Contents

- CHAPTER 1 Introduction to Axiom™ data analysis ................. 7
  - About this guide .................................................................... 7
  - Purpose ............................................................................... 7
  - Prerequisites ......................................................................... 7
  - Support ................................................................................. 7
  - Analysis software ................................................................... 7
  - Introduction ........................................................................... 11

- CHAPTER 2 Background ...................................................... 12
  - Axiom™ array terminology .................................................. 12
  - Marker .................................................................................. 12
  - What is a SNP cluster plot for AxiomGT1 genotypes? ......... 13

- CHAPTER 3 Best Practices Genotyping Analysis Workflow ...... 16
  - Design the study to avoid experimental artifacts ................. 16
  - Execute the required steps of the workflow ......................... 17
    - Step 1: Group sample plates into batches ....................... 18
    - Step 2: Generate sample “DQC” values ......................... 19
    - Step 3: QC the samples, based on DQC ......................... 19
    - Step 4: Generate sample QC call rate using Step1.AxiomGT1 ........................................ 19
    - Step 5: QC the samples based on QC call rate ............... 19
    - Step 6: QC the plates ......................................................... 20
    - Step 7: Genotype passing samples and plates over Step2.AxiomGT1 SNPs .................. 22
    - Step 8: Execute SNP QC .................................................... 22
      - Step 8A: Create SNP QC metrics .................................. 23
      - Step 8B: Classify SNPs using QC metrics ...................... 23
      - Step 8C: Create a recommended SNP list ..................... 26
  - Evaluate SNP cluster plots ................................................. 29
    - Well-clustered vs mis-clustered SNP cluster plot patterns .... 29
    - Multi-cluster SNP cluster plot patterns ......................... 31
    - Allopolyploid SNP cluster plot pattern ......................... 33
    - SNP cluster plot patterns for inbred populations ............. 34
CHAPTER 4  Additional genotyping methods  .................................................. 35
Manually change genotypes  ........................................................................ 35
Adjust genotype calls for OTV SNPs  ....................................................... 35
Genotyping autotetraploids  ...................................................................... 36
Increase the stringency for making a genotype call  .................................. 37
Genotyping inbred samples  ...................................................................... 37
  Identifying if an inbred penalty is needed  ............................................ 37
  How to use the inbred penalty setting  .................................................. 38
Axiom™ Analysis Suite  .......................................................................... 38
APT  ....................................................................................... 39

CHAPTER 5  Additional sample and plate QC  ............................................. 40
Additional sample QC  ............................................................................. 40
Detecting sample mix-ups  ...................................................................... 40
Unusual or incorrect gender calls  ............................................................. 40
Genotyping gender call process: cn-probe-chrXY-ratio_gender  ............. 40
Detecting mixed (contaminated) DNA samples  ....................................... 41
  Samples have relatively high DQC and low QC Call Rate (QCCR) values   41
  Samples have a high percentage of unknown gender calls  ...................... 42
  Samples tend to fall between the genotype clusters formed by the
    uncontaminated samples  .................................................................. 42
  Unusual patterns of relatedness  ............................................................. 43
  Increased computed heterozygosity  ..................................................... 43
Additional plate QC  .............................................................................. 43
Evaluate pre-genotyping performance with DQC box plots  ..................... 43
Monitor plate controls  ........................................................................... 44
Check for platewise MAF differences  ..................................................... 45

CHAPTER 6  SNP QC metrics  ................................................................. 46
SNP metrics used in the Ps_Classification Step (Step 8C)  ......................... 46
  SNP Call Rate (CR)  ........................................................................... 46
  Fisher’s Linear Discriminant (FLD)  ...................................................... 47
  Heterozygous Strength Offset (HetSO)  ............................................... 48
  Homozygote Ratio Offset (HomRO)  .................................................. 50
Additional SNP metrics that may be used for SNP filtering  ....................... 51
  Hardy-Weinberg p-value  ..................................................................... 51
  Mendelian Trio Error  ......................................................................... 52
  Genotyping call reproducibility  ......................................................... 52
CHAPTER 7 Executing Best Practices steps with Axiom™ Analysis Suite .......................... 53

Execute steps 1-8 with Axiom™ Analysis Suite .......................................................... 53
Axiom™ Analysis Suite setup ................................................................. 53
Step 1: Group samples into batches .......................................................... 54
Setup step 2, 3, 5, 6 and 8A, B: Set Sample Metrics, Plate Metrics, and SNP Metrics ... 54
Step 4 and 7: Generate sample QC call rate using Step1.AxiomGT1 and genotype passing samples and plates over Step2.AxiomGT1 SNPs ........................................ 56
Run analysis and Review data ................................................................. 57
Visualize SNPs and Change Calls through Axiom Analysis Suite cluster graphs ... 60
Display a particular SNP ................................................................. 62
Select a single sample ................................................................. 62
Select multiple samples ................................................................. 63
Manually change a sample’s call .......................................................... 64
Lasso function ................................................................. 65
Saving a cluster plot ................................................................. 66
Step 8C: Create a recommended SNP list .................................................. 67
Running OTV Caller or Classification Supplemental ........................................ 68
Exporting data from Axiom™ Analysis Suite ........................................... 69

CHAPTER 8 Executing Best Practices steps with command line software ....................... 70

