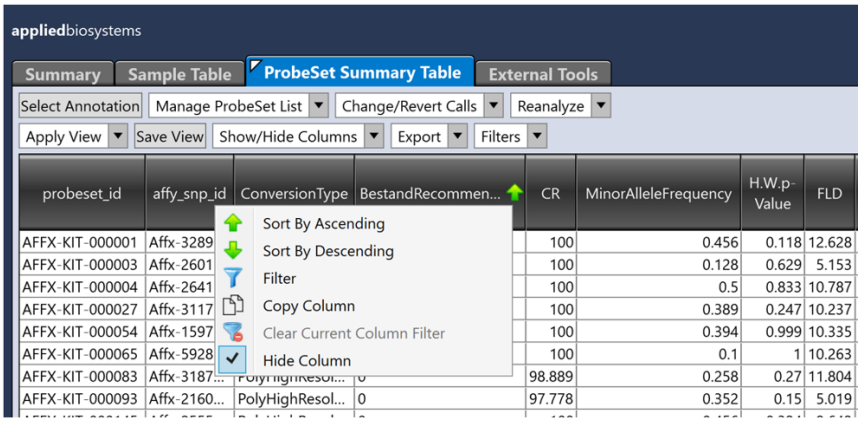


Figure 53 Filter conversion type column.

Run otv-caller or supplemental classification options

Axiom™ Analysis Suite includes the otv-caller function and the supplemental classification options as part of the software. To access these features, click the **Reanalyze** dropdown list in the SNP Summary Table (Figure 54).

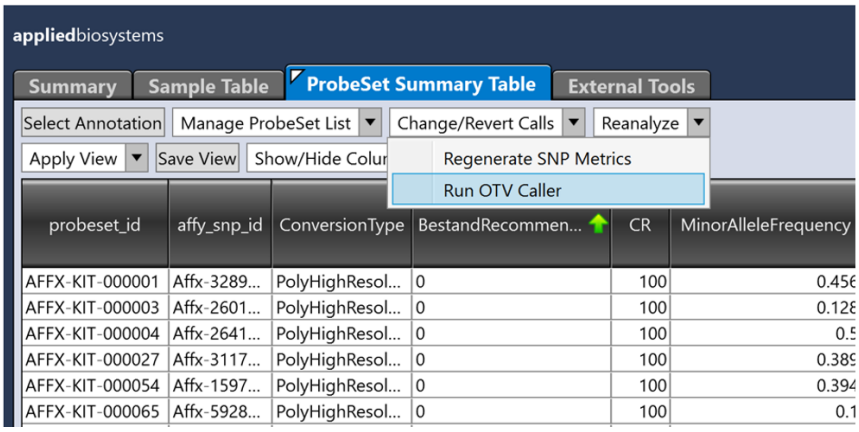


Figure 54 Reanalyze dropdown menu.

Note: Create a new recommended probeset list after running any addition analysis.

Export data from Axiom™ Analysis Suite

The genotype calls for passing samples and recommended SNPs can be exported from Axiom™ Analysis Suite for downstream analysis with third-party software in three different formats: TXT, PLINK, and VCF. To export the genotype calls, do the following:

1. Import the recommended SNP list (see “Step 8C: Create a recommended SNP list” on page 94).

2. In the Probeset Summary Table, click the **Export** dropdown list, then select **Export Genotyping Data.**

   The **Export Genotype Data** window appears (Figure 55).

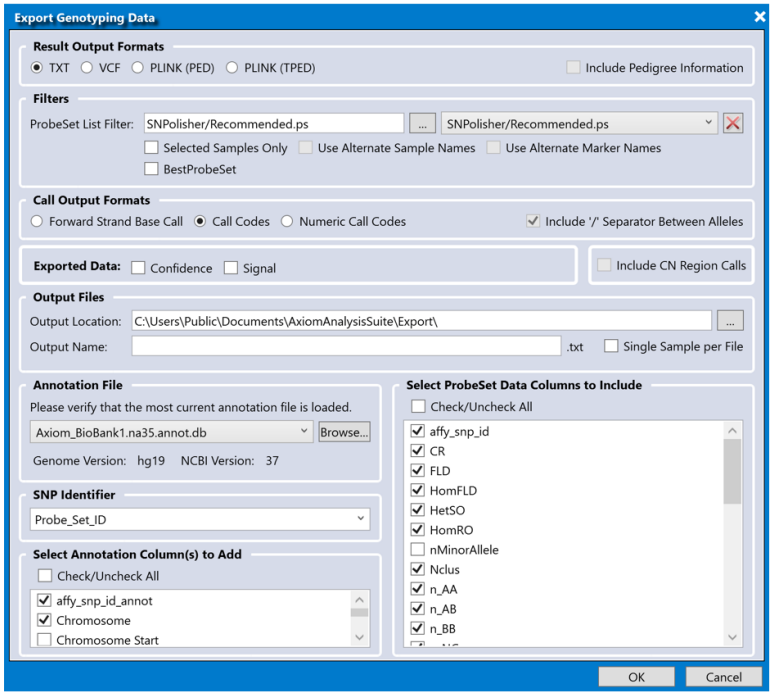


Figure 55 Export genotyping data window.

3. Select the Results Output Format: TXT, VCF, PLINK (PED), or PLINK (TPED).

4. Select the Call Output Format: Forward Strand Base Call, Call Codes, or Numeric Call Codes. Call Output Formats are not available for all Results Output Formats.

5. Click the dropdown list next to **SNP List Filter** , then select the created recommended SNP list.

6. Select the **Output Location** and **Output Name.**

7. Select **SNP Identifier.**

   Note: VCF file formats that are for use with Axiom™ HLA Analysis require the SNP Identifier to be **AFFY_SNP_ID.**

8. Include any desired annotation information by checking the boxes at the bottom of the menu. Additional annotation information is not available for all Results Output Formats.

9. Click **OK** to start the export.

8 Executing Best Practices steps with command line software

Execute best practice steps 1-7 with APT software

In this chapter, we provide instructions for executing steps 1-8 of the best practice analysis workflow (see Figure 7) using Applied Biosystems™ Array Power Tools (APT) combined with some simple scripts (to be written by the user).

The example scripts below are directly usable by Linux™ users. For readability, the Linux™ commands are broken into several lines with the backslash character: "\". The backslash character is not recognized by the Windows™ OS.

The APT commands can also be executed on a Windows™ computer.

1. Remove the backslashes ("\") and put the given command on one line.
2. Change the forward slash ("/") to a backslash ("\") when the input is a directory path.
3. Enter the command in the Windows™ command prompt window.

Best practices step 1: Group samples into batches

In preparation for step 2 of the best practice analysis workflow with APT (the 'Generate Sample DQC values' step), .CEL files corresponding to each batch must be collected into a file (we will refer to the files within each array batch as the 'cel_list') with the full path to each .CEL file in each row and with a header line = "cel_files". We will refer to this list as "cel_list1.txt". Below is a useful Linux™ one-liner for making cel_lists.

(echo cel_files; \ls -1 <DIRECTORY CONTAINING .CEL FILES>/*.CEL) > <OUTDIR>\cel_list1.txt

Best practices step 2: Generate the sample "DQC" values with APT

DQC values are produced by the program apt-geno-qc. Below is an example script for a Linux™ command-line shell. In this and other example scripts with APT programs, we assume that the analysis files were downloaded together in the same directory called <ANALYSIS_FILES_DIR>.

Example apt-geno-qc script for step 2 of the best practice analysis workflow

```
apt-geno-qc \
```

```
--analysis-files-path <ANALYSIS_FILES_DIR> \
```

```
--xml-file <ANALYSIS_FILES_DIR>/<axiom_array>.r<#>.apt-geno-qc.AxiomQC1.xml \
```

```
--cel-files <OUTDIR>/cel_list1.txt \
```

```
--out-file <OUTDIR>/apt-geno-qc.txt \
```

```
--log-file <OUTDIR>/apt-geno-qc.log
```

The generation of "cel_list1.txt" is discussed in step 1.

Best practices step 3: Conduct sample QC on DQC

Remove samples with a DQC value less than the default DQC threshold of 0.82. To execute this filter step, refer to the column "axiom_dishqc_DQC" in the file <OUTDIR>/apt-geno-qc.txt (produced by step 2 of the best practice analysis workflow). When executing the workflow with the APT system (Axiom™ Analysis Suite automates this step), you must write a script to remove .CELs from the <OUTDIR>/cel_list1.txt with DQC values that are < 0.82. We refer to filtered .CEL list from this step as cel_list2.txt.

Best practices step 4: Generate sample QC call rates with APT

Genotype calls are produced by the program apt-genotype-axiom using .xml files that end with AxiomGT1.apt2.xml. Below is an example script for a Linux™ command-line shell. In this and other example scripts with APT programs, we assume that the analysis files were downloaded together into the same directory called <ANALYSIS_FILES_DIR>.

Note: Do not use the program apt-probeset-genotype with Axiom™ arrays.

Example apt-genotype-axiom script for step 4 of the best practice analysis workflow using APT.

apt-genotype-axiom \

```
--log-file <OUTDIR>/apt-genotype-axiom.log \
```

```
--arg-file
<ANALYSIS_FILES_DIR>/<axiom_array>_96orMore_Step1.r<#>.apt-
genotype-axiom.AxiomGT1.apt2.xml \
```

```
--analysis-files-path <ANALYSIS_FILES_DIR> \
```

```
--out-dir <OUTDIR>/step1 \
```

```
--dual-channel-normalization true \
```

```
--cel-files <OUTDIR>/cel_list2.txt
```

The generation of "cel_list2.txt" is discussed in step 3.

Note: Select <axiom_array>_**LessThan96_Step1.r**<#>.apt-genotype-axiom.AxiomGT1.apt2.xml when available to perform QC genotyping with *SNP-specific models* if batch size is fewer than 96 samples. The LessThan96 xml is not an option for all Axiom™ arrays.

Best practices step 5: QC the samples based on QC call rate in APT

Remove samples with a QC call rate value less than the default threshold of 97%. To execute this filter step, refer to the column "call_rate" in the file "<OUTDIR>/step1/AxiomGT1.report.txt" produced by step 4. When executing the workflow with APT (Axiom™ Analysis Suite automates this step), you must write a script to remove .CELs from the <OUTDIR>/step1/cel_list2.txt whose call rate values are less than 97%. We refer to this .CEL list as *cel_list3.txt.* Note that the AxiomGT1.report.txt file has several header lines starting with #. The file can be read directly into a table (data.frame) using the R "read.table" function, which ignores lines starting with #.

Best practices step 6: QC the plates

In this section, we provide instructions for calculating the basic plate QC metrics and guidelines for identifying plates to remove from the analysis.

Note that you must write a script or use Excel™ to calculate the plate QC metrics.

- Group the .CEL files by plate, then for each plate:
  - Calculate plate pass rate.
    - $$\text{Plate Pass Rate} = \frac{\text{\# of samples passing DQC and 97\% QC call rate}}{\text{Total \# of samples on the plate}} \text{ X } 100$$

  - Calculate the average QC call rate of passing samples on the plate.
    - Remove the samples that failed the sample QC tests in steps 3 and 5.
    - Calculate the average QC call rate of the remaining samples for the given plate.
- Guidelines for passing plates in:
  - average QC call rate of passing samples > 98.5%

    If non-passing plates are identified in step 6, then all samples from these plates must also be removed in the process of creating *cel_list3.txt*

    Note: Plates with average QC call rate <98.5% are of low quality and can result in lower genotyping performance of samples on plates with high average QC call rates. Samples on such plates should be removed from the final genotyping run and considered for reprocessing.

Best practices step 7: Genotype passing samples and plates using AxiomGT1.Step2

For arrays with copy number-aware genotyping (CNAG) enabled, two extra steps should be executed as part of step 7 of the Best Practices Workflow. For CNAG enabled arrays, perform steps 7a and 7b below. If CNAG is not enabled for the array type, proceed to step 7c.

Step 7a: After all samples with a low sample QC call rate have been removed, summary intensity signals should be generated for all probesets.

```
apt-genotype-axiom \
```

```
--analysis-files-path <ANALYSIS_FILES_DIR> \
```

```
--arg-file <ANALYSIS_FILES_DIR>/<axiom_array>.r<#>.apt-
genotype-axiom.AxiomCN_PS1.apt2.xml \
```

```
--cel-files <DIR>/cel_list3.txt
```

```
--out-dir <OUTDIR>/summary \
```

```
--log-file <OUTDIR>/summary/apt2-axiom.log \
```

Where <DIR>/cel_list3.txt is the path to a text file with the header “cel_files” and each subsequent row a path to each CEL file, which should all pass the DQC threshold AND the QC call_rate threshold from the previous two steps.

The second CNAG step is to run the copy number analysis. *apt-copynumber-axiom-cnvmix* is the command to run the fixed region copy number analysis pipeline (step 7B). If a sample does not pass both *mapd-max* and *waviness-sd-max*, copy number results are not reported for that sample.

```
apt-copynumber-axiom-cnvmix \
```

```
--analysis-files-path <ANALYSIS_FILES_DIR> \
```

```
--arg-file <ANALYSIS_FILES_DIR>/<axiom_array>.r<#>.apt-
copynumber-axiom-cnvmix.AxiomCNVmix.apt2.xml \
```

```
--reference-file <ANALYSIS_FILES_DIR>/
<axiom_array>.r<#>.cn_models
```

```
--mapd-max 0.35 \
```

```
--waviness-sd-max 0.1 \
```

```
--summary-file <OUTDIR>/summary/AxiomGT1.summary.a5 \
```

```
--report-file <OUTDIR>/summary/AxiomGT1.report.txt \
```

```
--out-dir <OUTDIR>/cn \
```

```
--log-file <OUTDIR>/cn/apt-copynumber-axiom.log
```

Step 7C produces genotype calls for all SNPs and passing samples. Genotype calls are produced by the program apt-genotype-axiom. Below is an example script for a Linux™ command-line shell. In this and other example scripts with APT programs, we assume that the analysis files were downloaded together into the same directory called <ANALYSIS_FILES_DIR>.

Example apt-genotype-axiom script for step 7C.

```
apt-genotype-axiom \
```

```
--log-file <OUTDIR>/step2/apt-genotype-axiom.log \
```

```
--arg-file
<ANALYSIS_FILES_DIR>/<axiom_array>_96orMore_Step2.r<#>.apt-
genotype-axiom.AxiomGT1.apt2.xml \
```

```
--analysis-files-path <ANALYSIS_FILES_DIR> \
```

```
--out-dir <OUTDIR>/step2
```

```
--dual-channel-normalization true \
```

```
--summaries \
```

```
--genotyping-node:snp-posteriors-output true \
```

```
--batch-folder <OUTDIR>/suitefiles \
```

```
--cel-files <OUTDIR>/cel_list3.txt
```

If the array has multiallelic genotype calling enabled, then the following command must be added to the script:

```
--multi-genotyping-node:multi-posteriors-output true \
```

The generation of "cel_list3.txt" is discussed in steps 5 and 6.

Note that this example script for step 7 executes:

```
<axiom_array>_96orMore_Step2.r<#>.apt-genotype-
axiom.AxiomGT1.apt2.xml
```

whereas the example script for step 4 executes:

```
<axiom_array>_96orMore_Step1.r<#>.apt-genotype-
axiom.AxiomGT1.apt2.xml
```

The step 7 genotyping script includes options to write out several files to <OUTDIR>/step2.

The default files are:

- *AxiomGT1.calls.txt* contains the genotype calls for each probe set and sample.
- *AxiomGT1.confidences.txt* contains the confidence score (described in "What is a SNP cluster plot for AxiomGT1 genotypes?" on page 19) for each genotype call in the *Axiom™GT1.calls.txt* file.
- *AxiomGT1.report.txt* contains information about each sample.

Note: The output is per probeset, not per SNP. Although most SNPs are interogated by only one probeset, some SNPs are interrogated by more than one probeset. These SNPs will have more than one line of output in any file that generates output by probeset. The ps2snp file contains the mapping between a SNP and its probesets on an array.

The example script also includes options for additional output files. The posteriors and summary file must be created for use in Step 8.

- The *AxiomGT1.snp-posteriors.txt* file is enabled by *--write-models* option and includes the location and variance of the genotype clusters per probe set.
- The *AxiomGT1.summary.txt* file is enabled by *--summaries* option and includes the summarized intensity for the A and B allele of each probe set and sample.
- The --batch-folder option creates the files required to view the genotyping data and cluster plots in Axiom™ Analysis Suite.

Note: Select <axiom_array>_**LessThan96_Step2.r**<#>.apt-genotype-axiom.AxiomGT1.apt2.xml if available to perform genotyping with SNP-specific models if batch size is fewer than 96 samples.

Note: If the extra CNAG steps were run for the PMDA algorithm, the *apt-genotype-axiom* command should have these extra options enabled:

```
--copynumber-probeset-calls-file <OUTDIR>/cn/
AxiomCNVMix.cnpscalls.txt \
```

```
--allele-summaries true \
```

Best practices step 8A: Run ps-metrics

Ps-metrics uses three output files from Best Practices Step 7 (AxiomGT1.Step2 above) as inputs: the summary, calls, and posteriors file. The calls (AxiomGT1.calls.txt) and summary (AxiomGT1.summary.txt) files are required for all data sets. If a data set has biallelic probesets, the biallelic posteriors file (AxiomGT1.snp-posteriors.txt) must be input. If a data set has multiallelic probesets, the multiallelic posteriors file (AxiomGT1.snp-posteriors.multi.txt) must be input. Data sets with biallelic and multiallelic probesets must have both types of posteriors file as inputs.

Ps-metrics can be run on a subset of probesets or samples. A list of probesets, a list of samples, or both lists can be input to ps-metrics. See "Step 8A: Create SNP QC metrics" on page 30 for more information about ps-metrics.

A biallelic metrics file is produced for biallelic probesets, and a multiallelic metrics file is produced for multiallelic metrics. Ps-metrics produces two metrics files for data sets with both biallelic and multiallelic probesets.

To run ps-metrics on posterior and calls files in directory

<OUTDIR>/step2/ and generate output file metrics.txt:

```
ps-metrics --posterior-file .<OUTDIR>/step2/AxiomGT1.snp-
posteriors.txt \
```

```
--call-file <OUTDIR>/step2/AxiomGT1.calls.txt \
```

```
--metrics-file <OUTDIR>/SNPolisher/metrics.txt
```

If data set has multiallelic probesets, add the following input:

```
--multi-posterior-file <OUTDIR>/step2/AxiomGT1.snp-
posteriors.multi.txt \
```

```
--multi-metrics-file <OUTDIR>/SNPolisher/metrics.multi.txt \
```

The output from ps-metrics is a plain text file containing the SNP QC metrics. The default output file name is "metrics.txt". Each row is a SNP and each column is a QC metric. The output should look similar to Figure 56. This output file is one of the input files for other functions, so you must know the file's name and location on the computer.

probeset_id	CR	FLD	HomFLD	HetSO	HomRO	nMinorAllele	Nclus	n_AA	n_AB	n_BB	n_NC	hemizygous
AX-89778337	100	10.918	NA	0.48121	2.49899	54	2	230	54	0	0	0
AX-89778338	99.6479	5.9653	NA	0.38863	2.69981	89	2	194	89	0	1	0
AX-89778339	100	NA	NA	NA	1.67245	0	1	0	0	284	0	0
AX-89778340	99.6479	5.6528	NA	0.13211	1.39684	55	2	0	55	228	1	0
AX-89778341	96.4789	4.6887	NA	0.13635	1.1806	68	2	0	68	206	10	0
AX-89778342	98.2394	4.0849	8.53028	-0.0504	-0.2476	275	3	26	231	22	5	0
AX-89778343	100	11.243	24.09191	0.3905	1.08782	183	3	25	133	126	0	0
AX-89778344	99.6479	6.4202	NA	0.25466	2.06435	91	2	192	91	0	1	0
AX-89778345	100	5.7499	NA	0.429	1.54745	13	2	271	13	0	0	0
AX-89778346	97.1831	5.0842	NA	0.19844	1.94264	59	2	0	59	217	8	0
AX-89778347	99.6479	5.9784	NA	0.37153	2.35007	76	2	207	76	0	1	0
AX-89778348	100	8.7509	NA	0.65794	2.30562	1	2	283	1	0	0	0
AX-89778349	98.5915	4.8053	NA	0.22793	1.85181	119	2	0	119	161	4	0

Figure 56 An example of the output file from ps-metrics.