Execute Best Practice steps 1-7 with APT software ............................................. 70
Best Practices Step 1: Group samples into batches ............................................. 70
Best Practices Step 2: Generate the sample “DQC” values using APT .................... 71
Best Practices Step 3: Conduct sample QC on DQC ............................................. 71
Best Practices Step 4: Generate sample QC call rates using APT .......................... 71
Best Practices Step 5: QC the samples based on QC call rate in APT .................... 72
Best Practices Step 6: QC the plates .......................................................... 72
Best Practices Step 7: Genotype passing samples and plates using AxiomGT1.Step2 ........................................................................................................... 73
Best Practices Step 8A: Run Ps-Metrics ......................................................... 74
Best Practices Step 8B: Run Ps_Classification .................................................... 75
Visualize SNP cluster plots with SNPolisher Ps_Visualization function .................. 76
Contents

Documentation and support .............................................. 79
Related software documentation ......................................... 79
Customer and technical support ......................................... 79
References ........................................................................ 80
1 Introduction to Axiom™ data analysis

About this guide

Purpose
This guide provides information and instructions for analyzing Axiom™ genotyping array data. It includes the use of Axiom™ Analysis Suite, Applied Biosystems™ Power Tools (APT) and SNPolisher R 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 intended 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
Users should contact their 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 above, (2) and Applied Biosystems™ Power Tools (APT) version 1.18 and above, (3) the SNPolisher R package version 1.5.0 and above. The workflow utilizing these software systems is shown in the section "Execute the required steps of the workflow” on page 17.

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 a user 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 (Applied Biosystems Power Tools information). APT programs are intended for “expert users” who prefer programs that can be utilized 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, please refer to the Axiom Analysis Suite software user manual, and the APT help.
SNPolisher R functions provide visualization tools and advanced genotyping methods. All functions necessary for best practices are incorporated into Axiom Analysis Suite and APT. Usage of SNPolisher functions requires the user to have some familiarity with the programming language R. The R package files to install SNPolisher are available on the Thermo Fisher Scientific website (www.thermofisher.com).

**Download SNPolisher installation files**

SNPolisher is available in the **Analysis Tools** section of the DevNet Tools page. You must be registered with Thermo Fisher Scientific to download the necessary files.

1. If necessary, select **Register** from the **Sign In** drop-down box 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 page 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 four folders with example data for running in R. Note that this zipped folder is not a package binary for installing in R. Users must unzip the file to extract the SNPolisher folder, which contains the tar.gz package file. For instructions on R basics, installation, and usage of the R functions, (including additional function not discussed in this document), see the SNPolisher User Guide.

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** lists the names of all analysis files used to process Axiom genotyping arrays in Axiom Analysis Suite, or APT. Some arrays will have more files than those listed in the table in their library file package. An *annotation* file is an additional file 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.
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.

<table>
<thead>
<tr>
<th>Analysis library files</th>
<th>Axiom Analysis Suite</th>
<th>APT</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;axiom_array&gt;.analysis_settings</code></td>
<td>Required</td>
<td>N/A</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.ax_package</code></td>
<td>Required</td>
<td>N/A</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.r&lt;#&gt;.ps2snp_map.ps</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;_96orMore_Step1.r&lt;#&gt;.apt-genotype-axiom.AxiomGT1.apt2.xml</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;_GenericPriors.r&lt;#&gt;.apt-genotype-axiom.AxiomGT1.apt2.xml</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;_96orMore_Step2.r&lt;#&gt;.apt-genotype-axiom.AxiomGT1.apt2.xml</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;_LessThan96_Step1.r&lt;#&gt;.apt-genotype-axiom.AxiomGT1.apt2.xml</code></td>
<td>Required for small sample size</td>
<td>Required for small sample size</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;_SNPSpecificPriors.r&lt;#&gt;.apt-genotype-axiom.AxiomGT1.apt2.xml</code></td>
<td>Required for small sample size</td>
<td>Required for small sample size</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.r&lt;#&gt;.cdf</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.r&lt;#&gt;.qca</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.r&lt;#&gt;.qcc</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.r&lt;#&gt;.step1.ps</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.r&lt;#&gt;.generic_prior.txt</code></td>
<td>Required</td>
<td>Required</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.r&lt;#&gt;.AxiomGT1.sketch</code></td>
<td>• Required for human genomes • Optional for non-human genomes</td>
<td>• Required for human genomes • Optional for non-human genomes</td>
</tr>
<tr>
<td><code>&lt;axiom_array&gt;.r&lt;#&gt;.chrXprobes</code></td>
<td>• Required for mammalian genomes • N/A for non-mammalian genomes</td>
<td>• Required for mammalian genomes • N/A for non-mammalian genomes</td>
</tr>
</tbody>
</table>
### 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.apt.xml= Axiom_BioBank1_96orMore_Step2.r2.apt-genotype-axiom.AxiomGT1.apt.xml.  (Continued)

<table>
<thead>
<tr>
<th>Analysis library files</th>
<th>Axiom Analysis Suite</th>
<th>APT</th>
</tr>
</thead>
</table>
| <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 |
| <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 |

**Note:** Library files for some Axiom arrays may use alternative names for the Step1 and Step2.xml files:
- apt-probeset-genotype.AxiomGT1.apt.xml
- apt-axiom-genotype.AxiomGT1.apt.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 discovered through human GWAS has been steadily increasing, most of the variants detected to date have small effects, and very large sample sizes have been required to identify and validate these findings (Manolio, TA 2009; de Bakker PI, et al. 2008; Baker M. 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 CC, 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 (human, plant, and animal) Axiom™ Genotyping Arrays. The Axiom™ Genotyping Solution produces calls for both SNPs and indels (insertions/deletions). For simplicity, in this document, the term SNPs will refer to both SNPs and indels. Additional chapters in the document include:

- **Chapter 2, "Background"** provides information that is needed for understanding the rest 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 as well as 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 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).
- **Chapter 6, "SNP QC metrics"** describes metrics that are used in the Best Practices workflow (Chapter 3) for SNP classification as well as additional metrics used in the field for SNP QC.
- **Chapter 7, "Executing 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 with APT is provided in this chapter.
Background

Axiom™ array terminology

Marker

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 probe set.

Most Axiom markers are interrogated with one or two probe sets, 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 “probe set used to interrogate the SNP or indel”.

Figure 1  Illustration of a marker interrogated by forward and reverse strand probe sets.
What is a SNP cluster plot for *AxiomGT1* genotypes?

A SNP cluster plot corresponds to one probe set, designed to interrogate a given SNP; and each point corresponds to one sample whose A and B allele array intensities have been transformed into the X vs Y coordinate space 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 3) and (2) the SNPolisher package, via the *Ps_Visualization* function (example shown in Figure 2). Instructions for the *Cluster Graph* function usage are provided in Chapter 7, and instructions *Ps_Visualization* function usages are provided in Chapter 8.

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

*AxiomGT1* clustering is carried out in two dimensions, dimension Y is calculated as \([\log_2(A_{\text{signal}}) + \log_2(B_{\text{signal}})]/2\) and dimension X is calculated as \(\log_2(A_{\text{signal}}/B_{\text{signal}})\). X carries the main information for distinguishing genotype clusters. The X dimension is called *Contrast* and the Y dimension is called *Size* in cluster plots produced SNPolisher and Axiom Analysis Suite.

*AxiomGT1* genotype calls are made by identifying the genotype intensity distribution (AA, AB, or BB) each sample is most likely to belong to. The samples are colored and shaped by these *AxiomGT1* genotype calls. The Axiom Analysis Suite *SNP Cluster Graph* and SNPolisher *Ps_Visualization* defaults are set to have BB calls as blue upside down triangles, AB calls as gold circles, AA calls as red triangles. Note, in Axiom Analysis Suite it is possible to color and shape the data according to other sample attributes, which are shown in the legend for the graphs.

*AxiomGT1* genotype *NoCalls* 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 Confidence Score Threshold, the genotype call for the sample is converted to a NoCall. Axiom Analysis Suite *SNP Cluster Graph* and SNPolisher *Ps_Visualization* defaults are set to have No Calls as gray squares.

The *AxiomGT1* cluster variances are used to create ellipses around the cluster means in the SNP cluster plots. Ellipses based on priors are dashed and ellipses based on posteriors are solid for all cluster plots.

Unless specified otherwise, all cluster plots in the document have been produced by *Ps_Visualization* and use the sample colors and shapes as described above.
Chapter 2 Background

What is a SNP cluster plot for AxiomGT1 genotypes?

Figure 2  SNP cluster plot produced by the SNPolisher package, via the 
*Ps_Visualization* function.

\[
\begin{align*}
X(\text{Contrast/LogRatio}) &= \log_2(A_{\text{signal}}/B_{\text{signal}}) \\
Y(\text{Size/Strength}) &= [\log_2(A_{\text{signal}}) + \log_2(B_{\text{signal}})]/2
\end{align*}
\]
Figure 3  SNP cluster plot produced by Axiom Analysis Suite via *SNP Cluster Graph* function.
Design the study to avoid experimental artifacts

Good experimental design practices (Pluzhnikov A, et al. 2008; Cardon LR, and Palmer LJ, 2003; Clayton DG, et al. 2005; Zondervan KT, and Cardon LR, 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 (e.g., 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 may 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 there is no population substructure present that could confound the analysis of data from cases and controls (e.g., all of the controls are CEU, while the cases are YRI). For non-human samples the same principles apply, and samples should be randomized with regards to breeds, species, and subpopulations for genome under study. In addition, researchers should ensure their experiments are sufficiently powered to answer the question of interest. Again, it is best to examine all of these questions prior to the initiation of the project.

For a non-ideal study design, for which cases and controls are not randomized, the SNPolisher package provides the BalleleFreq_Test function to identify and remove SNPs with inconsistent genotypes due to shifts in intensity in probe sets across samples that were processed in the separate case and control batches. See the Ballele_Freq_Test in the SNPolisher User Guide.
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 4).

**Step 1:** Group samples into batches. For each batch, perform the following:

- **Step 2:** Generate Sample DQC values
- **Step 3:** QC samples based on DQC values
- **Step 4:** Generate sample QC call rate
- **Step 5:** QC samples based on QC call rate over QC SNPs in the step1.Axiom GT1 probe set list
- **Step 6:** QC the plates
- **Step 7:** Genotype passing samples & plates over recommended SNPs in the step2.Axiom GT1 probe set list
- **Step 8:** QC the SNPs and sort into six SNP categories
- **Step 9 (as needed):** OTV caller and Supplemental analysis for further classification

**Legend:**
- Step completed in Axiom Analysis Suite.

Note: APT 1.16.0 or higher will generate all appropriate QC metrics but Sample filtering in Steps 3 and 5 must be performed with Excel or R script in Windows or Linux environment.