The first 13 rows of the output file from ps-metrics ("metrics.txt"), opened with the Excel™ application.

Best practices step 8B: Run ps-classification

After the SNP QC metrics have been generated, SNPs can be classified using ps-classification. Ps-classification has four required arguments and 56 optional arguments. The four required arguments are:

1. the name and location of the biallelic metrics file,
2. the name and location of the multiallelic metrics file,
3. the location of the preferred output directory, and
4. the species (or genome) type: human, non-human diploid, or polyploid.

Ps-classification will combine biallelic and multiallelic probesets into one performance file when both a biallelic and multiallelic metrics file are provided as inputs. It is not required to provide both metrics files: if a data set has only biallelic or multiallelic probesets or if the user chooses to only supply one type of metrics file, ps-classification will still run.

A ps2snp file is needed for arrays that include SNPs that are interrogated with more than one probe set. This file, *<axiom_array>.r<#>.ps2snp_map.ps*, should be provided with the Analysis Library Files for the array (Table 1). If this file has not been provided, contact your local Thermo Fisher Scientific Field Application Scientist or **thermofisher.com/support.**

Below is an APT command example for ps-classification which:

1. uses output from ps-metrics metrics results in metric.txt,
2. stores classification results in the folder <OUTDIR>/SNPolisher,
3. the genotype data are *human*
4. *ps2snp.ps* file = <ANALYSIS_FILES_DIR>/<axiom_array>.r<#>.ps2snp_map.ps. <ANALYSIS_FILES_DIR> means the full path to the Analysis Library file directory.

   If the array has multiallelic genotype calling enabled, <axiom_array>.r<#> .ps2multisnp_map.ps file must be used.

```
ps-classification\
```

```
--species-type human \
```

```
--metrics-file <OUTDIR>/SNPolisher/metrics.txt \
```

```
--output-dir <OUTDIR>/SNPolisher \
```

```
--ps2snpfile <ANALYSIS_FILES_DIR/
<axiom_array>.r<#>.ps2snp_map.ps
```

If data set has multiallelic probesets, also provide the following:

```
--multi-metrics-file <OUTDIR>/SNPolisher/metrics.multi.txt \
```

Eight of the optional arguments are basic classification thresholds for the QC metrics. If only a species type is given, ps-classification will use the default thresholds for that genome type (see Table 4). SNPs classified as PolyHighResolution must have SNP QC values that pass all these thresholds. See Appendix B, “Complete set of classification thresholds used by ps-classification” for more information on the full set of arguments for ps-classification.

There is one logical indicator: hom-ro indicates if HomRO thresholds should be used (default is TRUE). Polyploid genotypes do not use either of the HomRO thresholds so the hom-ro flag should be FALSE for polyploid species.

Priority-order is used when performing best probeset selection: the best probeset is selected according to the priority order of probeset conversion types. These are based on the default category order: PolyHighResolution, NoMinorHom, MonoHighResolution, OTV, UnexpectedGenotypeFreq, CallRateBelowThreshold, Other, and OtherMA. The priority-order argument allows the user to change the order of categories when determining which probesets are selected as the best probeset for a SNP. All possible conversion types must appear in priority-order, and the user specifies the ranked order that the conversion types fall into when selecting between probesets for a SNP.

Ps-classification accepts a list of probesets, and categorizes the SNPs in this file only. The first row of this file should always be "probeset_id".

See “Step 8B: Classify SNPs using QC metrics” on page 31 for more information on ps-classification, and Appendix B, “Complete set of classification thresholds used by ps-classification” for more information on the full set of inputs, arguments, and thresholds for ps-classification.

Visualize SNP cluster plots with Ps_Visualization

The Ps_Visualization function is used to produce SNP cluster plots. The list of probesets for plotting can be generated using the results from ps-metrics, ps-classification, and otv-caller, or from other probeset properties that are of interest. Plots include the posterior (default) and prior (optional) distributions. Reference genotypes may also be displayed in adjoining plots. The cluster plots can help quality check SNPs and diagnose underlying genotyping problems. The genotype cluster colors can be changed by the user, and selected samples can be highlighted.

Ps_Visualization is available in the SNPolisher™ package and is not available in APT or AxAS. For more details on installing R and the various options with Ps_Visualization, see the *SNPolisher™ Package User Guide* (Pub. No. MAN0017790).

Ps_Visualization takes four required arguments and many optional arguments. The four required arguments are the probeset id file (pidFile), which contains the probeset_ids for plotting; the name of the output file, with the full output path if the user wants a location different from the default location of the working directory;

the name and location of the summary file; and the name and location of the calls file. The first line of the pidFile should always be "probeset_id".

Ps_Visualization uses a temporary directory (temp.dir) to store intermediate data files created from the summary and calls files. If no directory is given, the default that is used is "Temp/". The temporary directory is deleted by default after plotting is completed. If the user knows that a probeset will be plotted more than once, the temporary directory can be kept (keep.temp.dir) and the intermediate files can be read from the temporary directory instead of remade (use.temp.dir). Using these options with repeated plotting will speed up Ps_Visualization.

Ps_Visualization has many optional arguments and settings. See the *SNPolisher™ Package User Guide* (Pub. No. MAN0017790) for a complete description of all options. Several options are shown in Figure 57, Figure 58, and Figure 59.

Figure 57 shows a probeset that has been plotted with and without reference genotypes. Subfigure 1 shows the same set of samples colored by called genotypes and by reference genotypes. The reference plot only shows samples with reference genotypes - any sample with a reference value of NoCall has not been plotted (ref.NoCalls). Subfigure 2 shows how reference plots are created for probesets without reference genotypes: all samples are plotted and assigned a value of NoCall. Reference plots should not be included when there isn't any reference data (plot.ref=FALSE).

Figure 58 shows a probeset where a list of samples have been highlighted in one color, in contrast to the cluster colors. Subfigure 1 shows the highlighted samples in green with the default colors for the AA, AB, and BB clusters. Subfigure 2 shows the highlighted samples in red and the clusters plotted in other colors: AA is cyan, AB is dark green, and BB is purple.

Figure 59 shows more complex examples of sample highlighting and assigning colors to clusters. Subfigure 1 shows samples that have been highlighted in 2 different colors (green and turquoise) with the default cluster colors. Subfigure 2 shows the same highlighted samples but all cluster colors have been set to light grey so that the highlighted samples are very clear.

The assigned shapes for a cluster cannot be changed.

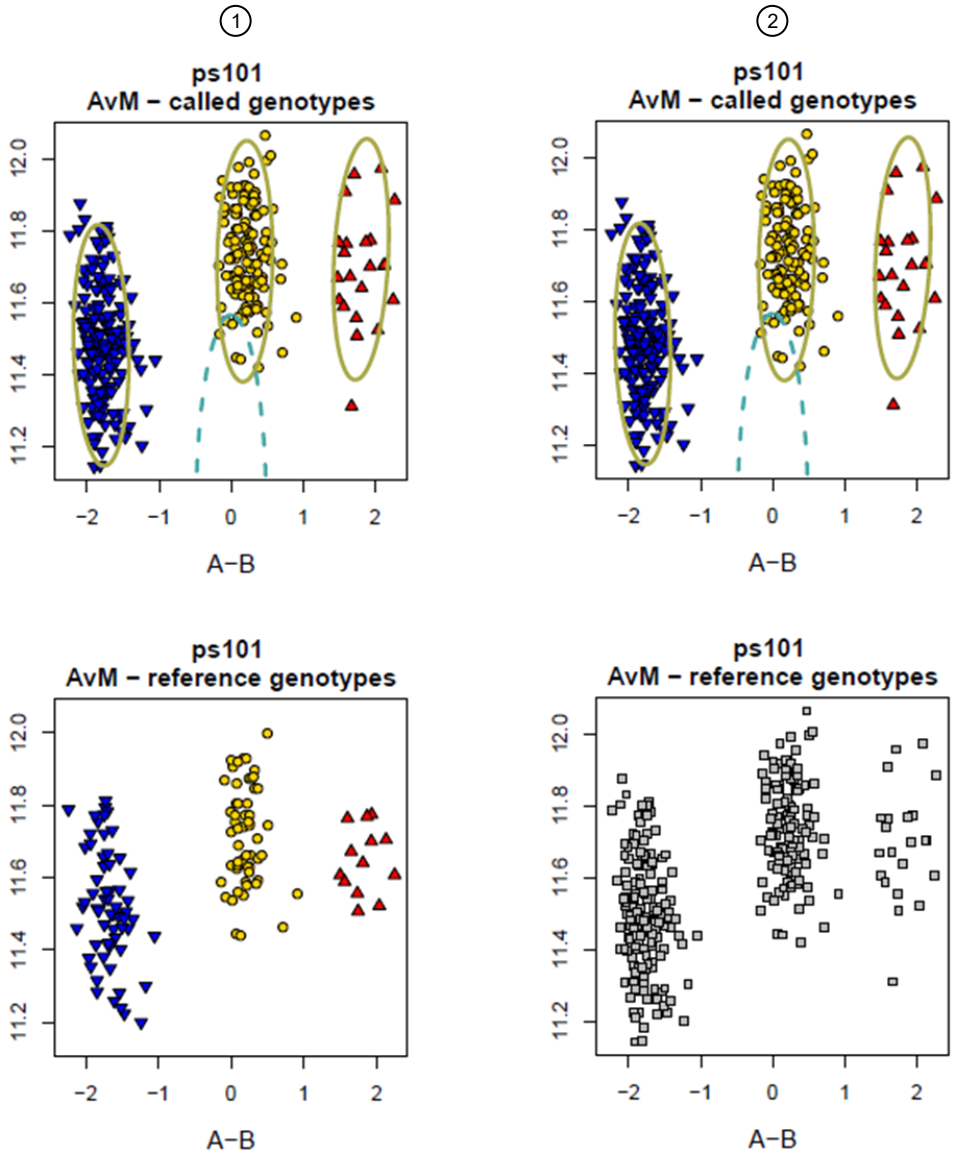


Figure 57 One SNP plotted with and without reference genotype.

(1) A SNP plotted with reference genotypes.
(2) The same SNP plotted without reference genotypes.

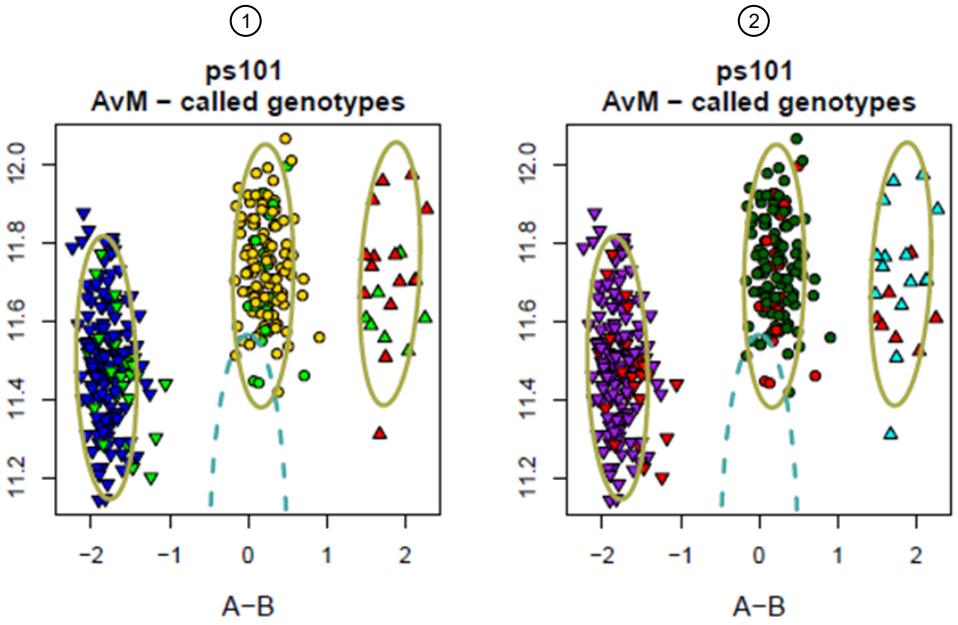


Figure 58 One SNP plotted with color and sample highlight options.

(1) One SNP plotted with default color options and highlighted samples.
(2) The same SNP plotted with different colors and the same highlighted samples.

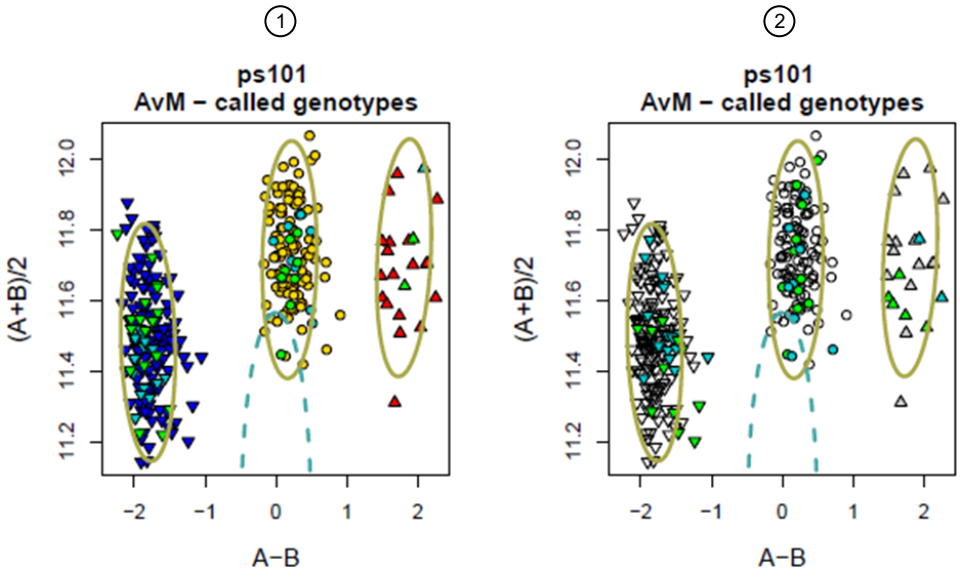


Figure 59 One SNP with samples highlighted in two colors.

(1) One SNP plotted with default color options and samples that are highlighted in green and turquoise.
(2) The same SNP with all non highlighted samples plotted in light gray.

Gender-separated plotting

When a special SNPs file and a report file are provided, Ps_Visualization will produce gender-separated plots for non-PAR X, Y, Z, and W SNPs. Non-PAR X and Z SNPs have two plots, one with only male samples and one with only female samples. Y and W SNPs have two plots, one with only male (or female for W SNPs) samples and one with all samples. Samples with unknown gender are plotted as a black asterisk and appear in all plots.

Plotting the genders separately can make it easier to visually inspect a SNP. For a non-PAR X SNP, the female samples can be diploid, haploid, or ZeroCN. The male samples should be haploid. This means that the male samples are expected to be clustered below the diploid female samples, and should be closer to X = 0 in the transformed data space. For a Y SNP, the male samples should be haploid and the female samples should be ZeroCN or NC. This means that the female samples are expected to be clustered below the male samples and to sit approximately at X = 0 in the transformed data space. The same patterns are expected in Z and W SNPs with the genders reversed.

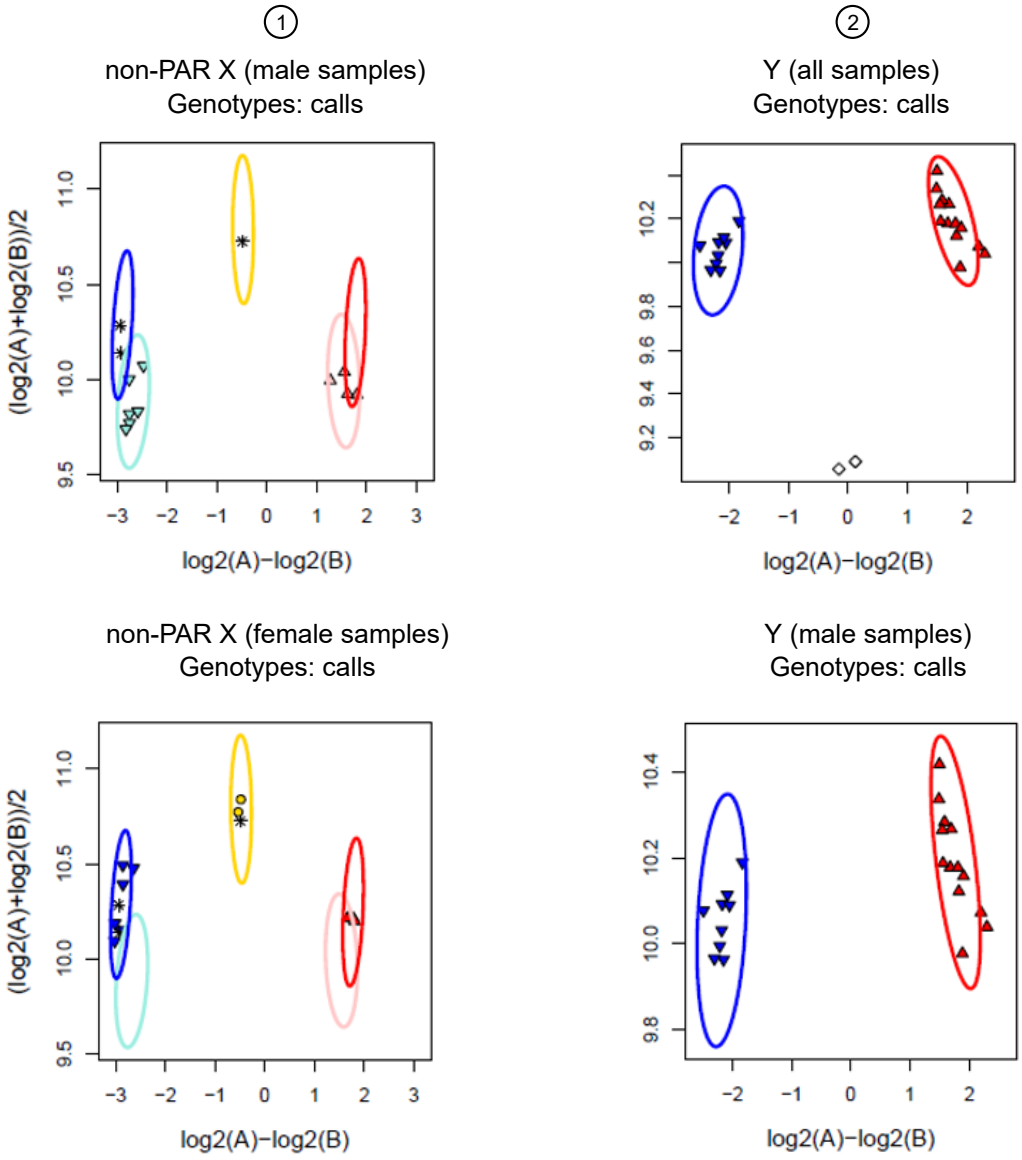


Figure 60 Plots of a non-PAR X SNP and a Y SNP.

① Left: Plot of a non-PAR X SNP: The male samples plot shows that the male samples are all haploid as expected and there are several diploid samples with unknown gender, and the female samples plot shows that the females are above the males.

② Right: Plot of a Y SNP: the plot with all samples shows that the female samples are below the males and centered at 0, and the male samples plot shows that the male samples are well clustered.

Multiallelic plotting

SNPs with 3 or more unique alleles in their assigned genotyped calls are plotted in the multiallelic plotting system: the 3 most common alleles are used as the X, Y, and Z axes. Color and shape assignments are dynamically allocated for each SNP. A legend is plotted with each multiallelic plot so the user knows which genotype call is represented by which color-shape combination. The standard default color-shape combinations from biallelic plots are used for the calls which appear in the biallelic plotting system: AA (red triangle), AB (yellow circle), BB (blue upside down triangle), OTV (cyan diamond), and NoCall (grey square). Users cannot change the color-shape assignments in multiallelic plots. Multiallelic SNPs are plotted in $\log_2$ space rather than Size vs Contrast space: $\log_2(X)$ vs $\log_2(Y)$ vs $\log_2(Z)$.

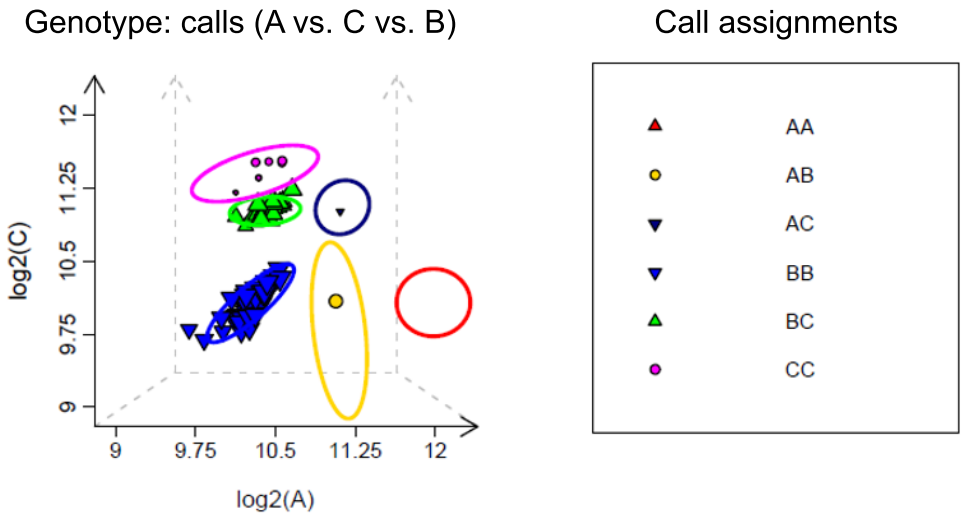


Figure 61 A multiallelic SNP plotted in the 3D system with accompanying legend.

If a SNP was designed to be multiallelic but the assigned genotype calls only have 1 or 2 unique alleles, then the SNP is plotted in a 2D biallelic plot that is in $\log_2$ space. The most common allele is on the X axis and the second allele is on the Y axis. If only one allele is present in the genotype calls, then it is on the Y axis if it is not A and A is on the X axis, or it is on the X axis if it is A and B is on the Y axis.

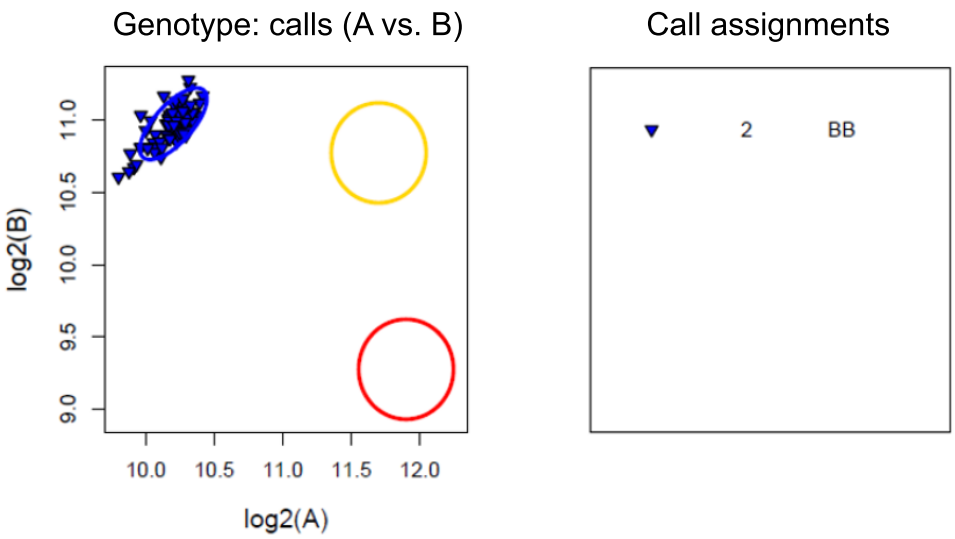


Figure 62 A multiallelic SNP with only 1 unique allele in the genotyping calls with accompanying legend.

While the default plotting method for multiallelic SNPs is 3D plotting, users can select the multiallelic pairwise option instead. This option plots all biallelic pairwise combinations of the alleles that appear in the genotype calls in the biallelic $\log_2$ plots, as well as a plot with all samples in $\log_2$(A) vs $\log_2$(B) space. For example, if alleles A, C, and D appear in the genotype calls for a SNP, then four plots are made: all samples, $\log_2$(A) vs $\log_2$(C) with only samples that have A and/or C alleles in their calls, $\log_2$(A) vs $\log_2$(D) with only samples that have A and/or D alleles in their calls, and $\log_2$(C) vs $\log_2$(D) with only samples that have C and/or D alleles in their calls. Figure 63 displays the biallelic pairwise combination plots for the SNP plotted in 3D in Figure 62 above.

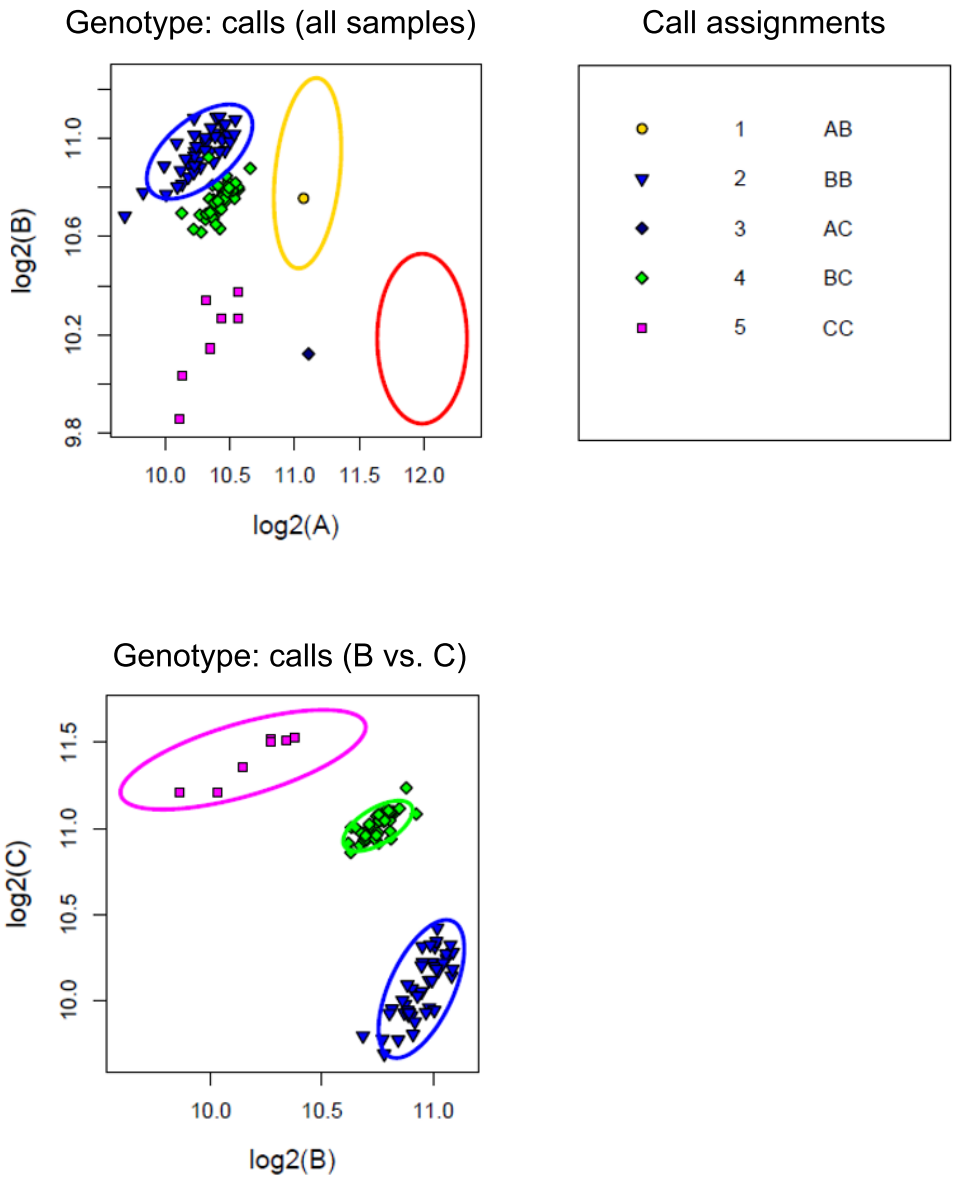


Figure 63 A biallelic pairwise combination plots of a multiallelic SNP.

Color-shape assignments are consistent when a SNP has more than one plot: biallelic pairwise plots, gender-separated plots for special SNPs, batch plots, and genotype calls and references plots. One legend is supplied for the full set of plots for a SNP.

Batch plotting

When there is a large number of samples, genotyping may be split into batches. This will result in multiple calls, summary, and posteriors files being produced for the same SNPs with different samples in each batch. Batch plotting can be used to investigate the behavior of samples across batches for a SNP, and visual inspection may be used to detect problems with a batch such as cluster flips or splits or unexpected allele frequencies. Ps_Visualization produces multiple batch plots per SNP when batch files are supplied as input.

Ps_Visualization handles the layout of batch plots automatically. A new page is started in a file at the beginning of each set of batch plots for every SNP. All plots are produced per batch and then the next batch is started. For example, if a user is plotting batches and requests intensity, genotype, and reference plots, then Ps_Visualization will produce an intensity plot, an intensity reference plot, a genotype plot, and a genotype reference plot before starting to plot the next batch. This ensures that the same plots for each batch are plotted on the same row for easy comparison.

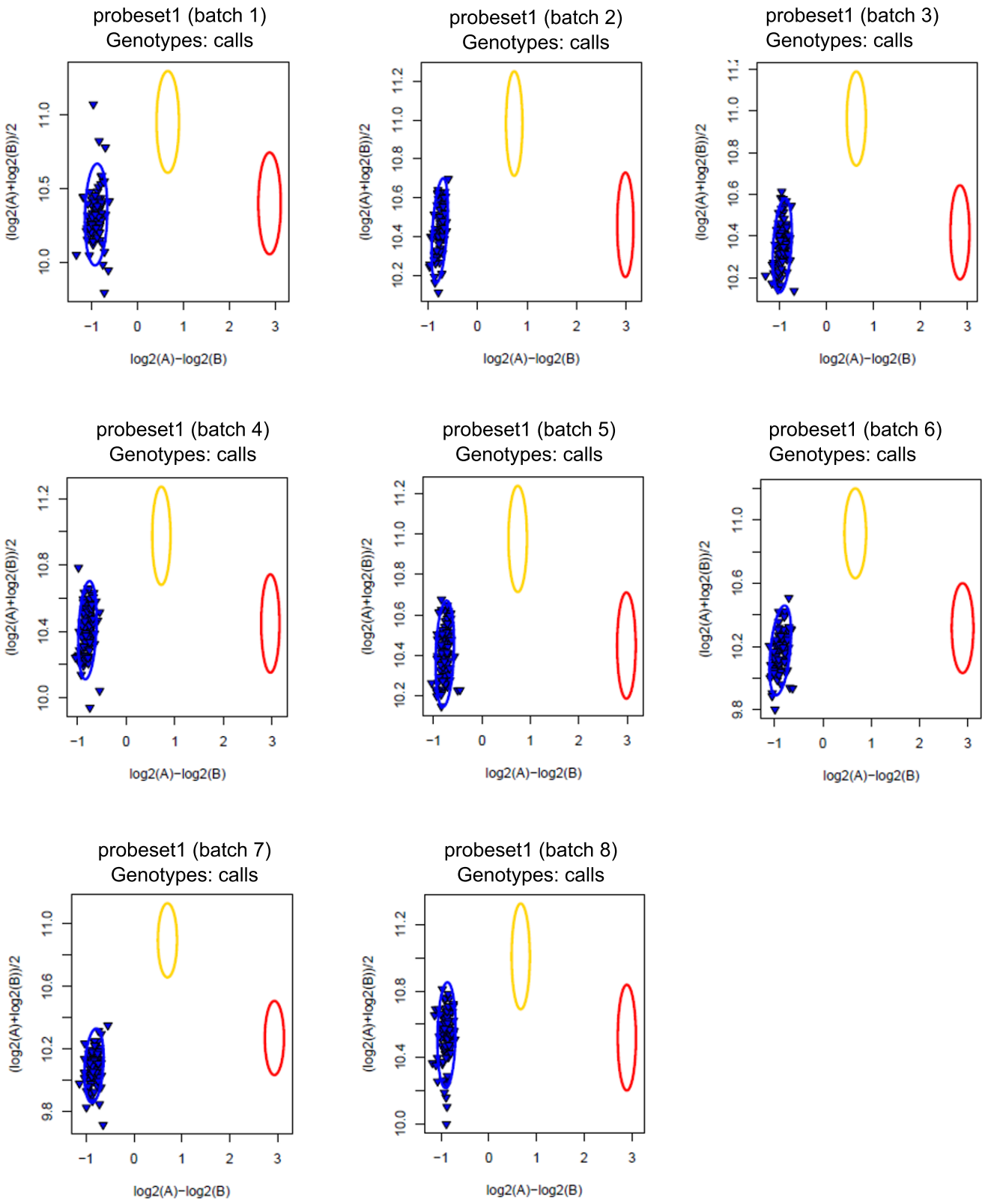


Figure 64 Batch plots for one biallelic SNP with 8 batches of samples.

Multiallelic SNPs with more than one batch are plotted similarly to biallelic SNPs. The legend displays the assigned colors and shapes for every call that appears in the batch plots, and all colors and shapes are assigned to the same genotype calls across all batches.

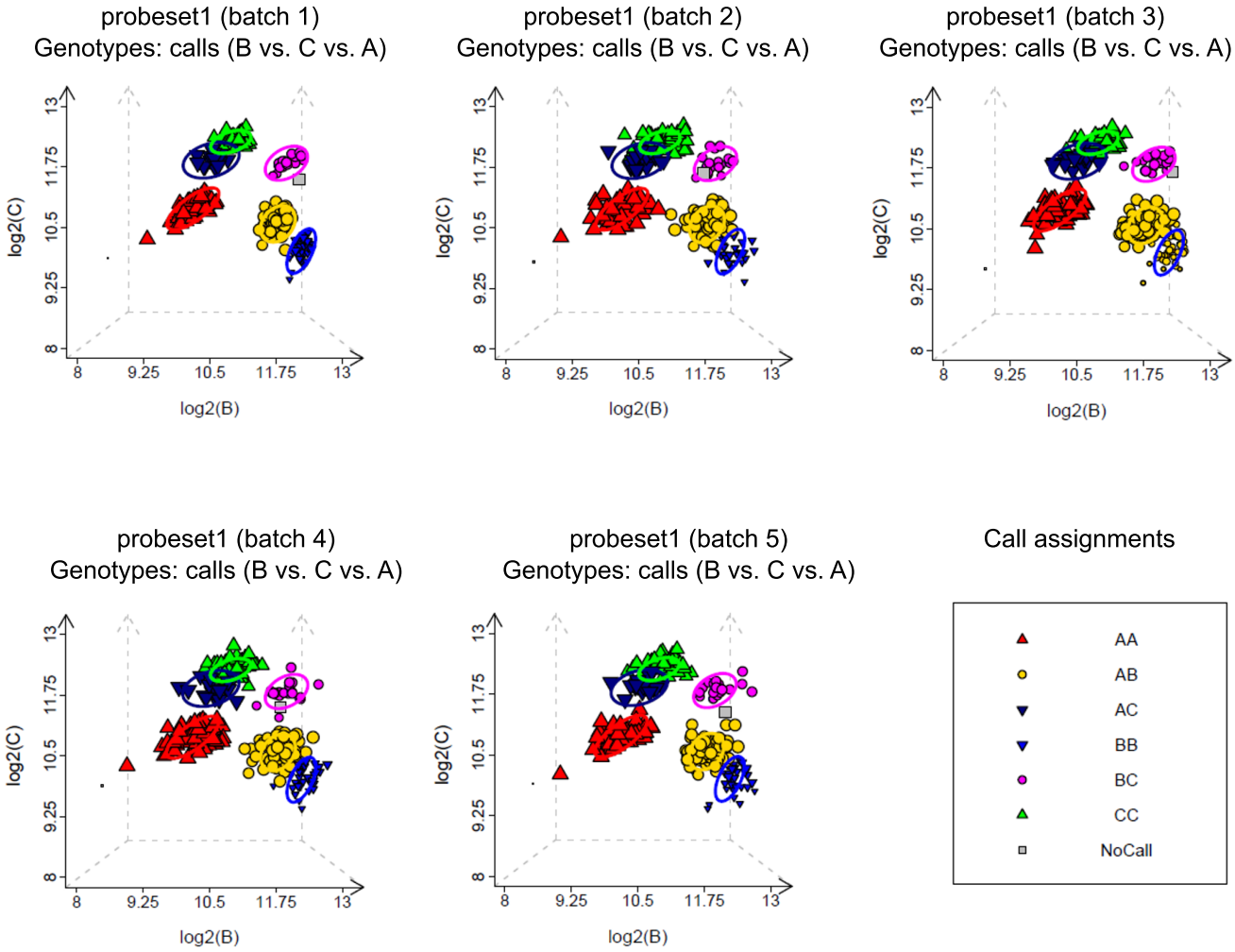


Figure 65 A multiallelic SNP with 5 batches.

All shape and color assignments are consistent across the batches, and are displayed in the single legend that accompanies the entire set of batch plots for the SNP.

Batch plotting can be used to easily identify if there are any batches with genotyping results that are considerably different from the majority of batches. The multiallelic SNP in Figure 65 has similar cluster locations across all 5 batches, but the genotype call assignments in batch 3 are different than in the other 4 batches. Plotting all batches genotyped for a marker or SNP is a useful diagnostic tool when samples have been split up between batches.

A Complete Set of SNP QC Metrics Produced by ps-metrics

Base workflow metrics

There are 17 basic QC metrics. CR, FLD, HomFLD, HetSO, and HomRO were described in Step 8B. Minor allele count (nMinorAllele), minor allele frequency (MAF), number of clusters (Nclus), number of AA calls (n_AA), number of AB calls (n_AB), number of BB calls (n_BB), and number of no calls (n_NC) are self-explanatory. Note that MAF is calculated using diploid and haploid calls if copy-number genotyping was used. The Hardy-Weinberg test statistics and p-values were described in “Hardy-Weinberg p-value” on page 79. Hemizygous is a logical operator that indicates if a SNP appears to be hemizygous (e.g., Y chromosome in human male). Hemizygosity is indicated with a 1; otherwise 0 is returned.

Minimum Mahalanobis Distance (MMD): The minimum Mahalanobis distance is the minimum of the distance between all AA points and all BB points to the plane created by the AB cluster, and can be thought of as the minimum distance from all AA and BB points to the mean of the AB distribution. Probesets that do not have samples in the AA, AB, and BB clusters have an MMD value of 0. MMD is named for P. C. Mahalanobis. See Figure 66 for a visual example of the MMD.