**Figure 4** 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. 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.
Step 1: Group sample plates into batches

In general, group plates in as large a batch size as is computationally feasible, or up to 4800 samples (50 96-array plates or 12 384-array plates). Group plates in the order in which the plates were processed (e.g., if using batches of 8 plates, it is usually preferable to group together the first 8, the second 8, etc.). The minimum batch size when using generic priors is 96 samples comprising 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 chosen in all workflows. Each batch should contain either 15 or more distinct female samples or zero female samples. In other words, 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. Users may attempt to co-cluster plates with such differences, but plate QC guidelines ("Step 6: QC the plates" on page 20, and "Additional plate QC" on page 43) must be followed carefully.
- DNA samples 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 comprises “greatly divergent” is a gray area and may 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 4800 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, as well as achieving reasonable computation performance of the command line analysis programs (APT and SNPPolisher, see Chapter 8) with the system that will analyze the array plate batches. 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 4) using the apt-genotype-axiom command ("Best Practices Step 7: Genotype passing samples and plates using AxiomGT1.Step2" on page 73) on a Linux server with the following configuration: x86_64 architecture, 16 x 3GHz XEON core, and 128 GB of RAM. Note that this is without any computational parallelization.
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 (i.e., 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 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:

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

Distributions of contrast values are computed 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 will be 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.

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". These samples should be either reprocessed in the laboratory or dropped from the study. The default DQC threshold value is 0.82 for all Axiom arrays except Axiom_BOS1 which is 0.95.

Not all problematic samples are detectable by the DQC metric prior to the first round of genotyping (see "Detecting mixed (contaminated) DNA samples" on page 41). 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 requests the user to provide at least a plate of Axiom data to identify probe sets that meet this criteria. Thermo Fisher will then provide the Axiom Analysis Library package (Table 1) for the array. Users should contact their local Thermo Fisher Scientific Field Application Scientist or thermofisher.com/support when such data is available.

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

Samples with a QC call rate value less than the default threshold of 97% 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 40 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 computed 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, may contribute to mis-clustering events (described in "Evaluate SNP cluster plots on page 29), whether processed separately or processed with all other plates in the batch. These differences may manifest themselves as putative differences in the MAF of SNPs over these samples relative to the rest 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 may 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 will decrease the power and/or increase false positive rates.

The metrics and guidelines for plate performance are as follows:

**Metrics:**

- **Plate pass rate** = \( \frac{\text{Samples passing DQC and 97% QC call rate}}{\text{Total samples on the plate}} \times 100 \)

- **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 \( \geq 95\% \) for samples derived from tissue, blood or cell line, and \( \geq 93\% \) if sample source is saliva

- Average QC call rate of passing samples \( \geq 99\% \)

**Guideline for Passing Plates**

- Average QC call rate of passing samples \( \geq 98.5\% \)
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 5). Ideally all plates in the batch will pass the guidelines for high-quality plates (green zone Figure 5). Passing plates in the gray zone should be reviewed for plate processing problems. If there are no known plate processing problems, the user may proceed with caution to include passing samples from such plates. Low sample pass rates may 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 may be included. All samples on non-passing plates (red zone Figure 5) 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 5  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 hand 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 43 for additional Axiom specific methods and general methods 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. This Best Practices Step 7 is referred to as Step2.AxiomGT1 genotyping in the instructions provided for genotyping with Axiom Analysis Suite (Chapter 7) and APT (Chapter 8).

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

As discussed in “What is a SNP cluster plot for AxiomGT1 genotypes?” on page 13, 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 generally recommended for large batches (>96 samples) when 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 on page 9) for genotyping large batches and some arrays are provided with analysis files for genotyping small batches.

Certain arrays may benefit from usage of SNP-specific priors, even when the sample size is large. These may 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: The 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). This is because only tested, working SNPs are used for Step 4 QC call rates; whereas Step 7 call rates are often computed over untested probe sets 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 29) 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 will be identified. The overall approach is to sort the best probe set per SNP into categories 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 certain types of genomes.

Step 8 uses the Ps_Metrics and Ps_Classification functions. These functions are available in the APT software version 1.18 or greater and fully integrated into Axiom Analysis Suite. Instructions for Axiom Analysis Suite usage see “Setup step 2, 3, 5, 6 and 8A, B: Set Sample Metrics, Plate Metrics, and SNP Metrics” on page 54, and for APT usage see Best Practices Step 8A and 8B on page 74.
Step 8A: Create SNP QC metrics

The \textit{Ps\_Metrics} function is used on the output files from the Best Practices Step 7 genotyping run (also referred to as Step2.AxiomGT1), and computes twelve SNP QC metrics for each probe set (probeset\_id) that was genotyped in Step 7: Call Rate (CR), Fisher’s Linear Discriminant (FLD), HomFLD, Heterozygous Strength Offset (HetSO), Homozygous Ratio Offset (HomRO), minor allele count (nMinorAllele), 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), and a hemizygous indicator (hemizygous). Values for five of these metrics: CR, FLD, HetSO, HomRO, and nMinorAllele form the basis of the SNP classifications (discussed below). The CR, FLD, HetSO, HomRO SNP QC metrics are described in Chapter 6, "SNP QC metrics". SNP QC metrics are used in the \textit{Ps\_Classification} Step (Step 8B).

Additional SNP QC tests used in the field are discussed in "Additional SNP metrics that may be used for SNP filtering" on page 51.

Step 8B: Classify SNPs using QC metrics

The \textit{Ps\_Classification} function is used to sort each probe set into seven classes based on the five SNP QC metrics generated by the \textit{Ps\_Metrics} function. The classes are described in Figure 6. The seven classifications are based on default QC thresholds shown in Table 2 for different genome types. Note, the user can change the thresholds if desired.