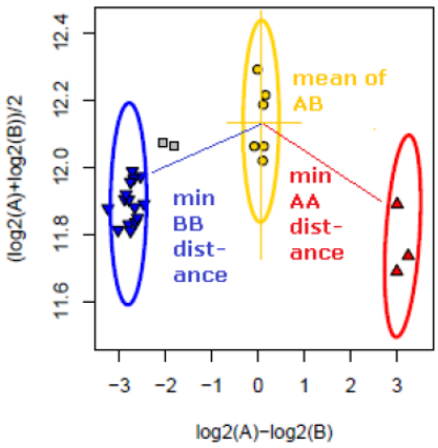


Figure 66 The minimum Euclidean distance from the AA and BB points to the mean of the AB cluster.

In this example, the MMD is the smaller of the two minimum distances (BB).

Genotype Frequency P-values: If a genotype frequency file is provided, then p-values are calculated from a test of expected genotype frequencies. The genotype frequency file holds the expected frequency of a cluster for a SNP; this value is tested against the observed frequency of the cluster to see if the observed frequency is

much larger than the expected frequency. The p-value is calculated using Hoeffding's inequality, which gives the probability that the number of samples observed in a cluster is different from the expected number of samples in that cluster. The calculation is only performed on clusters where there are more samples observed than expected. This restriction keeps p-values from being calculated on clusters where the number of observed samples is less than or equal to the expected number of samples.

Although the one-sided and two-sided p-values are different, this metric is used to identify clusters with very small p-values, where the difference between one-sided and two-sided p-values are greatly reduced. The formula is p = exp(-2*(observed - expected)^(2/n)).

Figure 67 shows the differences between the expected and observed genotype frequencies for a SNP where there is a large difference. The reference genotypes show an expected frequency of 100% for AA and 0% for AB and BB, but the observed genotypes show the opposite. This SNP has a very small p-value for the test between expected and observed genotype frequency.

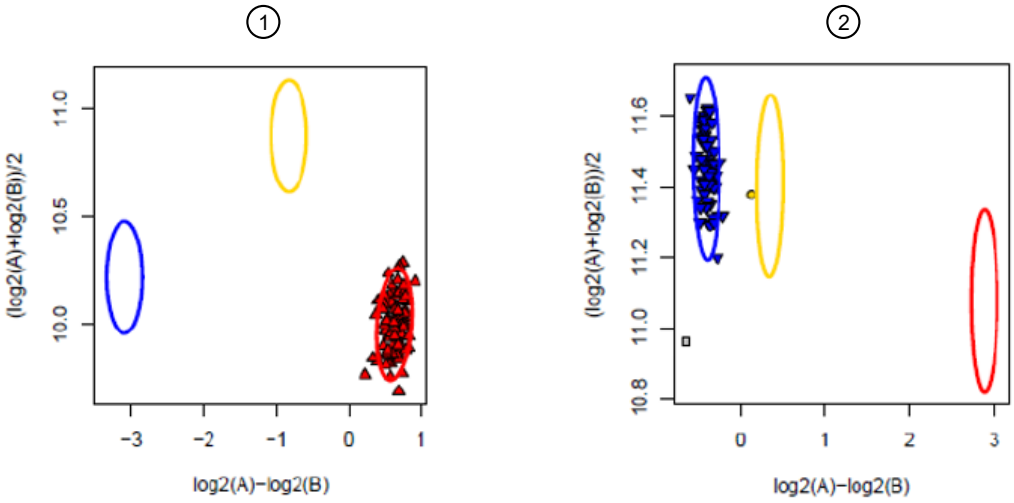


Figure 67 Expected versus observed genotype frequencies.

1. The expected genotype frequency based on this reference plot is 100% for AA and 0% for BB and BA.
2. The called genotypes show an observed frequency of 0% for AA and AB and 0% for BB.

This metric holds all p-values calculated for a probeset. The metric has the cluster name and then the p-value, separated by a colon. Multiple clusters for a probeset are handled by separating the values with a semicolon (e.g., AA:0.15;AB:0.25). Any clusters without a provided expected frequency value or where the observed value is smaller than the expected value produce NA for the p-value.

P-values will not be calculated unless the minimum number of samples with an assigned genotype per probeset has been met (i.e. has a genotype call other than NoCall). The larger the number of samples used in the calculations, the more reliable the p-values will be. We strongly suggest that the genotype frequency p-values are not calculated on SNPs with fewer than 20 genotyped samples. For gender-separated metrics, the minimum number of samples will be checked for each gender. The p-values are calculated separately, so if a probeset has more than enough

female samples but not enough male samples, a p-value will be calculated for the females-only row in the metrics file.

Calculations will be performed for any cluster that has an observed frequency, an expected frequency, and has enough genotyped samples. This includes haploid clusters and multiallelic clusters.

The genotype frequency file should contain three columns named “probeset_id”, “cluster”, and “frequency”. Each row of the file should list the SNP, the cluster of that SNP, and the matching frequency. If there is more than one frequency supplied per SNP, then that SNP will have multiple rows. All genotype frequencies must be between 0 and 1. If genotype frequencies are supplied for all clusters of a SNP, the sum of the genotype frequencies must be greater than or equal to 1.

Cluster means and variances

Two means are given in the posteriors file for each cluster: the center of the posterior ellipse on the X axis (meanX) and the center of the posterior ellipse on the Y axis (meanY). All clusters except copy number zero (CN0) have posterior values (AA.meanX, AA.meanY, AB.meanX, AB.meanY, BB.meanX, BB.meanY, A.meanX, A.meanY, B.meanX, B.meanY). CN0 does not have a posterior and does not appear in the posteriors file, so the summary file is used to calculate the mean X and Y locations of the CN0 samples (CN0.meanX, CN0.meanY). Likewise, NoCalls do not have a posterior and do not appear in the posteriors file and the mean locations are calculated from the calls and summary files (NC.meanX, NC.meanY).

Cluster variances may also be recalculated in addition to the means. Variances are calculated directly from the summary and calls data instead of being taken from the posteriors file when gender-separated metrics are calculated for special SNPs and with the supplemental and SSP workflows. The required minimum value for the variances is 0:03 and any calculated variance that is less than 0:03 is set to 0:03.

Gender separated metrics

If a special SNPs file and a report file are provided, some metrics are calculated directly from the summary and calls data on the gender-separated samples for non-PAR X and Z SNPs and on the male samples only for Y SNPs and on the female samples on for W SNPs. The metrics file will have the endings “_male”, “_female”, and “_all” appending to the probeset names to indicate which gender of samples was used in calculating the metrics. Non-PAR X, Y, Z, and W SNPs have “_male” or “_female” added for each gender, and have two rows in the metrics file. MT and CP SNPs have “_all” added to each probeset name and have one row in the metrics file.

Y and W probesets

Y and W SNPs should not have heterozygous genotype call assignments, and therefore HetSO and FLD have a value of NA for all Y and W SNPs. Male samples should have haploid genotype calls on Y SNPs and female samples should have haploid genotype calls on W SNPs.

The historic convention for genotype call assignments on the Axiom platform was to assign only diploid calls for all SNPs. This resulted in the male samples being assigned the homozygous diploid call that corresponded to the actual haploid call, e.g. a true genotype call of A was assigned a value of AA. When copy-aware genotype calls are used, the true haploid genotype calls are assigned. When copy-number aware genotype calls are not used, the convention of assigning diploid calls to male samples on Y SNPs and female samples on W SNPs is used.

Female samples on Y SNPs and male samples on W SNPs are expected to be either NC (without copy-number aware genotypes) or CN0 (with copy-number aware genotypes). Users should check if the samples have non-zero values for the number of assigned diploid and/or haploid genotype calls. Genotype call assignments to female samples on Y SNPs and male samples on W SNPs may indicate a problem with genotype call assignments. The location of the female samples on Y SNPs and male samples on W SNPs are used in ps-classification when categorizing the performance of the Y and W SNPs. CN0 clusters do not appear in the posteriors file and so the summary file is used to calculate the mean X and Y locations of the copy number zero samples. Likewise, NoCalls do not have a cluster assignment and do not appear in the posteriors files. NC.meanX and NC.meanY are calculated from the summary and calls files.

Figure 68 shows the expected sample behavior by gender for a Y SNP. The male samples should be haploid calls and the female samples should either be NoCalls or CN0. The female samples should sit below the male samples and close to 0 in the X dimension.

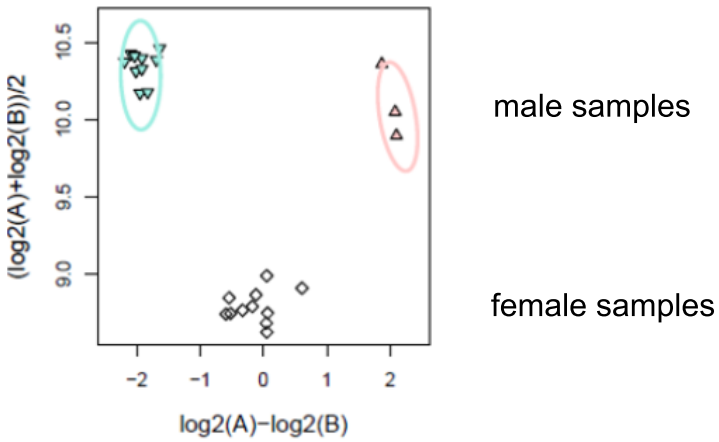


Figure 68 Plot of a Y SNP.

In a Y SNP, the male samples are haploid and the female samples are NoCall or ZeroCN. The males should sit above the females, and the females should sit above 0 in the X dimension. Metrics are calculated separately on the male and female samples in order to evaluate their locations in ps-classification.

These metrics are calculated on female samples for Y SNPs and male samples for W SNPs, including extra metrics when copy-number aware genotype calls are present:

Table 11 Metrics for female samples for Y SNPs and male samples for W SNPs.

Metrics for female samples for Y SNPs and male samples for W SNPs	
n_AA (should be 0)	n_NC
n_AB (should be 0)	NC.meanX
n_BB (should be 0)	NC.meanY
Additional copy-number metrics	
n_A (should be 0)	n_CN0
n_B (should be 0)	CN0.meanX
	CN0.meanY

Male samples on Y SNPs and female samples on W SNPs are expected to have diploid calls without copy-number aware genotyping and haploid calls with copy-number aware genotyping. Most of the metrics calculated for Y SNPs use the male samples.

Variances are only calculated for samples with more than *clustermin* samples; otherwise the default value of 0.03 is used. If copy-number aware genotype calls are present, additional metrics are also calculated:

Table 12 Metrics for male samples for Y SNPs and female samples for W SNPs.

Metrics for male samples for Y SNPs and female samples for W SNPs	
CR	BB.meanX
Nclus	BB.meanY
nMinorallele	BB.varX
n_AA	BB.varY
n_AB (should be 0)	AA.meanX
n_BB	AA.meanY
n_NC	AA.varX
HomRO	AA.varY
HomFLD	
GenotypeFreqPvals	
Additional copy-number metrics	
n_A	A.meanX

Table 12 Metrics for male samples for Y SNPs and female samples for W SNPs. *(continued)*

Metrics for male samples for Y SNPs and female samples for W SNPs	
n_B	A.meanY
n_CN0	A.varX
B.meanX	CN0.meanX
B.meanY	CN0.meanY
B.varX	HomRO_hap
	HomFLD_hap

Non-PAR X SNPs and Z SNPs

Almost all metrics are calculated on gender-separated samples for non-PAR X SNPs and Z SNPs. Due to the historical convention of assigning diploid calls to true haploid calls, female and male samples are expected to have diploid calls without copy-number aware genotyping. If copy-number aware genotype calls are assigned, female samples on non-PAR X SNPs and male samples on Z SNPs should have diploid and haploid calls and male samples on non-PAR X SNPs and female samples on Z SNPs should only have haploid calls.

Because both male and female samples are expected to have genotype calls, metrics are calculated separately for both genders with the exception of any metric involving heterozygous calls. Male samples on non-PAR X SNPs and female samples on Z SNPs are truly haploid and should not produce a heterozygous call when the call assignments are diploid-only. If a male sample has a true haploid genotype value of A, the matching assigned diploid call is AA. If the true haploid genotype is B, then the assigned call is BB. There is no haploid call that matches up to a diploid call of AB. HetSO and FLD are calculated on female samples only on non-PAR X SNPs and male samples only on Z SNPs and are set to missing for male samples on non-PAR X SNPs and female samples on Z SNPs.

Figure 69 shows the expected sample behavior by gender for a non-PAR X SNP. The female samples should be diploid and the male samples should be haploid. The male samples should sit below the female samples and be somewhat closer to 0 in the X dimension than the female samples.

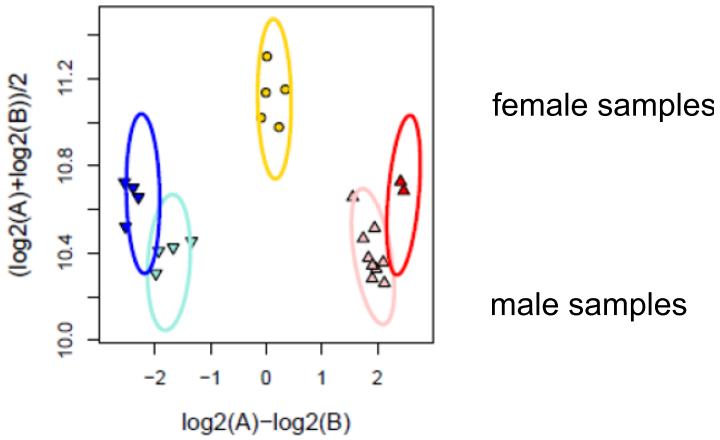


Figure 69 Non-PAR X SNP.

In a non-PAR X SNP, the male samples are haploid and the female samples are diploid. The females should sit above the males, and the males should be closer in towards 0 in the X dimension. Metrics are calculated separately on the male and female samples in order to evaluate their locations in ps-classification.

The metrics are calculated separately on the genders, including extra metrics when copy-number aware genotype calls are present. Variances are only calculated for samples with more than clustermin samples; otherwise the default value of 0.03 is used.

Table 13 Gender-separated metrics for non-PAR X SNPs and Z SNPs.

Gender-separated metrics for non-PAR X SNPs and Z SNPs	
CR	BB.meanX
Nclus	BB.meanY
nMinorAllele	BB.varX
n_AA	BB.varY
n_AB	AB.meanX (females/males only)
n_BB	AB.meanY (females/males only)
n_NC	AB.varX (females/males only)
HomRO	AB.varY (females/males only)
HetSO (females/males only)	AA.meanX
FLD (females/males only)	AA.meanY
GenotypeFreqPvals	AA.varX
	AA.varY
Additional copy-number metrics	
n_A	A.meanX

Table 13 Gender-separated metrics for non-PAR X SNPs and Z SNPs. *(continued)*

Gender-separated metrics for non-PAR X SNPs and Z SNPs	
n_B	A.meanY
N_CN0	A.varX
B.meanX	CN0.meanX
B.meanY	CN0.meanY
B.varX	HomRO_hap
	HomRO_hap (females/males only)

Additional copy-number aware metrics

Ps-metrics automatically calculates haploid and copy number zero (CN0) metrics if copy-number aware genotype calls are detected in the calls file. HomFLD_hap is a version of HomFLD computed only on haploid clusters (which are homozygous by definition). HomFLD_hap is undefined for SNPs without two haploid clusters. HomRO_hap is a version of HomRO computed only on haploid clusters, and is the location on the x-axis of the populated haploid cluster center that is closest to zero.

Table 14 Copy-number aware metrics.

Metric	
HomFLD_hap	A.varX
HomRO_hap	A.varY
n_A	B.meanX
n_B	B.meanY
n_CN0	B.varX
A.meanX	B.varY
A.meanY	CN0.meanX
	CN0.meanY

Multiallelic workflow

Two multiallelic metrics are designed to compare the signal and background intensities, and several other are derived from biallelic metrics.

Signal and background metrics

An allele can have different copy numbers per probeset, and the value depend on which samples are genotyped. Copy number values can be 0, 1, or 2.

Figure 70 shows the different copy number values for alleles A, B, C, and D for a SNP with clusters AA, AB, AC ,BB, BC, and CC. Although allele D is an alternate allele option for the SNP, none of the samples had genotypes with allele D. Allele A has copy number 2 for AA, copy number 1 for AB and AC, and copy number 0 for BB, BC, and CC. Allele B has copy number 2 for BB, copy number 1 for AB and BC, and copy number 0 for AA, AC, and CC. Allele C has copy number 2 for CC, copy number 1 for AC and BC, and copy number 0 for AA, AB, and BB.

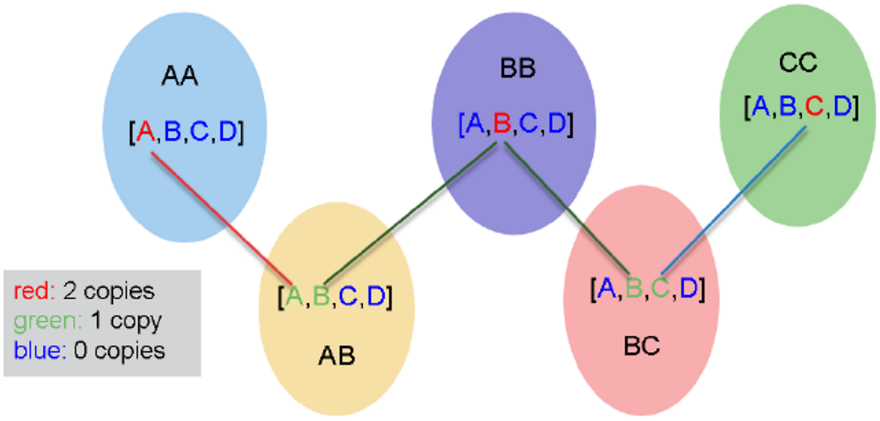


Figure 70 Clusters from a multiallelic SNP have a different copy number for each allele.

The copy number states for allele A are displayed in Figure 71.

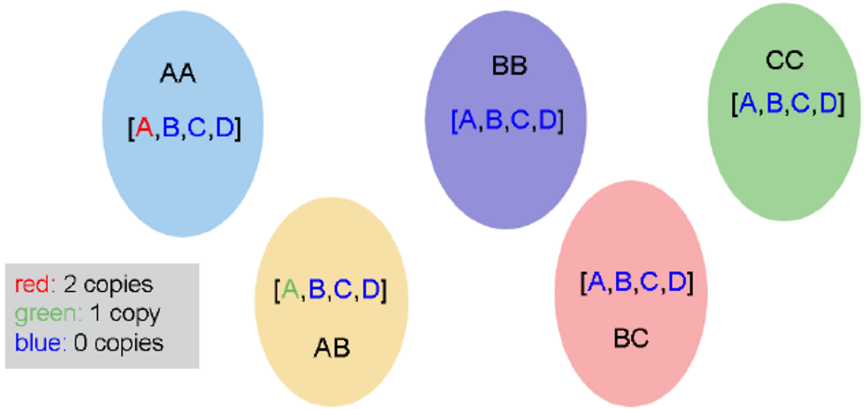


Figure 71 Copy number values for allele A for a multiallelic SNP.

Alleles that have copy number 0 do not contribute to the allele signal and hence are background signal. In Figure 71, clusters BB, BC, and CC do not have any A alleles and alleles B, C, and D are background for allele A.

HomMMA is the multiallelic metrics that measures signal versus background strength and is the minimum of the means for each allele's homozygous cluster. For allele A, the mean value is mean A from the row for cluster AA. If a probeset has alleles A,

B, and C, *HomMMA* is the minimum of AA-meanA, BB-meanB, and CC-meanC. The higher the *HomMMA* value, the stronger the allele signal is and the more defined the cluster.

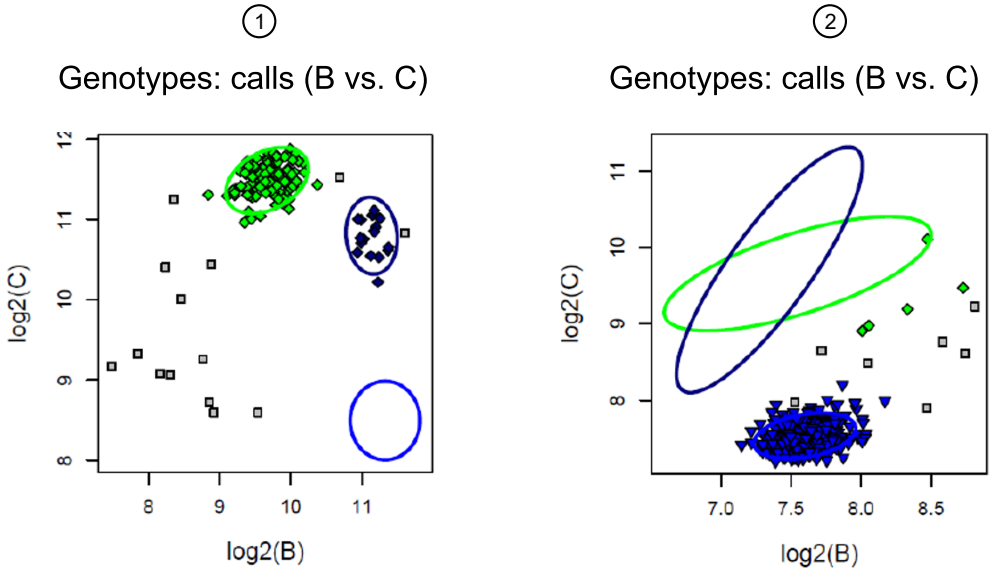


Figure 72 SNPs with high and low HomMMA.

① A SNP with high MomMMA (11.5)
② A SNP with low MomMMA (7.6)

Biallelic-derived multiallelic metrics

Several multiallelic metrics are analogous to bialellelic metrics and are created by transforming the multiallelic posterior cluster locations into size vs contrast space for each biallelic pair of alleles. For the FLD calculations, a weighted variance across all biallelic pairs is used to mimic the pooled variance found in the biallelic posteriors. The formulas for the transformed means and variances use means, variances, and covariances from the multiallelic posteriors file for each pair of alleles. For example, if we would like to calculate the transformed means and variances for alleles A and B, we would use the rows for clusters AA, AB, and BB and meanA, meanB, varA, varB, and covarAB from each row. For any two alleles i and j, the transformed means are:

$$mean(X) = log_2(i) - log_2(j)$$

$$mean(Y) = \frac{log_2(i) + log_2(j)}{2}$$

The values in the multiallelic posteriors file have already been transformed into log2 space, so $log_2(i)$ is just the meanI value and $log_2(j)$ is just the meanJ value. Figure 73 shows the posteriors for the example probeset AX-123456789. It has 3 alleles (A, B, C) and the order of the means and variances are given in the header lines. Using this ordering, we know which mean and variance values go with which alleles per cluster row.

#%data-order-mean-nalleles-3=A,B,C
#%data-order-covariance-nalleles-3=varA,covAB,covAC,varB,covBC,varC

probeset_id	copynumber	nAlleles	cluster	mean	nObsMean	covariance	nObsVar
AX-123456789	2	3	AA	12.163257,11.346342	279.2	0.042845,0.029852,0.0	280
AX-123456789	2	3	AB	11.487242,11.590495	6.3	0.065806,0.018384,0.0	16
AX-123456789	2	3	AC	11.658022,10.963805	27.3	0.041822,0.014533,-0.0	37
AX-123456789	2	3	BB	10.248318,13.027269	0.2	0.060000,0.000000,0.0	1
AX-123456789	2	3	BC	10.248318,12.827269	0.3	0.060000,0.000000,0.0	10
AX-123456789	2	3	CC	10.192613,10.108785	5.2	0.152987,0.060522,0.0	6

Figure 73 Posteriors for the multiallelic probeset AX-123456789.

If we want the transformed means for the biallelic pairing AB, we can use the A and B values and ignore the C values. We know the mean values for the AA, AB, and BB clusters, and we can easily calculate the transformed means:

Table 15 Multiallelic cluster means for alleles A and B.

	MeanA	MeanB
AA	12.16327	11.34634
AB	11.48724	11.59050
BB	10.24832	13.02727

meanX for cluster AA is meanA - meanB using the values from the AA row: 12.16327 – 11.34634 = 0.816925. meanX for cluster AB is meanA - meanB using the values from the AB row: 11.48724 – 11.59050 = -0.10325. meanX for cluster BB is meanA - meanB using the values from the BB row: 10.24832 – 13.02727 = -2.77895.

meanY for cluster AA is (meanA+meanB)/2 using the values from the AA row: (12.16327+11.34634)/2 = 11.7548. meanY for cluster AB is (meanA+meanB)/2 using the values from the AB row: (1.48724+11.59050)/2 = 11.53887. meanY for cluster BB is (meanA+meanB)/2 using the values from the BB row: (10.24832+13.02727)/2 = 11.63779.

For any two alleles i and j, the transformed variances are:

$$var(X) = var(log_2(i)) + var(log_2(j)) - 2cov(log_2(ij))$$

$$var(Y) = \frac{1}{4}\Big(var(log_2(i)) + var(log_2(j)) + 2cov(log_2(ij))\Big)$$

$$cov(XY) = \frac{1}{2}\Big(var(log_2(i)) - var(log_2(j))\Big)$$

The variance and covariance values given in the multiallelic posteriors file are already transformed into log_2 space, so the values can be used directly in these formulas.

Continuing with the example of probeset AX-123456789, if we want the transformed variances for the biallelic pairing AB, we can use the A and B values and ignore the C values. We know the variance and covariance values for the AA, AB, and BB clusters, and we can easily calculate the transformed variances and covariance:

Table 16 Multiallelic Cluster Variances for Alleles A and B.

	varA	covarAB	varB
AA	0.042845	0.029852	0.042786
AB	0.065806	0.018384	0.074312
BB	0.06	0.0	0.06

varX for cluster AA is varA + varB -2covAB using the values from the AA row: 0.042845 + 0.042786 – 2 * 0.029852 = 0.025927. varX for cluster AB is varA + varB - 2covAB using the values from the AB row: 0.065806 + 0.074312 - 2 * 0.018384 = 0.10335. varX for cluster BB is varA + varB - 2covAB using the values from the BB row: 0.06 + 0.06 - 2 * 0.0 = 0.12.

varY for cluster AA is ¼*(varA + varB + 2covarAB) using the values from the AA row: ¼*(0.042845 + 0.042786 + 2 * 0.029852) = 0.036334. varY for cluster AB is ¼*(varA + varB + 2covarAB) using the values from the AB row: ¼*(0.065806 + 0.074312 + 2 * 0.018384) = 0.044222. varY for cluster BB is ¼*(varA + varB + 2covarAB) using the values from the BB row: ¼*(0.06 + 0.06 + 2 * 0.0) = 0.03.

covarXY for cluster AA is ½*(varA-varB) using the values from the AA row: ½*(0.042845-0/042786) = 0.0000295. covarXY for cluster AB is ½*(varA - varB) using the values from the AB row: ½*(0.065806 – 0.074312) = -0.00425. covarXY for cluster BB is ½*(varA - varB) using the values from the BB row: ½*(0.06 - 0.06) = 0.

Note that a cluster may have different mean and variance values depending on which biallelic pairings are used when calculating the transformed means and variances. The example above used the alleles A and B and ignored allele C, but we could have easily used alleles A and C and ignored allele B:

Table 17 Multiallelic transformed example with alleles A and C.

	meanX	meanY	varX	varY	covarXY
AA	1.336064	11.49524	0.052142	0.032402	-0.00259
AC	0.264041	11.526	0.69138	0.014983	0.009555
CC	-1.25352	10.81937	0.104589	0.064636	0.062204

The last transformation formula that is used is the weighted variance. The weighted variance is a single variance value per probeset for X and Y that is a weighted average of all variances for the clusters. If a probeset has N clusters, then the weighted variances are:

$$var_{weighted}(X) = \frac{\sum_{i=1}^{N} nObsVar_i * varX_i}{\sum_{i=1}^{N} nObsVar_i}$$

$$var_{weighted}(Y) = \frac{\sum_{i=1}^{N} nObsVar_i * varY_i}{\sum_{i=1}^{N} nObsVar_i}$$

Using the values from the biallelic pairing A and B and C, the weighted variances are:

Table 18 Multiallelic weighted variances with alleles A, B, and C.

	varX	varY	nObsVar	Var(weighted)X	Var(weighted)Y
AA	0.025927	0.036334	280	0.04501586	0.03372433
AB	0.10335	0.044222	16	0.04501586	0.03372433
BB	0.12	0.03	1	0.04501586	0.03372433
AA	0.052142	0.032402	280	0.04501586	0.03372433
AC	0.069138	0.014983	37	0.04501586	0.03372433
CC	0.104589	0.064636	6	0.04501586	0.03372433

Notice that the weighted variances for X and Y are the same for all clusters.

The cluster AA appears twice in Table 18. While this is not the usual way for displaying the clusters produced by the biallelic pairings, it is important to show that the cluster AA has values generated using two different alleles: B and C. When calculating the weighted variances, both AA cluster contribute. This is not a double-counting of cluster AA because the means and variances reference two different biallelic planes: A versus B and A versus C. The mean and variance values for AA are with respect to the two different biallelic planes that the cluster is being measured in. The weighted variances are calculated across all of the biallelic planes that are possible from the higher-dimensional multiallelic plane so any cluster that can be measured in different biallelic planes must be included in the weighted calculations.

FLD_MA, *MinFLD_MA*, and *HomFLD_MA* can be used to determine if the clusters in a multiallelic probeset are well separated and tightly clustered. Figure 74 shows a multiallelic SNP with good cluster separation across all alleles. All clusters are populated in this example. *FLD_MA* and *MinFLD_MA* are calculated using the homozygous-heterozygous cluster pairs of AA-AB, BB-AB, AA-AC, CC-AC, BB-BC, and CC-BC. *HomFLD_MA* is calculated using the homozygous-homozygous cluster pairs of AA-BB, AA-CC, and BB-CC.

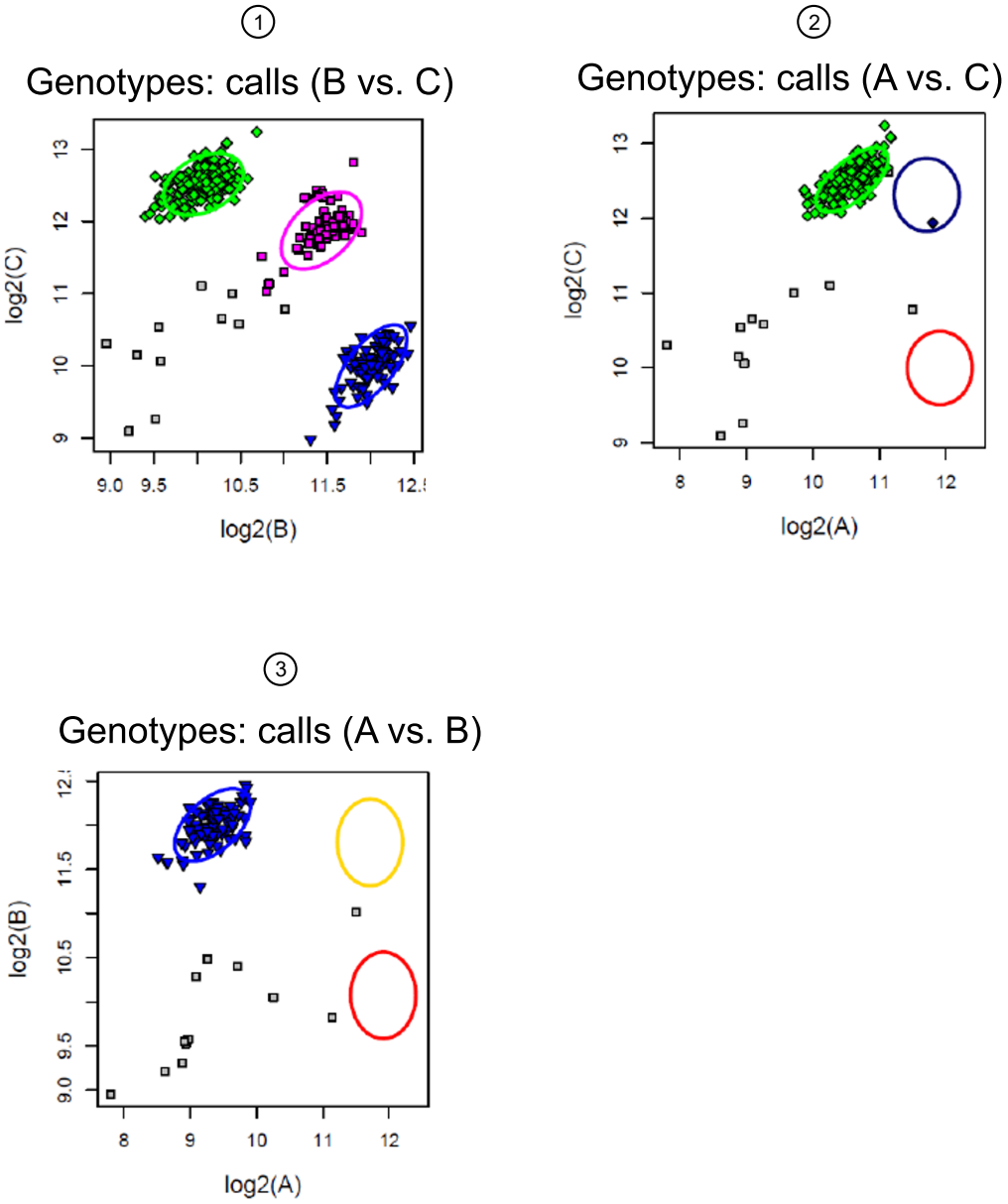


Figure 74 A multiallelic SNP with high FLD_MA, MinFLD_MA, and HomFLD_MA values has good cluster separation in all pairwise biallelic combinations.

1. One SNP with high FLD_MA (7.9).
2. The same SNP with high MinFLD_MA (5.4).
3. The same SNP with high HomFLD_MA (18.3).

Figure 75 shows a multiallelic SNP where the clusters are not well separated. The CC cluster is not populated in this example. *FLD_MA* and *MinFLD_MA* are calculated using the homozygous-heterozygous cluster pairs of AA-AB, BB-AB, AA-AC, and BB-BC. *HomFLD_MA* is calculated using the homozygous-homozygous cluster pair of AA-BB.

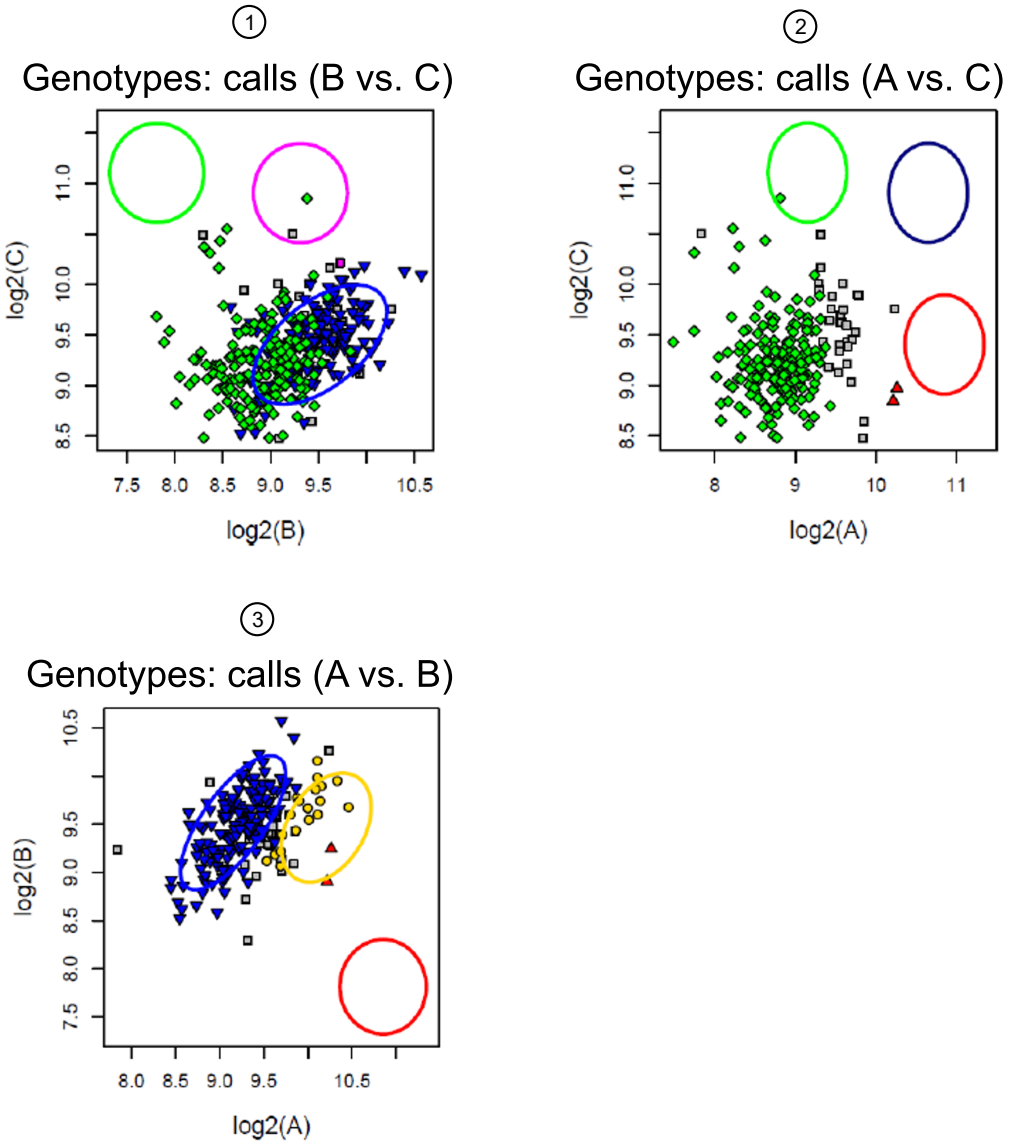


Figure 75 A multiallelic SNP with low *FLD_MA*, *MinFLD_MA*, and *HomFLD_MA* values has clusters that have not separated well.

1. One SNP with low FLD_MA (4.1).
2. The same SNP with low MinFLD_MA (3.6).
3. The same SNP with low HomFLD_MA (8.2).

B Complete set of classification thresholds used by ps-classification

Base workflow classification thresholds

Any biallelic probeset that has metrics values larger than the thresholds in the base workflow will be classified as PolyHighResolution, NoMinorHom, or MonoHighResolution depending on the number of genotype clusters. Probesets that pass all thresholds except for cr-cutoff are classified as CallRateBelowThreshold. Probesets that pass all thresholds except the genotype frequency p-value threshold are classified as UnexpectedGenoFreq for arrays which have a provided genotype frequency file; additionally, UnexpectedGenoFreq probesets can fail the HomRO thresholds and still be classified as UnexpectedGenoFreq. Probesets that pass all thresholds except the het-so-otv-cutoff threshold are classified as OTV. Probesets that fail any other thresholds and cut-offs or that fail multiple thresholds are classified as Other.

Table 19 The default threshold cut-off values for the base classification workflow.

	Human	Diploid	Polyploid
cr-cutoff	95	97	97
fld-cutoff	3.6	3.6	3.6
het-so-cutoff	-0.1	-0.1	-0.1
het-so-otv-cutoff	-0.3	-0.3	-0.3
hom-ro-1-cutoff	0.6	0.6	N/A
hom-ro-2-cutoff	0.3	0.3	N/A
hom-ro-3-cutoff	-0.9	-0.9	N/A
num-minor-allele-cutoff	2	2	2
hom-ro	True	True	False
genotype-p-value-cutoff	1e-06	1e-06	1e-06

There are three HomRO thresholds, one for the number of clusters that appear (e.g., hom-ro-1-cutoff is used when there is only one genotype cluster for a probeset). Hom-ro indicates if the HomRO thresholds should be used in classification. Polyploid

genotypes do not use any of the HomRO thresholds so hom-ro is automatically set to FALSE and cannot be set to TRUE.