The best probe set per SNP is determined by the classification priority order: PolyHighResolution, NoMinorHom, OTV, MonoHighResolution, and CallRateBelowThreshold. For a SNP with two probe sets, where one probe set is NoMinorHom and one probe set is MonoHighResolution, the probe set that has been classified as NoMinorHom will be selected as the best probe set. See the SNPolisher User Guide for more details on the \texttt{priority.order} argument for \texttt{Ps\_Classification}. The file \texttt{<axiom\_array>.r<#>.ps2snp\_map.ps} in the Analysis Library File package (Table 1 on page 9) contains the list of matched probe sets and SNPs.
Figure 6  Cluster plot examples and descriptions of the seven SNP classification categories. OTV SNPs are discussed further in "Adjust genotype calls for OTV SNPs" on page 35.
The *Ps_Classification* function outputs the *Ps.performance.txt* file, which contains the probeset_id’s, affysnp_id’s, QC metrics, hemizygous status, and an indicator if this probe set is the best for the SNP (BestProbe set), and which classification (Figure 6) the probe set belongs to (ConversionType) for each probe set. If all SNPs have one probe set, then every probe set is the best probe set by default. Column Names and Examples are shown below (Table 3).

### Table 2  Default QC thresholds for CR, FLD, HetSO HomRO (metrics defined in Chapter 6) and nMinorAllele (number of minor alleles in the batch). 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 35). 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. For more information see the SNPolisher User Guide.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Human</th>
<th>Diploid</th>
<th>Polyploid</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR</td>
<td>95</td>
<td>97</td>
<td>97</td>
</tr>
<tr>
<td>FLD</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>HetSO</td>
<td>−0.1</td>
<td>−0.1</td>
<td>−0.1</td>
</tr>
<tr>
<td>HetSO.OTV</td>
<td>−0.3</td>
<td>−0.3</td>
<td>−0.3</td>
</tr>
<tr>
<td>HomRO1</td>
<td>0.6</td>
<td>0.6</td>
<td>N/A</td>
</tr>
<tr>
<td>HomRO2</td>
<td>0.3</td>
<td>0.3</td>
<td>N/A</td>
</tr>
<tr>
<td>HomRO3</td>
<td>−0.9</td>
<td>−0.9</td>
<td>N/A</td>
</tr>
<tr>
<td>nMinorAllele</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

The *Ps.Classification* function outputs the *Ps.performance.txt* file, which contains the probeset_id’s, affysnp_id’s, QC metrics, hemizygous status, and an indicator if this probe set is the best for the SNP (BestProbe set), and which classification (Figure 6) the probe set belongs to (ConversionType) for each probe set. If all SNPs have one probe set, then every probe set is the best probe set by default. Column Names and Examples are shown below (Table 3).

### Table 3  *Ps.performance* column names and examples

<table>
<thead>
<tr>
<th>Column name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>probeset_id</td>
<td>AX-11481545</td>
</tr>
<tr>
<td>affy_snp_id</td>
<td>Affx-27771153</td>
</tr>
<tr>
<td>CR</td>
<td>99.2320</td>
</tr>
<tr>
<td>FLD</td>
<td>8.1936</td>
</tr>
<tr>
<td>HomFLD</td>
<td>17.9734</td>
</tr>
<tr>
<td>HetSO</td>
<td>0.4500</td>
</tr>
<tr>
<td>HomRO</td>
<td>2.5716</td>
</tr>
<tr>
<td>nMinorAllele</td>
<td>5123</td>
</tr>
<tr>
<td>Nclus</td>
<td>3</td>
</tr>
<tr>
<td>n_AA</td>
<td>3112</td>
</tr>
<tr>
<td>n_AB</td>
<td>3383</td>
</tr>
<tr>
<td>n_BB</td>
<td>870</td>
</tr>
</tbody>
</table>
The `Ps_Classification` function also selects the best probe sets from the `Ps.performance.txt` file and divides these into seven category files named: `PolyHighResolution.ps`, `NoMinorHom.ps`, `Hemizygous.ps`, `MonoHighResolution.ps`, `CallRateBelowThreshold.ps`, `Other.ps`, and `OffTargetVariant.ps`. Each category file is a 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 convention to indicate the file contains a list of probe set IDs.

**Step 8C: Create a recommended SNP list**

SNPs that are not sorted into recommended classes for the genome type should be excluded from further downstream analysis. Table 4 shows which classes are recommended for the given genome type. SNPs in recommended classes may also be referred to as converted in this document.

**Table 3**  
<table>
<thead>
<tr>
<th>Column name</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_NC</td>
<td>57</td>
</tr>
<tr>
<td>hemizygous</td>
<td>0</td>
</tr>
<tr>
<td>HomHet</td>
<td>0</td>
</tr>
<tr>
<td>ConversionType</td>
<td>PolyHighResolution</td>
</tr>
<tr>
<td>BestProbeset</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 4**  
<table>
<thead>
<tr>
<th>Genome type</th>
<th>SNP Class determined by <code>Ps.Classification</code> Function</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td><strong>Recommended</strong></td>
<td><strong>Recommended</strong></td>
</tr>
<tr>
<td>Diploid-inbred only</td>
<td><strong>Recommended</strong></td>
<td><strong>Not Recommended</strong></td>
</tr>
<tr>
<td>Diploid-outbred or mixture of inbred and outbred</td>
<td><strong>Recommended</strong></td>
<td><strong>Recommended</strong></td>
</tr>
<tr>
<td>Polyploid</td>
<td><strong>Recommended</strong></td>
<td>Requires additional genetic knowledge</td>
</tr>
</tbody>
</table>

MonoHighResolution SNPs are recommended with caution, especially if the best probe set for the SNP site has never been tested. An additional test for recommending MonoHighResolution SNPs is to require that both probe sets (if available on the array)
for the SNP site are classified as MonoHighResolution and that the genotypes agree. Hemizygous SNPs are recommended by default, but visual inspection is advised (see below). SNPs that are classified as OTV may also be considered converted after the OTV Call function has been used to re-label the genotype calls (see “Adjust genotype calls for OTV SNPs” on page 35) and after visual inspection of the recalled genotypes.

A total list of unique probe sets for recommended SNPs can be created manually by combining the category files (described above) for the default recommended (yellow) and/or chosen by the user. Or Ps.Classification can be executed with output.converted = TRUE (the default is FALSE), and PolyHighResolution, NoMinorHom, MonoHighResolution, and Hemizygous classes are combined to create a category file called converted.ps. Therefore converted.ps contains all probe sets, one per SNP, that are recommended for downstream analysis if the Genome/Species Type is human or Diploid-outbred or mixture of inbred and outbred.

**Visual SNP analysis for hemizygous SNPs**

Chromosome Y, W, and mitochondrial and other hemizygous genomes produce only two genotype clusters (i.e., one representing A and one representing B). These two clusters should be easily resolved from one another and so are recommended by default. Thermo Fisher recommends that customers perform a visual check of the cluster plots to confirm this assumption. The small number of SNPs from chromosome Y and the mitochondrial genome make it possible to visually inspect all of their SNP cluster graphs.

![Cluster plots of mitochondrial and Y chromosome SNPs.](image)

Panels A through F of Figure 7 show the expected pattern of homozygous genotype clusters for mitochondrial SNPs, and panels G through L of 6 show the expected pattern of homozygous genotype clusters produced by the Y chromosome SNPs of male samples. In Figure 7 Panel G-L, a cluster of No Call data is 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 essentially equivalent to background signal that are automatically set to No Call in female samples.

**Note:** It is important to exclude Y chromosome SNPs from the QC tests for X chromosome and autosomal SNPs, because the inclusion of female samples in the data set will incorrectly cause Y chromosome SNPs to fail the Call Rate and HetSO tests.

For both Y and MT SNPs, well resolved clusters are ideal. For Y SNPs, the called male samples should be slightly shifted in cluster space from the no call female samples. If the clusters are merging, then these SNPs should be excluded. Figure 8 shows examples of Y SNPs to include and exclude for different number of calls of the minor allele. Similarly for MT SNPs, the clusters should be separated in the X dimension of cluster space and should not be merging together. Additionally, the clusters should not be sitting at 0 in contrast space. Figure 9 shows examples of MT SNPs to include and exclude for cases where with and without two alleles.

**Y Probeset nMinorHom=0**

**Y Probeset nMinorHom>0**

---

**Figure 8** Y SNPs
Chapter 3 Best Practices Genotyping Analysis Workflow
Evaluate SNP cluster plots

**Figure 9** 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 13) 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 6) 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 4) should have clusters that are reasonably 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 class 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, and Chapter 8, respectively. Cluster plots in this section were produced by the *Ps_Visualization* function.

**Well-clustered vs mis-clustered SNP cluster plot patterns**

Figure 10 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 the 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).

Figure 10  Examples of a well-clustered SNP (left) and misclustered SNP (right) in contrast vs size space.
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 may be produced (Figure 11). This multi-cluster pattern may be due to genuine genetic differences in the clustered samples, especially when genotyping plant and animal genomes; or the pattern may 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 may 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 may have interfering SNPs and indels relative to the array probe sequences that decrease the genotype intensities. OTV SNP sites (Didion JP, 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 11-B. The AxiomGT1 genotyping algorithm assumes a maximum of three genotype clusters for just two alleles and thus will merge additional intensity clusters into three genotype states, resulting in unpredictable mis-calling of the true, complex genetic states.

SNP classification should classify multi-cluster SNPs as Other, CallRateBelowThreshold or OTV. In some cases, 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 4), Supplemental filters can be applied. SNPs in the OTV class can be correctly re-labeled with four genotype states including OTV (see "Adjust genotype calls for OTV SNPs" on page 35). Both Ps_Classification_Supplemental and OTV caller 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 SNPolisher software. If samples in outlier intensity clusters can be colored based on a common variable (for example a common Plate ID or a common sub-species) the potential root cause may be identified. The user may want to repeat Best Practices Step 7 genotyping, excluding the samples that form outlier intensity clusters.
Figure 11  Examples of multi-cluster SNPs. A. Diploid plant SNP with 8 intensity clusters. B. Diploid plant SNP with 5 intensity clusters; one is an OTV.
Allopolyploid genomes contain more than two paired sets of chromosomes, where each set is referred to as a sub-genome, and the sub-genomes are derived from different species. The alleles of allopolyploid SNP sites usually segregate in just one sub-genome, while remaining fixed in the homeologous sites in the other sub-genomes. Allopolyploid genomes occur in some plant and fish species and produce expected differences in SNP cluster patterns (Figure 12), relative to diploid genomes (Figure 10 left). The intensity contributions of fixed sub-genomes do not create additional clusters but they shift and compress the clusters formed by the sub-genome 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 the Ps_Classification step (see SNPolisher User Guide for more information) or in the Threshold Settings in the Axiom Analysis Suite (see Axiom™ Analysis User Guide Pub. No. 703307 for more information).