Num-minor-allele-cutoff is the threshold for the minimum number of minor alleles that must appear in the total calls for a probeset to be classified as NoMinorHom or PolyHighResolution instead of MonoHighResolution.

The default value for genotype-p-value-cutoff is set to be very small because this threshold should only be used to flag SNPs with very unexpected genotype frequency values. We recommended leaving this threshold value very small. If a user wishes to have the p-values output in the performance file but not have any SNPs classified as UnexpectedGenotypeFreq, this threshold can be set to 0 to ensure that all SNPs pass. Genotype-p-value-cutoff is the only threshold value that is used with both biallelic and multiallelic SNPs. If genotype frequency p-values were calculated by ps-metrics, then any SNP with one or more p-value that is smaller than genotype-p-value-cutoff is classified as UnexpectedGenotypeFreq.

Copy-number aware metrics

When copy-number aware genotype metrics are available, an extra set of thresholds are used in classifying SNPs. The additional thresholds are used to check the positions of diploid, haploid, and CN0 clusters. The default values for these thresholds do not change based on species type.

Table 20 The default threshold values used in the biallelic base classification workflow with copy-number aware genotype metrics.

Threshold	Default
hom-ro-hap-1-cutoff	0.4
hom-ro-hap-2-cutoff	0.2
hom-hap-X-cutoff	-0.2
hom-hap-Y-upper-cutoff	0.2
hom-hap-Y-lower-cutoff	1
CN0-hap-X-cutoff	-0.2
CN0-hap-Y-cutoff	0
CN0-dip-X-cutoff	0.2
CN0-dip-Y-cutoff	0

HomRO hap1.cut and HomRO hap2.cut are the minimum HomRO values when a biallelic SNP has one or two haploid clusters. Hom-hap-X-cutoff is the threshold for the maximum distance on the X axis between diploid and haploid clusters (i.e., haploid clusters should not be outside diploid clusters). Hom-hap-Y-upper-cutoff and hom-hap-Y-lower-cutoff are the range for the maximum distance on the Y axis between diploid and haploid clusters (i.e., haploid clusters should be below

diploid clusters). CN0-hap-X-cutoff, CN0-hap-Y-cutoff, CN0-dip-X-cutoff, and CN0-dip-Y-cutoff are the thresholds for the maximum distance between haploid CN0 clusters and diploid and CN0 clusters in the X and Y dimensions (i.e., the CN0 clusters should sit at 0 on the X axis and below the haploid and diploid clusters).

Special SNPs classification

If the biallelic metrics file has gender-separated metrics, additional thresholds are used in classification of MT, CP, Y, W, non-PAR X, and Z SNPs. These thresholds are used to check the positions of haploid and CN0 clusters for MT, CP, Y, and W SNPs, positions of clusters by gender for Y, W, non-PAR X, and Z SNPs, and positions of diploid, haploid, and CN0 clusters for non-PAR X and Z SNPs. Non-PAR X and Z SNPs can be PolyHighRes, NoMinorHom, MonoHighRes, CallRateBelowThreshold, UnexpectedGenotypeFreq, and Other. If non-PAR X or Z SNPs are in the dataset but gender-separated metrics were not produced, then all non-PAR X or Z SNPs will be classified as though they were autosomal SNPs. Y, W, MT, and CP SNPs can be PolyHighRes, MonoHighRes, CallRateBelowThreshold, UnexpectedGenotypeFreq, and Other. If Y, W, MT, or CP SNPs are in the dataset but gender-separated metrics were not produced, then all Y, W, MT, or CP SNPs will be classified as hemizygous. Note that MT and CP SNPs use the MT thresholds because there is no gender difference.

Table 21 The default threshold values in the biallelic base classification workflow with special SNPs metrics.

Threshold	Default	Gender
hom-ro-hap-1-MTChr-cutoff	0.4	All
hom-ro-hap-2-MTChr-cutoff	0.2	All
homfld-YChr-cut	6.5	Male
min-YChr-samples-cut	5	All
hom-ro-hap-1-XChr-cutoff	0.1	Male
hom-ro-hap-2-XChr-cutoff	0.05	Male
homfld-XChr-cut	6.5	Male
fld-XChr-cut	4	Female
het-so-XChr-cutoff	-0.1	Female
aaf-XChr-cut	0.36	All
hom-ro-hap-1-ZChr-cutoff	0.1	Female
hom-ro-hap-2-ZChr-cutoff	0.05	Female
homfld-ZChr-cut	6.5	Female

Table 21 The default threshold values in the biallelic base classification workflow with special SNPs metrics. *(continued)*

Threshold	Default	Gender
fld-ZChr-cut	4	Male
het-so-ZChr-cutoff	-0.1	Male
aaf-ZChr-cut	0.36	All

Hom-ro-hap-1-MTChr-cutoff and hom-ro-hap-2-MTChr-cutoff are the minimum HomRO values when an MT or CP SNP has one or two haploid clusters. Homfld-YChr-cut is the minimum HomFLD hap value on male samples for Y probesets. Min-YChr-samples-cut is the minimum number of male and female samples that must be assigned genotypes to check the distances between the male and female samples: the females should be below the males and towards the center of the X axis; and if there are two male clusters, they should be well separated. HomFLD_W.cut is the minimum HomFLD hap value on female samples for W probesets. min_W_samples is the minimum number of male and female samples that must be assigned genotypes to check the distances between the male and female samples: the males should be below the females and towards the center of the X axis; and if there are two female clusters, they should be well separated.

FLD and HomRO for female samples are used as part of the base classification workflow but male HomRO values are only checked as part of the extra classification for non-PAR X SNPs. Hom-ro-hap-1-XChr-cutoff and hom-ro-hap-2-XChr-cutoff are the minimum HomRO values for male samples only for non-PAR X SNPs with 1 or 2 haploid clusters. Homfld-XChr-cut is minimum HomFLD hap value for male samples only for non-PAR X SNPs. Het-so-XChr-cutoff is the minimum HetSO value for female samples only on non-PAR X SNPs. Aaf-XChr-cut is the threshold for the difference between the female and male values of AAF (A-allele frequency) for non-PAR X SNPs. There should not be too large of a difference between AAFs calculated on the two genders.

FLD and HomRO for male samples are used as part of the base classification workflow but female HomRO values are only checked as part of the extra classification for Z SNPs. Hom-ro-hap-1-ZChr-cutoff and hom-ro-hap-2-ZChr-cutoff are the minimum HomRO values for female samples only for Z SNPs with 1 or 2 haploid clusters. Homfld-ZChr-cut is minimum HomFLD hap value for female samples only for Z SNPs. Het-so-ZChr-cutoff is the minimum HetSO value for male samples only on Z SNPs. Aaf-ZChr-cut is the threshold for the difference between the female and male values of AAF (A-allele frequency) for Z SNPs. There should not be too large of a difference between AAFs calculated on the two genders.

Supplemental/SSP workflows

To run the supplemental or SSP classification workflows, the supplemental or SSP flag must be set to TRUE in APT. The Z-scores are calculated and are compared to the Z-score thresholds if the user does not supply thresholds for the variance values. If thresholds for variances are supplied, the raw variance values are used instead of the Z-scores. When all Z-score and variance thresholds are set to NA, the only supplemental categories used in classification will be UnexpectedHeterozygosity and HomHomResolution.

The SSP workflow is identical to the supplemental workflow except it include the 3 extra SSP metrics in the output performance file.

Table 22 Default values for the optional threshold arguments in the supplemental and SSP workflows.

Threshold	Description	Default
BB.varY.Z.cut	Variance Z-score threshold for BB cluster in the Y dimension	3
BB.varX.Z.cut	Variance Z-score threshold for BB cluster in the X dimension	3
AB.varY.Z.cut	Variance Z-score threshold for AB cluster in the Y dimension	3
AB.varX.Z.cut	Variance Z-score threshold for AB cluster in the X dimension	3
AA.varY.Z.cut	Variance Z-score threshold for AA cluster in the Y dimension	3
AA.varX.Z.cut	Variance Z-score threshold for AA cluster in the X dimension	3
BB.varX.cut	Variance threshold for BB cluster in the X dimension Will not be used if BB.varX.Z.cut is supplied	N/A
BB.varY.cut	Variance threshold for BB cluster in the Y dimension Will not be used if BB.varY.Z.cut is supplied	N/A
AB.varX.cut	Variance threshold for AB cluster in the X dimension Will not be used if AB.varX.Z.cut is supplied	N/A
AB.varY.cut	Variance threshold for AB cluster in the Y dimension Will not be used if AB.varY.Z.cut is supplied	N/A
AA.varX.cut	Variance threshold for AA cluster in the X dimension Will not be used if AA.varX.Z.cut is supplied	N/A
AA.varY.cut	Variance threshold for AA cluster in the Y dimension Will not be used if AA.varY.Z.cut is supplied	N/A

Table 22 Default values for the optional threshold arguments in the supplemental and SSP workflows. *(continued)*

Threshold	Description	Default
clustermin	Minimum number of samples per cluster	5
nMinorHom.cut	Threshold for the nMinorHom category	0
HetvMAF.cut	Threshold for the UnexpectedHeterozygosity category	10
HomFLD.cut	Threshold for homozygous cluster resolution (HomFLD)	6.5

The Z.cut thresholds are used with Z-scores, and the default values (3) are very large for Z-scores. These defaults should only capture SNPs that have a cluster with a very unusually large variance (i.e., cluster split). The Z.cut thresholds cannot be negative.

The default values for the .cut thresholds are NA. If a Z-score threshold is not supplied and values are not supplied for the .cut values, then the category is turned off (see below for more details). The .cut arguments are used with the size (Y) or contrast (X) data. We suggest using a value of 10, which is very large for cluster variances and should only capture SNPs that have a cluster with a very unusually large variance. The .cut thresholds should only be used when it is clear that the data is not close to normally distributed. We recommend using the Z-score threshold values, if possible.

To turn off filtering for a cluster/dimension, do not supply values for both types of arguments. When values are not supplied for Z.cut and .cut thresholds, a message is output to remind the user that the category has been removed from the list of possible categorizations.

Non-numeric arguments for ps-classification

There are several non-numeric input arguments for ps-classification that affect the classification results.

Table 23 Non-numeric input arguments to ps-classification.

Argument	Default	Workflow
hom-ro	True	Base
priority-order	PHR, NMH, MHR, OTV, UnexGenoFreq, CRBT, Other, OtherMA	All biallelic workflows
priority-order-multi-allele	PHR, NMH, MHR, Hemizygous, UnexGenoFreq, CRBT, Other, OtherMA	Multiallelic
variance-class	PolyHighResolution	Supplemental

As stated above, hom-ro controls whether the HomRO thresholds are used in classification. The default is TRUE.

Priority-order and priority-order-multi-allele are used when selecting the best probeset for a SNP. If more than one probeset is tied in the selection algorithm for best probeset, then the best probeset is selected according to the priority order: PHR is selected over NMH, etc. The priority.order argument allows the user to change the order of categories when determining which probesets are selected as the best probeset for a SNP. All of the listed categories must appear in priority.order when the user species the order. Priority-order-multi-allele is the same argument for the multiallelic workflow and allows the user to select different priority orders for biallelic and multiallelic probesets that are in the same data set. The OTV category cannot be listed in priority-order-multi-allele because it is not a category that multiallelic SNPs can belong.

The argument variance-class holds the list of categories on which to calculate the z-scores (default is PolyHighResolution). Note that the Z-scores change as different categories are used with variance-class. Changing variance-class from just PolyHighResolution to PolyHighResolution and NoMinorHom will change the Z-scores, and some of the filtering results will change as well. The user should select which categories are interesting and use those in variance-class.

Multiallelic workflow

The multiallelic workflow uses the multiallelic metrics and the thresholds are separate from the biallelic workflows. HomMMA.cut, FLD_MA.cut, MinFLD_MA.cut, and HomRO_MA.cut are the minimum values of these metrics necessary for a multiallelic SNP to pass and be sorted into a recommended category. For multiallelic SNPs with only 2 alleles detected, the thresholds used are FLD_MA.2.cut, MinFLD_MA.2.cut, HetSO_MA.2.cut, and HomRO_MA.2.cut. HomRO_MA.1.cut is used when only 1 allele was detected for a multiallelic SNP. BestCR_MA.cut is part of the process for determining the best probeset when a ps2snp file is provided.

Table 24 Default threshold values for the multiallelic classification workflow.

Threshold	Default
HomMMA.cut	10
FLD_MA.cut	5.2
FLD_MA.2.cut	5.2
MinFLD_MA.cut	0
MinFLD_MA.2.cut	0
HetSO_MA.2.cut	-0.1
HomRO_MA.cut	0.2
HomRO_MA.2.cut	0.3

Table 24 Default threshold values for the multiallelic classification workflow. *(continued)*

Threshold	Default
HomRO_MA.1.cut	0.6
BestCR_MA.cut	90

If a multiallelic SNP had haploid calls (indicated by the ending "_diploid" appended to the probeset name), classification is performed using all of the metrics produced by Ps Metrics. These multiallelic SNPs are not classified into any special categories.

Multiallelic non-PAR X SNPs are classified similarly to multiallelic SNPs with haploid calls: classification is performed using the metrics from Ps Metrics, which were calculated to handle any haploid calls. Non-PAR X SNPs are classified into the same categories as autosomal multiallelic SNPs.

Multiallelic Y and MT SNPs are evaluated on a very small set of metrics: CR, Nclus, and GenotypeFreqPvals. Y and MT SNPs cannot be classified as NoMinorHom because of the lack of heterozygous clusters. Y and MT SNPs can be classified into any of the remaining categories.

The specialSNP_chr column indicates if a probeset is a special SNP and which chromosome it's on. The options are for multiallelic probesets are autosomal, X, Y, MT, and CP. Biallelic probesets can also have Z and W chromosomes.

If a special SNPS file is not provided when running ps-metrics, non-PAR X SNPs are treated as though they are autosomal SNPs and all Y, MT, and CP SNPs are categorized as Hemizygous. This special category is reserved for SNPs where the posteriors file indicates that only haploid calls are expected but without the knowledge of which chromosome the SNP is found on because of the lack of a special SNPs file. Hemizygous SNPs cannot be categorized into other categories without the special SNPs file.

Best probeset selection

A SNP may be interrogated by more than one probeset. If a SNP has more than one probeset, the ps2snp file contains the mapping of probesets to SNPs. Probeset names are found in the probeset_id column and SNP names are found in the snpid column. When the ps2snp file is supplied to ps-classification, one probeset per SNP is selected as the best performing probeset for that SNP. If there is only one probeset for a SNP, then that probeset is automatically labelled as the best probeset. When there are multiple probesets, the classification category is used to select which probeset is best.

The priority-order argument controls the order of categories for the best probeset selection. The default value for priority-order is PolyHighResolution, NoMinorHom, MonoHighResolution, OTV, UnexpectedGenotypeFreq, CallRateBelowThreshold, Other, OtherMA. If a SNP has two probesets that have been classified as PolyHighResolution and MonoHighResolution, then the PolyHighResolution probeset will be selected as the best probeset.

When the supplemental workflow is used, the default value for priority-order includes the supplemental categories as well: PolyHighResolution, nMinorHom, UnexpectedHeterozygosity, ABvarianceY, AAvarianceY, BBvarianceY, ABvarianceX, AAvarianceX, BBvarianceX, HomHomResolution, NoMinorHom, OTV, MonoHighResolution, UnexpectedGenotypeFreq, CallRateBelowThreshold.

When there is a tie for best probeset using categories, best probeset selection is performed on the tied probesets using the metrics values. The best probeset is selected using the best CR value, then the best FLD value, then the best HomFLD value, and then the best HetSO value. If there is still a tie between two probesets, then the probeset that is first alphabetically will be selected. If a SNP has two probesets that have been classified as PolyHighResolution and one probeset that has been classified as MonoHighResolution, the metrics values will be used to break the tie between the two PolyHighResolution probesets. If one probeset has a CR of 98.5% and the other probeset has a CR of 100%, then the second probeset will be selected as the best probeset.

Probesets that are selected as the best probeset and which are also in the recommended categories are labelled as BestandRecommended. The default values for the recommended categories are PolyHighResolution, NoMinorHom, MonoHighResolution, and Hemizygous for Human and Diploid classification and PolyHighResolution for Polyploid classification. Recommended categories are those where the genotyping quality results are high enough that all probesets in those categories are recommended for use in further studies. A probeset may be the best probeset for a SNP but not in a recommended category, and probesets that are in recommended categories may not the best probeset for SNP. In the first case, all probesets for a SNP may be categorized into non-recommended categories (e.g., Other), in which case the best probeset will not be recommended. In the second case, all probesets for a SNP may be categorized into recommended categories (e.g., PolyHighResolution and MonoHighResolution) so even the probesets that are not selected as the best probeset are still recommended. The Best and Recommended probeset list is recommended for use in downstream analysis to obtain the highest quality results.

Multiallelic best probeset selection is very similar to biallelic best probeset selection. If a multiallelic SNP is interrogated by multiple probesets, the best probeset will be selected using the multiallelic categories. The default value for priority-order-multi-allele is PolyHighResolution, NoMinorHom, MonoHighResolution, Hemizygous, UnexpectedGenotypeFreq, CallRateBelowThreshold, Other. If there is a tie for best probeset for a multiallelic SNP, the metrics values are used to select the best probeset from the ties. The best probeset is selected using the larger value of nAllelesDetected, then the best FLD MA value, and then the best HomFLD MA value.

When a four-column ps2snp file is supplied rather than a two-column ps2snp file, the additional information allows ps-classification to identify all biallelic and multialellic probesets that contribute genotypes to a multiallelic SNP. The algorithm for selecting the best probeset is more complex when there are both bilallelic and multiallelic probesets to select from. Multiallelic probesets that have detected three or more alleles are preferred over biallelic probesets. Biallelic probesets may be selected as the best probeset when multiallelic probesets detect fewer than three alleles or a biallelic probeset is in a recommended category and a multiallelic probeset is not.

The four columns in a ps2snp file are probeset_id, snpid (or affy_snp_id). multi_snp_id, and ordered_alleles. The ordered_alleles column shows the reference and alternative alleles. A probeset with only one reference and alternative is a biallelic probeset. For example, probeset AX-123456781 has alleles A/C/G/T, which indicates that the reference allele is A and that there are three separate alternative alleles, C and G and T. This is a multiallelic probeset.

The probeset_id and snpid columns are the same between the two- and four-column ps2snp files. For a multiallelic SNP, the snpid column has the same value as the multi_snp_id column. The multi_snp_id column holds the SNP name for a multiallelic SNP, and is used to identify which probesets contribute to the genotypes for a multiallelic SNP.

probeset_id	affy_snp_id	multi_snp_id	ordered_alleles
AX-123456780	Affx-101	Affx-1	A/G
AX-123456781	Affx-1	Affx-1	A/C/G/T
AX-123456782	Affx-103	Affx-1	C/G
AX-123456783	Affx-104	Affx-1	T/G
AX-123456784	Affx-105	Affx-2	C/G
AX-123456785	Affx-106	Affx-2	T/G
AX-123456786	Affx-107	Affx-2	A/G
AX-123456787	Affx-2	Affx-2	A/C/G/T
AX-123456788	Affx-109	Affx-3	C/CGTA
AX-123456789	Affx-3	Affx-3	C/CGTA/CTAC
AX-123456790	Affx-111	Affx-3	C/CTAC
AX-123456791	Affx-4	Affx-4	A/AT/TA
AX-123456792	Affx-113	Affx-4	A/AT
AX-123456793	Affx-114	Affx-4	A/TA
AX-123456794	Affx-5	Affx-5	A/AC/C
AX-123456795	Affx-116	Affx-5	AC/C
AX-123456796	Affx-117	Affx-5	A/C
AX-123456797	Affx-118	Affx-6	A/TA
AX-123456798	Affx-6	Affx-6	A/T/TA
AX-123456799	Affx-120	Affx-6	A/T
AX-123456800	Affx-121	Affx-7	C/G
AX-123456801	Affx-7	Affx-7	C/G/GG
AX-123456801	Affx-123	Affx-7	G/GG

Figure 76 An example of a 4-column ps2snp file.

For multiallelic SNP Affx-1, the biallelic probesets AX-123456780, AX-123456782, and AX-123456783 are used together to produce the final multiallelic genotypes for the multiallelic probeset AX-123456781.

When a four-column ps2snp file is supplied, the best probeset is selected with the same best probeset selection that is used with a two-column ps2snp file. Once a best multiallelic probeset is selected, ps-classification checks if the multiallelic SNP is really multiallelic or if it only has genotypes with 2 alleles (nAllelesDetected). All multiallelic probesets for a SNP are checked against three thresholds to determine if there are any multiallelic probesets which have detected 3 or more alleles that are of high enough quality: BestCR MA.cut (default=90%), FLD MA.cut (default=5.2), and MinFLD_MA.cut (default=0). Any multiallelic probesets that pass all 3 thresholds then are evaluated on nAllelesDetected. If there are any multiallelic probesets

that pass these 3 thresholds and which have 3 or more alleles detected, the SNP is multiallelic and the BestandRecommended probeset for the SNP remains the BestandRecommended probeset. If none of the multiallelic probesets are recommended even though at least one passes the 3 thresholds and has 3 or more alleles detected, then the best probeset remains multiallelic but there is no BestandRecommended probeset for the multiallelic SNP.

When all multiallelic probesets have only 1 or 2 alleles detected, the best probeset selection determines that the multiallelic probesets have produced biallelic genotypes. In this situation, the biallelic probesets that are BestandRecommended are evaluated to see if one of them should be used as the BestandRecommended probeset for the multiallelic SNP.

If none of the biallelic probesets are the BestandRecommended for their snpid, the multiallelic probesets with 1 or 2 alleles detected are reevaluated. If one of these multiallelic probesets with 1 or 2 alleles detected was originally the BestandRecommended probeset for the multiallelic SNP, it remains the BestandRecommended even though it has fewer than 3 alleles detected. If none of the multiallelic and biallelic probesets were originally BestandRecommended, then there is no BestandRecommended probeset for the multiallelic SNP.

When there is one or more biallelic probesets that are recommended, ps-classification evaluates them to determine if one should be the BestandRecommended probeset for the multiallelic SNP.

If there is one BestandRecommended biallelic probeset, then that probeset is the BestandRecommended probeset for the multiallelic SNP; the multiallelic probeset classifications are changed to Other and the original classifications are stored in the column OriginalCT.

If there is more than one BestandRecommended biallelic probeset for a multiallelic SNP, then these probesets are evaluated for best probeset selection. Any conflicts between biallelic probesets indicate that the multiallelic SNP has problematic results. If two of the biallelic probesets are classified as PolyHighResolution and Hemizygous, this indicates an issue determining if the multiallelic SNP was autosomal or a special SNP. In this case, all probesets for the multiallelic SNP will be reclassified as Other and the original classifications will be stored in the column OriginalCT.

If there is more than one BestandRecommended biallelic probeset that is classified as MonoHighResolution, then the homozygous cluster for all MonoHighResolution probesets should use the same allele. If the homozygous clusters in multiple MonoHighResolution probesets do not match, all probesets for the multiallelic SNP will be reclassified as Other and the original classifications will be stored in the column OriginalCT.

If there is more than one BestandRecommended biallelic probeset that is classified as PolyHighResolution, then the assigned genotypes are problematic for a multiallelic SNP and all probesets for the multiallelic SNP will be reclassified as Other and the original classifications will be stored in the column OriginalCT.

If all of the BestandRecommended biallelic probesets that detect 2 alleles have been classified as NoMinorHom, then the best probeset from the NoMinorHom probesets are selected using best call rate, then largest number of heterozygous samples, and then best FLD.

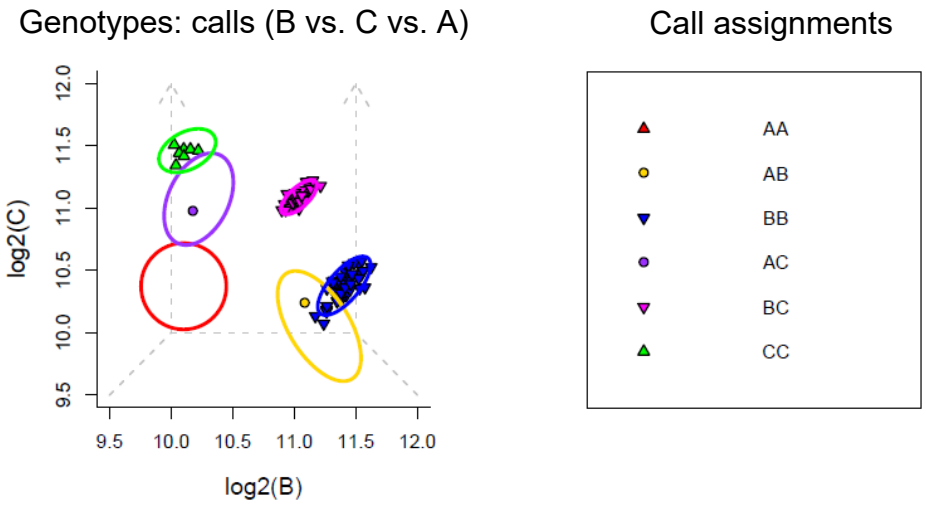


Figure 77 A multiallelic SNP that is BestandRecommended.

Figure 77 shows a multiallelic probeset that interrogates 3 alleles and has high CR and FLD values. This probeset is selected as the best probeset for a multiallelic SNP, and because it was classified as PolyHighResolution and has detected more than 2 alleles, it is also the BestandRecommended probeset for the SNP.

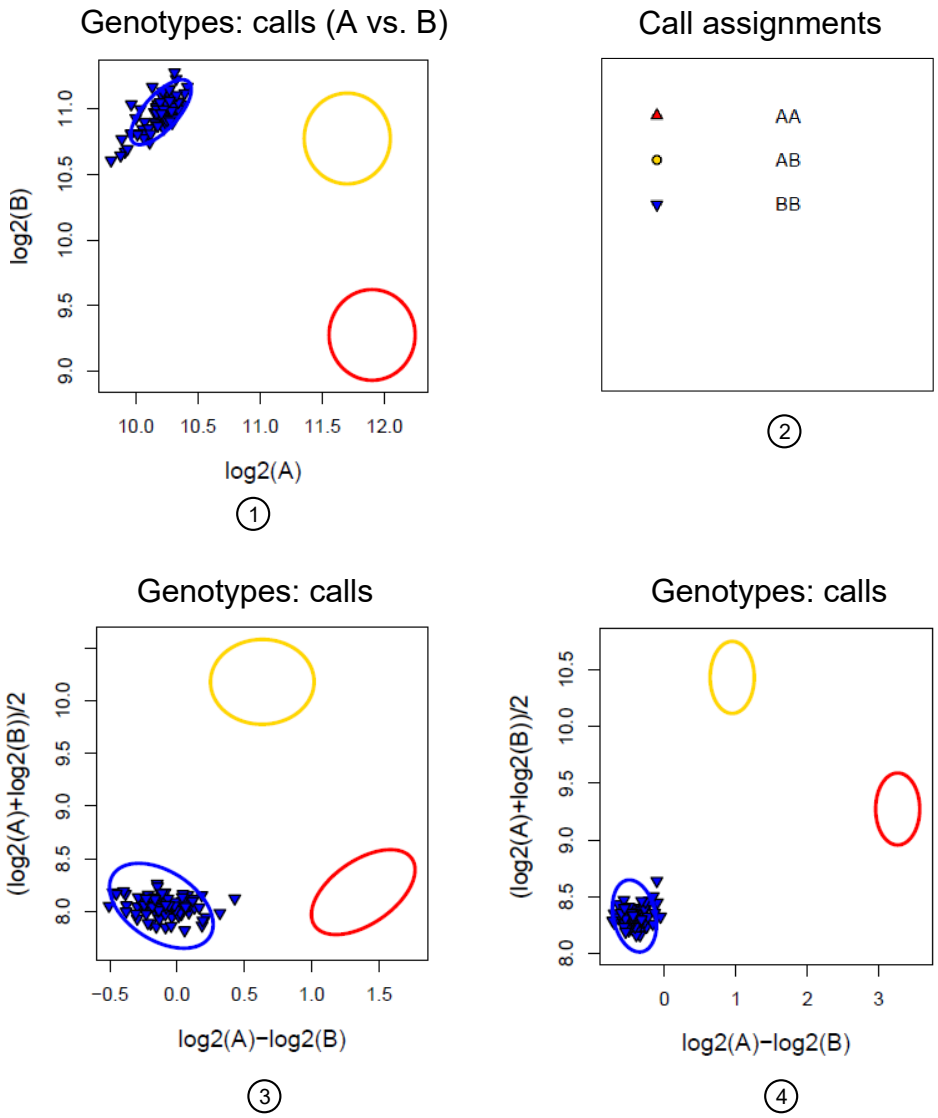


Figure 78 A multiallelic SNP that has one cluster.

1. A multiallelic probeset with one cluster. The A allele is T and the B allele is G.
2. Legend displaying the assigned shapes and colors.
3. A biallelic probeset with one cluster. The A allele is A and the B allele is G.
4. A biallelic probeset with one cluster. The A allele is C and the B allele is G.

Figure 78 shows three probesets for a multiallelic SNP: one multiallelic probeset and two biallelic probesets. The multiallelic probeset and legend are shown in sub-figures 1 and 2. The multiallelic probeset is classified as MonoHighResolution and has detected only one allele, so the more complex BestandRecommended probeset selection is performed. The two biallelic probesets are also monomorphic, but they are both classified as Other. All three probesets agree on the same genotype call for the cluster. While the multiallelic probeset has fewer than 3 alleles, it also has higher quality than the two biallelic probesets so it is selected as the BestandRecommended probeset for this multiallelic SNP.

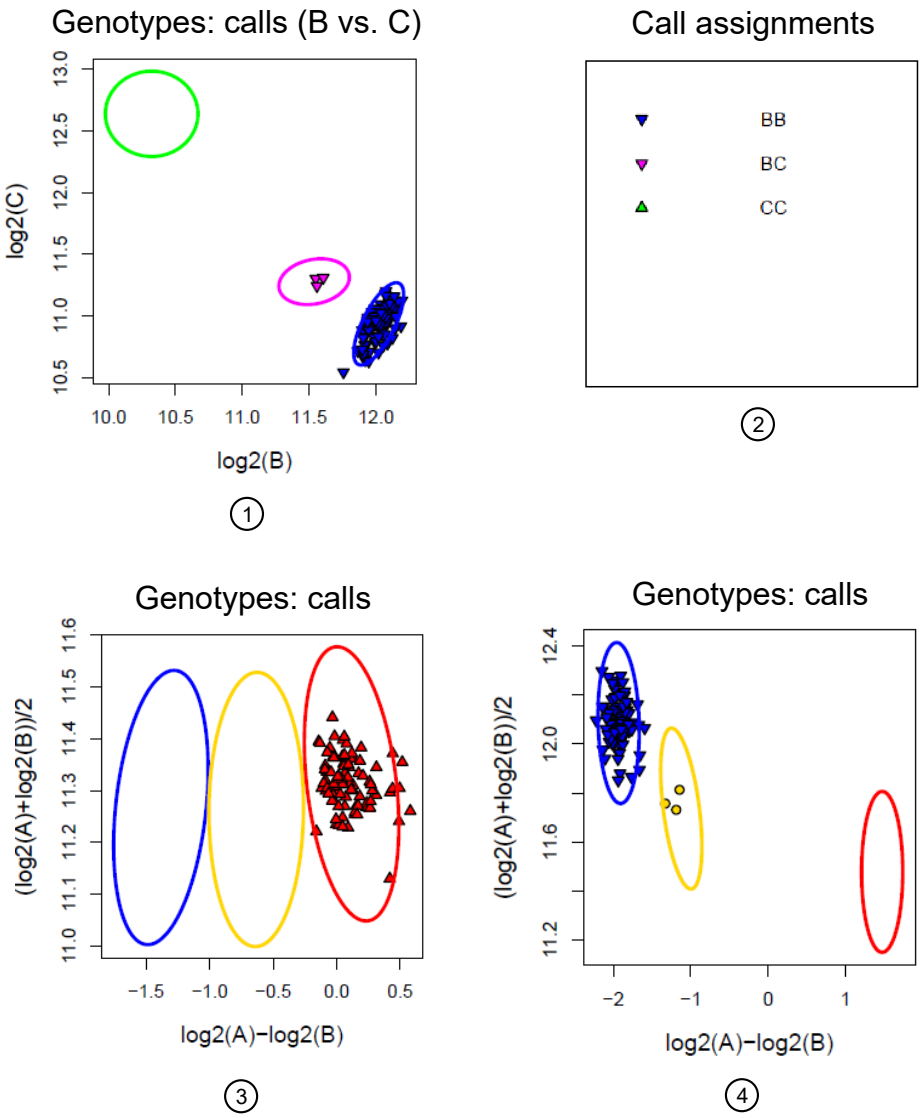


Figure 79 A multiallelic SNP that has two clusters.

① A multiallelic probeset with two clusters. The B allele is T and the C allele is G.
② Legend displaying the assigned shapes and colors.
③ A biallelic probeset with one cluster. The A allele is T and the B allele is C.
④ A biallelic probeset wtih two clusters. The A allele is G and the B allele is T.

Figure 79 shows three probesets for a multiallelic SNP: one multiallelic probeset and two biallelic probesets. The multiallelic probeset and legend are shown in sub-figures 1 and 2; it is classified as NoMinorHom and has detected two one alleles, so the more complex BestandRecommended probeset selection is performed. The two biallelic probesets have 1 and 2 clusters and are classified as Other (sub-figure 3) and OTV (sub-figure 4). While the multiallelic probeset has fewer than 3 alleles, it also has higher quality than the two biallelic probesets so it is selected as the BestandRecommended probeset for this multiallelic SNP.

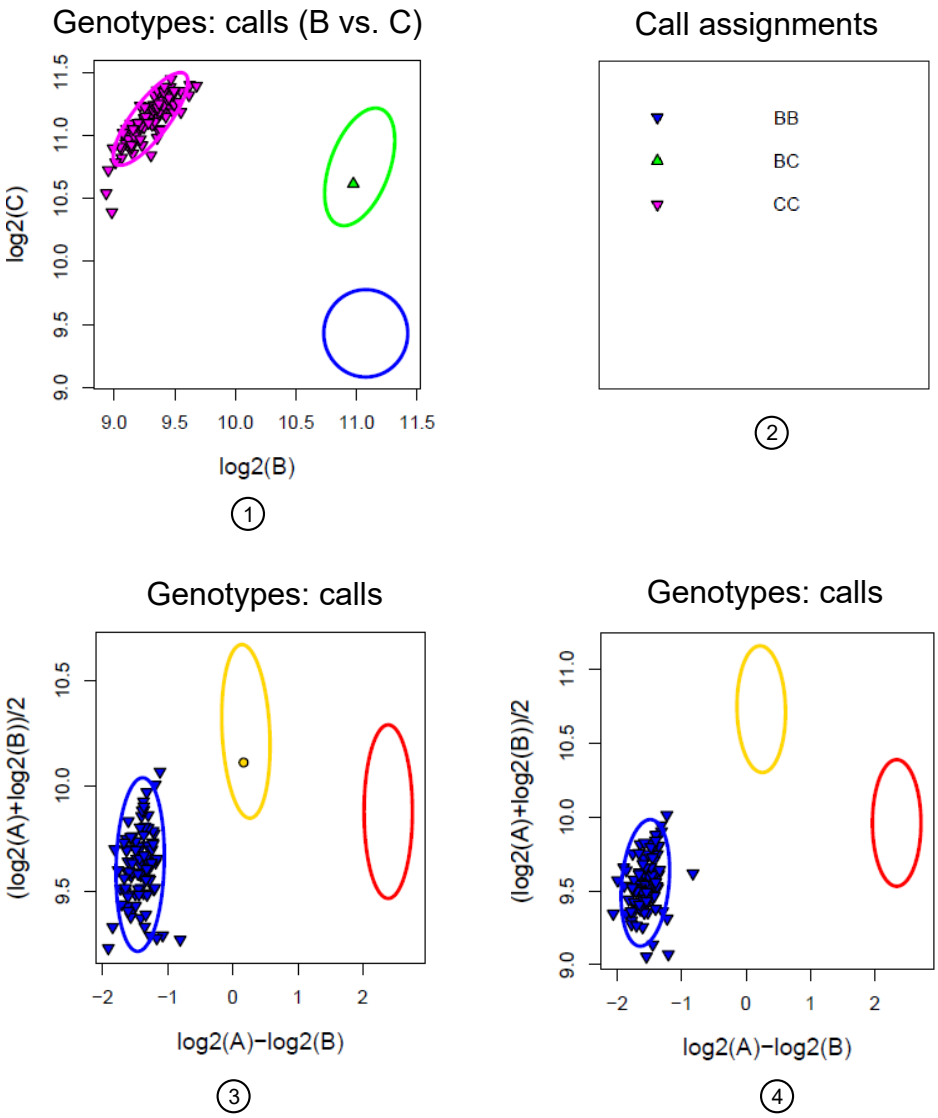


Figure 80 A multiallelic SNP that has two clusters.

1. A multiallelic probeset that has two clusters. The B allele is G and the C allele is T.
2. Legend displaying the assigned shapes and colors.
3. A biallelic probeset that has two clusters. The A allele is G and the B allele is T.
4. A biallelic probeset that has one cluster. The A allele is A and the B allele is T.

Figure 80 shows three probesets for a multiallelic SNP: one multiallelic probeset and two biallelic probesets. The multiallelic probeset and legend are shown in sub-figures 1 and 2; it is classified as NoMinorHom and has detected two one alleles. Because the multiallelic probeset only detected two alleles, the more complex BestandRecommended probeset selection is performed. The two biallelic probesets have 1 and 2 clusters and are classified as NoMinorHom (sub-figure 3) and MonoHighResolution (sub-figure 4). Because the two biallelic probesets are both in recommended categories and the multiallelic probeset detected fewer than three alleles, one of the biallelic probesets will be selected as the BestandRecommended probeset for this multiallelic SNP. The NoMinorHom biallelic probeset is selected over the MonoHighResolution probeset because it has more alleles, so the NoMinorHom

probeset is the new BestandRecommended probeset. The multiallelic probeset's category is updated to Other and the original classification type of NoMinorHom is stored in the column OriginalCT.