**Figure 12  Cluster plot for an allo-octoploid plant.** Each sample is colored by the AxiomGT1 genotype call (blue, gold, red) for the sub-genome with the segregating allele. Each genotype cluster is labeled by the likely allo-octoploid genotype using the following notation: the genotypes of 4 sub-genomes are separated by dashes, the genotype of the sub-genome with the segregating allele is noted first (red), followed by the genotypes of the sub-genomes whose alleles are fixed (black). It is likely that the genotypes of the fixed sub-genomes 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.
Inbred populations produce few or no heterozygous genotypes and there is often a high frequency of both of the homozygous genotypes (Figure 13). These cases will be classified as PolyHighResolution as long as the Polyploid or Diploid option is selected in the Ps_Classification step or Threshold Settings. However, a SNP with no heterozygous genotypes represented by the cluster plot shown in Figure 13-C will be classified as Other if the Human option is selected in the Ps_Classification step or Threshold Settings. AxiomGT1 analysis options should be set to include the inbred penalty when genotyping inbred populations. Additional information on using the inbred penalty can be found in the section “Genotyping inbred samples” in Chapter 4 of this guide.

Figure 13  Cluster plots for inbred populations. A. Allopolyploid plant. B and C. Diploid Plants.
**Additional genotyping methods**

### Manually change genotypes

In some cases, SNPs called incorrectly due to problematic cluster patterns can be corrected with expert manual intervention. Such cases include SNPs with cluster splits of multi-cluster SNPs such as OTV cluster patterns which escape the OTV classification. Instructions are provided in “Visualize SNPs and Change Calls through Axiom Analysis Suite cluster graphs” on page 60 for Axiom Analysis Suite.

### Adjust genotype calls for OTV SNPs

One of the SNP categories produced by the *Ps_Classification* function is OTV. The term “off-target variant” (OTV) refers to SNP sites (Didion JP, 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 interfere with genotyping of the targeted SNP. Interference may be caused by double deletions, sequence non-homology, or DNA secondary structures.

OTV SNPs display an OTV cluster with substantially lower hybridization intensities that are centered at zero in the X /Contrast dimension, and fall below the true AB cluster in the Y/Size dimension OTV clusters are often miscalled as AB (Figure 14-A).

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 re-labelled as OTV (Figure 14-B).

*OTV_Caller* intended usage is for SNPs that have been classified into the OTV class by the *Ps_Classification* function (“Step 8B: Classify SNPs using QC metrics” on page 23).

The *OTV_Caller* is fully integrated in the Axiom Analysis Suite; see the *Axiom™ Analysis Suite User Guide* (Pub. No. 703307) for more details on OTV_Caller.

Instructions for Generating SNP cluster plots for the recalled OTV genotypes and thus producing the 4th cluster colored cyan - are provided in the SNPolisher User Guide see *Ps_Visualization*.
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 auto-tetraploid genomes is different from the workflow for allopolyploid and diploid genomes.

The R package fitTetra (http://cran.r-project.org/web/packages/fitTetra/index.html) produces genotypes for autotetraploids and is recommended for Axiom arrays designed to interrogate such genomes. fitTetra was developed by Dr. RE Voorips at Wageningen University’s Plant Breeding section. The paper describing the fitTetra algorithm is available (Voorips RE, et al., 2011).

SNPolisher functions provide a workflow to (1) generate the needed Axiom data, (2) reformat Axiom data for fitTetra input, (3) use fitTetra R package for assigning genotype calls, and then (4) reformat fitTetra output for use of SNPolisher functions on the produced calls.

See Section 3.8 of the SNPolisher User Guide for detailed descriptions of the fitTetra input and output functions, as well as more information on the fitTetra package. Section 4.3 of the SNPolisher User Guide is a detailed example of how to run the functions in order to produce SNPolisher-compatible calls, confidences, and posteriors files for auto-tetraploid data.

Figure 14 Effect of OTV Calling on OTV cluster (arrow) genotypes. A. Before OTV genotyping the OTV cluster is mis-called as AB (gold). B. After OTV genotyping the OTV cluster has been identified and re-labeled as a 4th OTV genotype cluster (cyan).
Increase the stringency for making a genotype call

*Ps_CallAdjust* is a post-processing SNPolisher function for rewriting less reliable SNP calls to “No Call” by decreasing Confidence Score thresholds. Confidence Scores are discussed in “What is a SNP cluster plot for AxiomGT1 genotypes?” on page 13. A detailed description of *Ps_CallAdjust* is given in Section 3.6 of the SNPolisher User Guide, and examples of the effect of changing the threshold are described in Sections 4.1.5 and 4.2.7 in the SNPolisher User Guide.

Genotyping inbred samples

The Axiom GT1 algorithm is very 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.

### Identifying if an inbred penalty is needed

If you know your samples are inbred, the inbred penalty should be used. If you are unsure if your samples are inbred or not, proceed with genotyping without the inbred penalty and review the cluster plots for the PolyHighResolution categorized SNPs. If a very low number of heterozygous calls are present then the analysis should be redone using the inbred penalty. Figure 15 shows an example of a data set where the inbred penalty should be applied.

![Figure 15](image-url)

*Figure 15* The inbred penalty is recommended for use with samples where a low number of het calls are expected or observed.
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 5). 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 5  Example inbred file with penalty set to 4 for all samples.