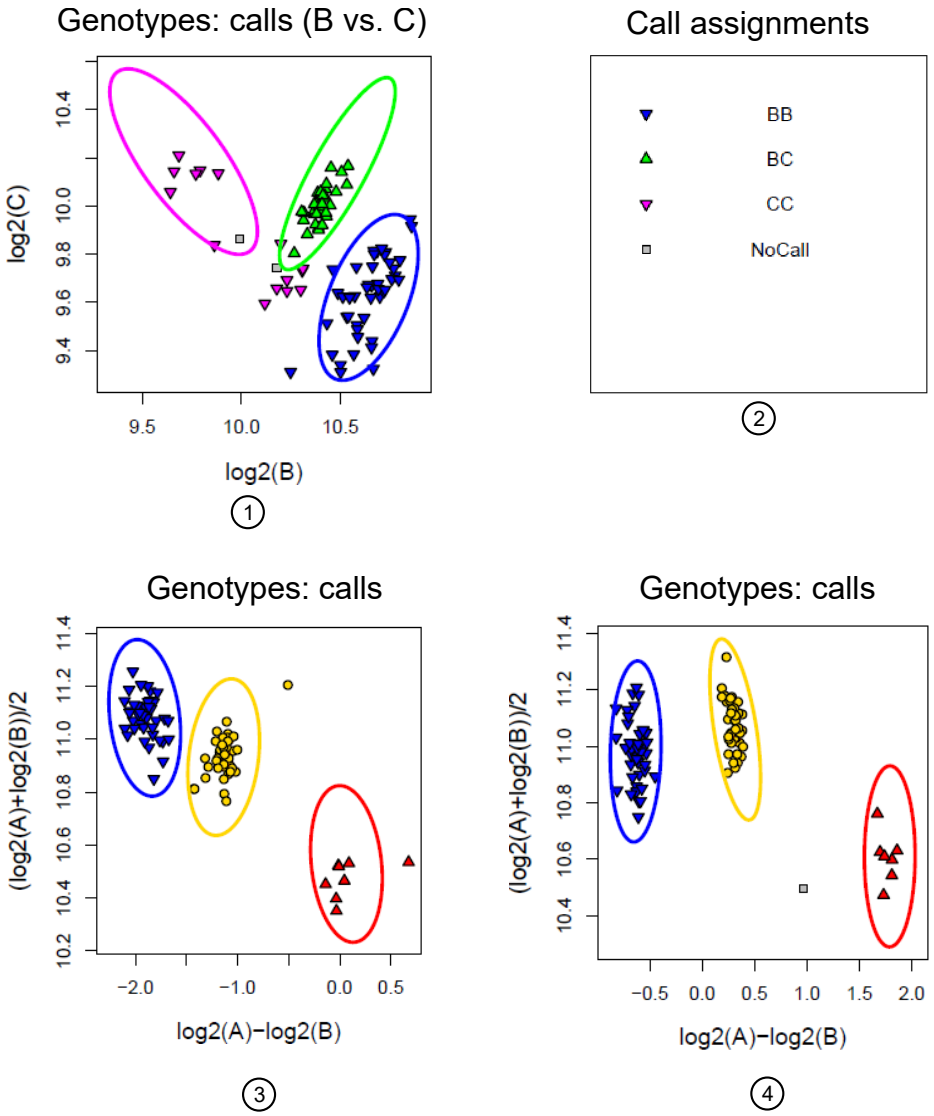


Figure 81 A multiallelic SNP that has three clusters.

1. A multiallelic probeset with three clusters. The A allele is C and the B allele is G.
2. Legend displaying the assigned shapes and colors.
3. A biallelic probeset with three clusters. The A allele is G and the B allele is T.
4. A biallelic probeset with three clusters. The A allele is C and the B allele is G.

Figure 81 shows three probesets for a multiallelic SNP: one multiallelic probeset and two biallelic probesets. The multiallelic probeset and legend are displayed in sub-figures 1 and 2; it is classified as Other, and has detected two one alleles and does not pass all 3 multiallelic thresholds, so the more complex BestandRecommended probeset selection is performed. Both biallelic probesets are categorized as PolyHighResolution, which indicates contradictory genotype assignments. The same samples in the different heterozygous clusters have been

assigned multiple different genotypes: GT in the first biallelic probeset (sub-figure 3) and CG in the second biallelic probeset (sub-figure 4). It is not clear how to discern the true genotypes. The multiallelic probeset was not a BestandRecommended probeset because it was not recommended and either of the biallelic probesets cannot be the BestandRecommended probeset because of the contradictory genotyping. This multiallelic SNP does not have a BestandRecommended probeset.

C Dual workflow

About the Dual Workflow

The Dual Workflow can be used to recover some genotype calls from samples that fail sample QC in the Best Practices workflow. This workflow should never be applied in circumstances in which the genotype calls might influence important decisions, since lower sample quality or processing issues will cause the calls to be less accurate than those from passing samples. The workflow is appropriate for individually unimportant samples or samples for which lower accuracy is acceptable, such as in some agricultural biology applications. In some cases, it may be appropriate to apply the workflow to rare samples that have been exhausted in the assay.

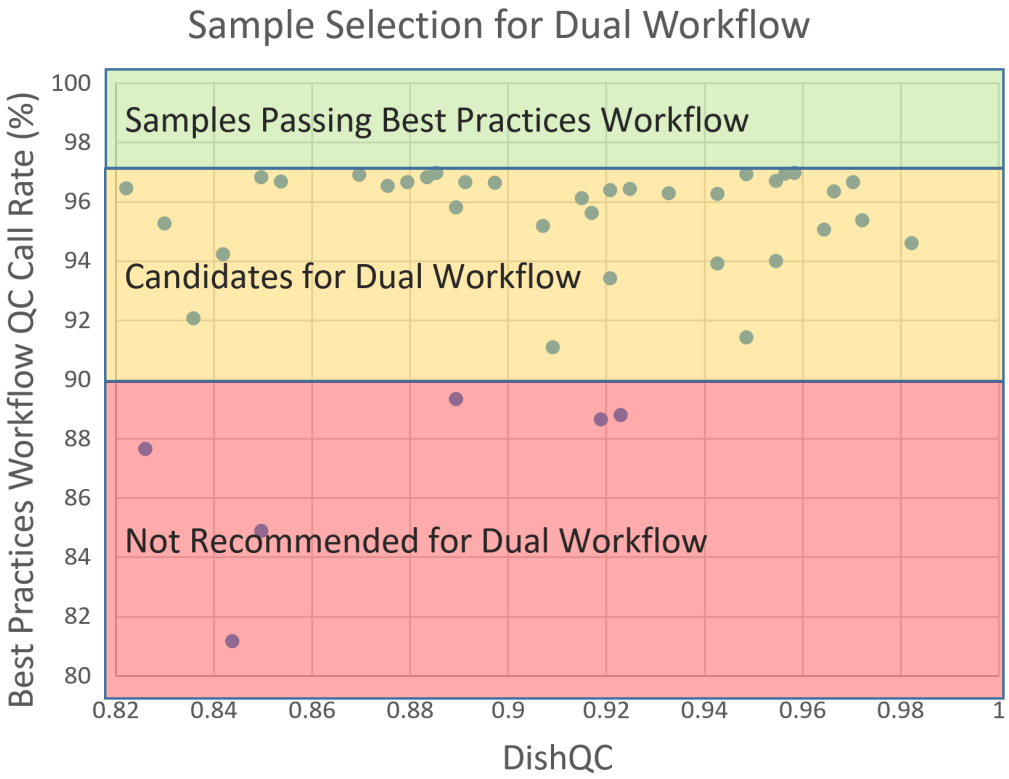


Figure 82 Sample selection guidelines for the Dual Workflow.

Samples should pass the DQC threshold (0.82 for most Axiom™ arrays) and have a call rate below the passing threshold for the Best Practices Workflow (97% for most Axiom™ arrays) but greater than 90%.

The Dual Workflow has been implemented in Axiom™ Analysis Suite version 5.0.

Note: The Dual Workflow requires an updated library file package.The Best Practices Workflow must be run using the updated library file package prior to the Dual Workflow analysis. To determine if a user has an updated library package or to request an updated library package, please contact the local support team.

Steps in the Dual Workflow

The Dual Workflow consists of the following steps.

1. Execute the Best Practices Workflow as described in Chapter 3, “Best Practices Genotyping Analysis Workflow”. The Dual Workflow is intended for use with single-plate batches of samples but may also be applied to larger batches.
2. Open batch in Viewer.
3. In the **Sample Table** tab, select samples for the **Dual Workflow**, according to the following rules and procedures (see Figure 83).
   a. The intended analysis must be appropriate for less-accurate results for these samples.
   b. The sample(s) were assayed on a plate that passes plate QC (see “Step 6: QC the plates” on page 26) and has a minimum of 75% passing samples.
   c. The samples have a passing DQC single-sample QC metric (usually ≥ 0.82).
   d. The samples have a QC Call Rate below passing (usually 97%) but ≥ 90% (as shown in Figure 83
   e. Highlight to select samples that pass cut offs listed above.
   f. In **Reanalyze** drop down, select **Best Practices with Dual Workflow**.

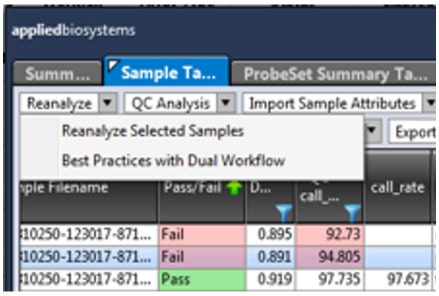
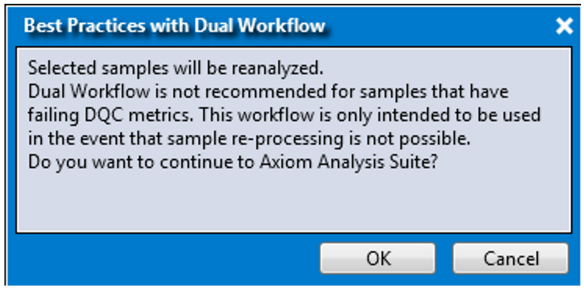


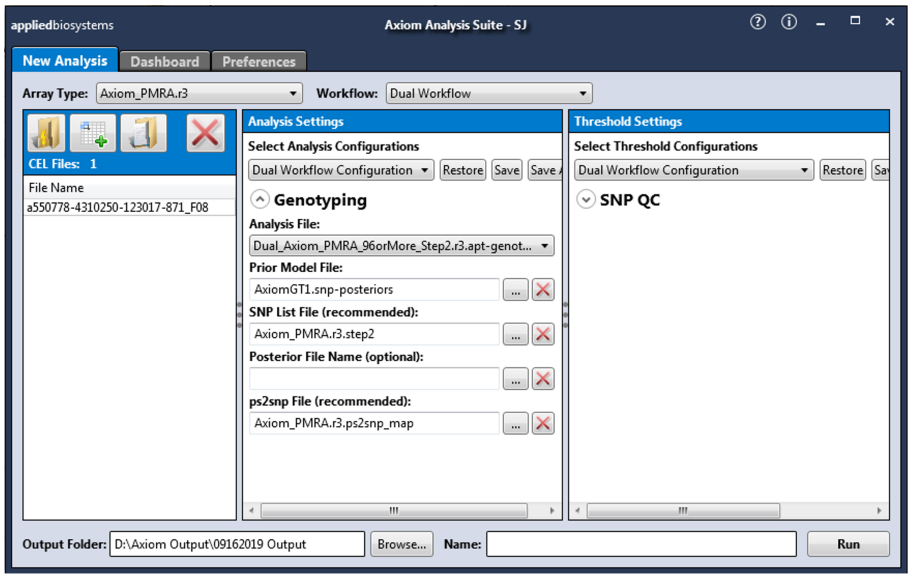
Figure 83 Reanalyze drop down list box.

A warning message appears.



g. Click **OK** to proceed.

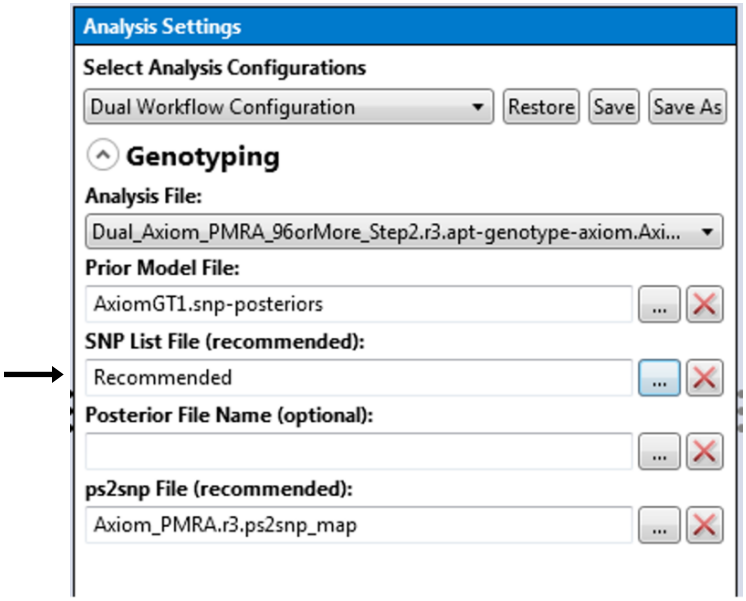
A **New Analysis** window appears.



The Dual Workflow Analysis Settings have been modified.

- Dual Workflow Configuration is automatically selected.
- The Dual Workflow Configuration contains the Genotyping section with the following changes:
  - The Analysis File for Dual Workflow has updated parameters for the following:
    - Genotyping mode has been changed to Single Sample.
    - The confidence threshold has been updated to 0.01, increasing stringency to the genotyping calls.
  - The **Prior Model File** section is automatically populated to AxiomGT1.snp-posteriors.txt (the posteriors from the Best Practices workflow.

- **IMPORTANT!** The **SNP List File** contains the Step2 probeset list. This needs to be changed to the **Recommended.ps** from the Best Practices batch.



1. Click the browse to button.
2. In the popup window, navigate to the SNPolisher folder of your Best Practices batch.
   For example,
   `C:\Users\Public\Documents\AxiomAnalysisSuite\Output\BestPracticeBatch\SNPolisher`
3. Select Recommended.ps and click **OK**.
   The SNP List File now shows your recommended.ps list.

4. Name the batch, then click **Run**.
   The selected samples generate a separate batch that can be opened in the Viewer.

5. Filter samples according to their call rate over best-and-recommended probesets.
   A threshold of 80% is recommended for accuracy of the resulting genotype calls. The relationship of call rate to concordance is shown in Figure 84.

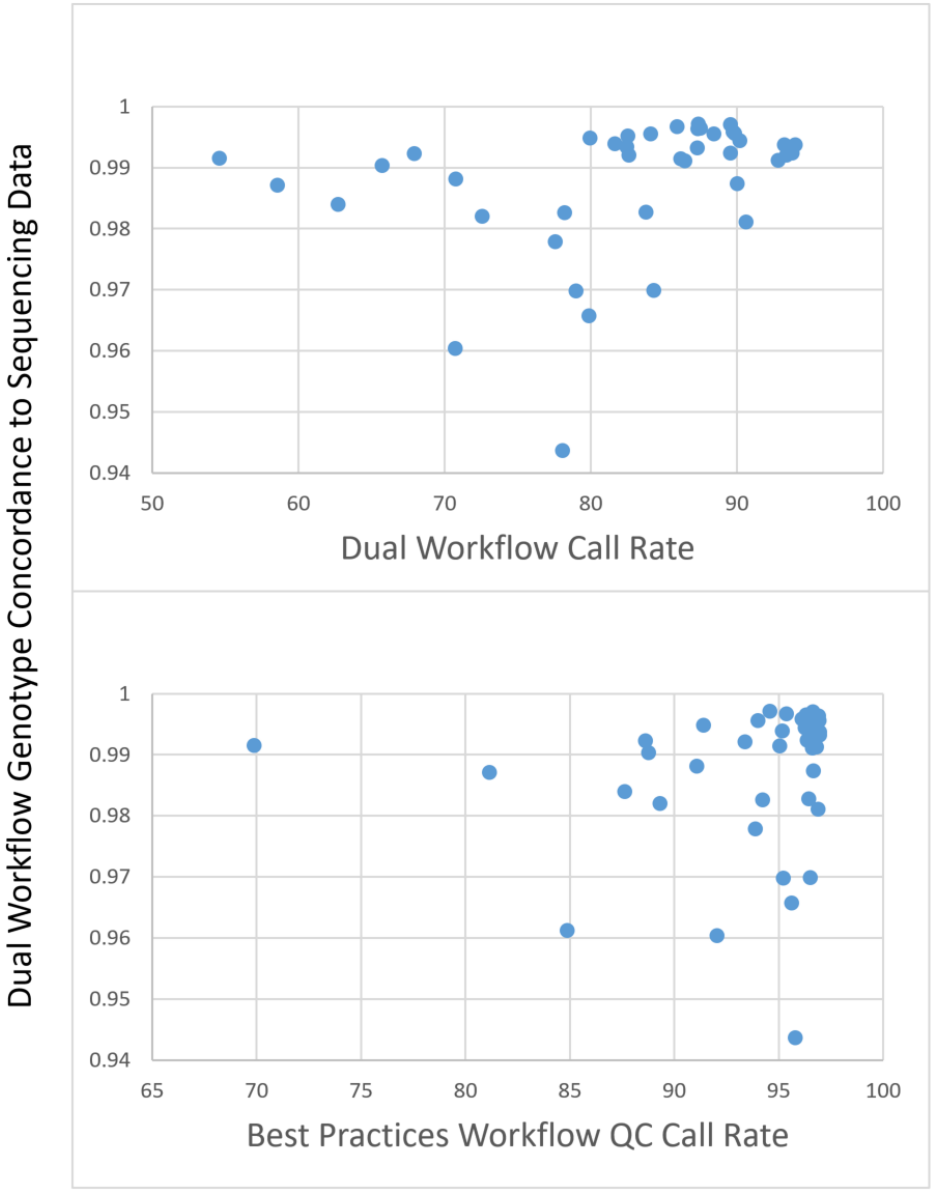


Figure 84 Concordance of Dual Workflow samples vs. two measures of call rate.

42 cell-culture samples with reference genotypes from sequencing were used. *All samples failed Best Practices Workflow QC Call Rate*. Genotyping was carried out in single-plate batches according to the Dual Workflow. (A) Concordance to sequencing data (~530k markers) vs. Dual Workflow call rate (call rate calculated over the list of best-and-recommended probesets from the Best Practices Workflow processing of the same plate); (B) Concordance to sequencing data (~530k markers) vs. Best Practices QC Call Rate.

Documentation and support

Related documentation

Document	Publication number	Description
Axiom™ Analysis Suite User Guide	703307	The Axiom™ Analysis Suite software integrates single nucleotide polymorphism (SNP) genotyping, insertion/deletion (indel) detection, and off-target variant (OTV) calling of simple and complex genomes in an easy-to-use graphical interface.
SNPolisher™ Package User Guide	MAN0017790	The SNPolisher™ Package is an R package that provides advanced analysis, and visualization tools.
Applied Biosystems™ Array Power Tools	http://media.affymetrix.com/support/developer/powertools/changelog/index.html	

Customer and technical support

Visit **thermofisher.com/support** for the latest service and support information.

- Worldwide contact telephone numbers
- Product support information
  - Product FAQs
  - Software, patches, and updates
  - Training for many applications and instruments
- Order and web support
- Product documentation
  - User guides, manuals, and protocols
  - Certificates of Analysis
  - Safety Data Sheets (SDSs; also known as MSDSs)

    Note: For SDSs for reagents and chemicals from other manufacturers, contact the manufacturer.

Limited product warranty

Life Technologies Corporation and/or its affiliate(s) warrant their products as set forth in the Life Technologies' General Terms and Conditions of Sale at **www.thermofisher.com/us/en/home/global/terms-and-conditions.html**. If you have any questions, please contact Life Technologies at **www.thermofisher.com/support**.

References

Affymetrix (2007) BRLMM-P: a genotype calling method for the SNP 5.0 array. Technical Report.

Baker M (2010) Genomics: The search for association. *Nature* 467(7319):1135-8.

Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. *Lancet* 361(9357):598-604.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 37(11):1243-6.

de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 17(R2):R122-8.

Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, de Villena FP, Churchill G (2012) Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics* 13:34.

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS (2010) GENEVA Investigators. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol* 34(6):591-602.

Manolio TA, Collins FS (2009) The HapMap and genome-wide association studies in diagnosis and therapy. *Annu Rev Med* 60:443-56.

Pluzhnikov A, Below JE, Tikhomirov A, Konkashbaev A, Nicolae D,Cox NJ (2008) Differential bias in genotype calls between plates due to the effect of a small number of lower DNA quality and/or contaminated samples. *Genet Epidemiol* 32:676.

Voorrips RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12:172.

Zondervan KT, Cardon LR (2007) Designing candidate gene and genome-wide case-control association studies. *Nat Protoc* 2(10):2492-501.