<table>
<thead>
<tr>
<th>cel_files</th>
<th>inbred_het_penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT0011_001_a.CEL</td>
<td>4</td>
</tr>
<tr>
<td>GT0011_002_a.CEL</td>
<td>4</td>
</tr>
<tr>
<td>GT0011_003_a.CEL</td>
<td>4</td>
</tr>
<tr>
<td>GT0011_004_a.CEL</td>
<td>4</td>
</tr>
</tbody>
</table>

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 16). Click both “Inbred” radio buttons and provide the inbred file by clicking the “…” button. The suite will automatically use the inbred file provided during analysis. Please see Chapter 7 for more information on running best practices with Axiom Analysis Suite.
To use the inbred het penalty in APT, simply pass the --read-inbred <inbred_file> command when running both Step1 and Step2 genotyping. Please see Chapter 8 for more information on running best practices with APT.
Additional sample and plate QC

Additional sample QC

Detecting 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 will 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 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 CC, et al., 2010)

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

In Axiom, the gender calling algorithm 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.
Detecting 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 17-A); 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 will cause QC call rates to decrease. Figure 17-B shows the effect of deliberately mixing 4 samples (enclosed in box). Figure 17-C 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 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 associated with these samples. If the source of sample contamination is understood, it’s possible to proceed with the study after eliminating just those samples that obviously fall into the contamination zone. Note that contamination will produce the pattern in Figure 17-C, but it has also been observed that large image artifacts on the array surface can produce this pattern as well.
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 18 include deliberately mixed samples (red) and these points fall between the cluster locations for pure BB, AB, and AA genotypes. See the SNPolisher User Guide and usage of Ps_Visualization for instructions to color specific samples in a cluster plot.

Figure 17 DCQ vs QC Call Rate (QCCR) Plots. A. Representative data set of 10 plates with no obvious contamination problems. B. One plate including 4 samples (enclosed in box) where DNAs were deliberately mixed. C. Five plates, one plate (green) contains samples that were accidentally contaminated during pipetting.
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 described in the “Relatedness” section of Quality control and quality assurance in genotypic data for genome-wide association studies (Laurie CC, et al., 2010).

Increased computed heterozygosity

Cross-contamination of samples will increase the computed 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 will also exhibit increased computed 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 19) 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 computed early and quickly for every sample on every plate (“Step 2. Generate sample “DQC” values” on page 19).

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 obviously much lower than the rest 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 20). The 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).

Figure 19  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 (i.e., the 25th% of the DQC mean of this sample plate) is lower than 2 standard deviations below the median of medians.

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 (e.g., a HapMap sample). The calls 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 issues. A less robust but acceptable indicator of performance is to measure reproducibility by genotyping duplicate samples (the genotypes of which may not 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.
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 may 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 A, 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 CC, et al., 2010) and summarized here.

To detect batch effects on allelic frequencies, we use a homogeneity test suggested by N. J. Cox (Pluzhnikov et al., 2008). If $\tilde{p}_i$ is the sample minor allele frequency for a SNP on the i-th plate (with $n_i$ samples), $\bar{p}_i$ is the average frequency over all plates except the i-th (a total of $N_{(i)}$ samples), and $\bar{p}$ is the average over all plates (a total of $N_i$ samples), then a 1 degree of freedom chi-squared test statistic is given by $\frac{(\tilde{p}_i - \bar{p}_{(i)})^2}{\bar{p}_i(1-\bar{p}_i)}$ for each SNP. 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 may be difficult in projects with multiple sources of ethnic variation and/or relatedness among samples.
SNP QC metrics

SNP metrics used in the *Ps_Classification* Step (Step 8C)

SNP Call Rate (CR)  

\[
\text{SNP Call Rate} = \frac{\text{#Samples Called}}{N}
\]

#Samples Called = the number of samples assigned a genotype call of either AA, BB or AB at the SNP locus. That is the number of samples that do not have a “No Call” assignment.

N = the number of samples over which a genotype call is attempted for the SNP.

SNP Call Rate (CR) is the ratio of the number of samples assigned a genotype call of either AA, BB or AB for the SNP (i.e., the number of samples that do not have “No Call”) to the number of 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 20-A*). Poor cluster resolution may produce inaccurate genotypes in the samples that are called or a non-random distribution of samples with no-calls and may lead to false positive associations in a GWA study.

**Figure 20**  SNPs with different SNP Call Rates (CR).  

A. SNP with low (93.0%) CR.  

B. SNP with high (99.4%) CR.

Although SNP Call Rate 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 may result in the inclusion of some SNPs with CR just
below the threshold as well as the removal of some SNPs with CR just above the threshold. See Table 2 on page 25 for default CR thresholds used in the Ps_Classification step.

Fisher’s Linear Discriminant (FLD)

\[
Fisher’s \ Linear \ Discriminant \ (FLD) = \text{Min}(i = aa, bb) \left\{ \frac{M_{ab} - M_i}{sd} \right\}
\]

Where: \(M_{ab}\) = center of het cluster in X or Contrast dimension; \(M_{aa}, M_{bb}\) = center of hom a,b cluster in X or Contrast dimension; \(sd\) = square root of variance pooled across all three distributions.

FLD is undefined when there is only one genotype cluster.

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 generally correlated, but in some cases FLD will detect problems that are not captured by CR.

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

Figure 21-A 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 an FLD of 2.95. Conversely, the well-clustered SNP in Figure 21-B has a high CR and separated cluster centers, producing an FLD of 7.97.
Chapter 6 SNP QC metrics

SNP metrics used in the Ps_Classification Step (Step 8C)

Heterozygous Strength Offset (HetSO)

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

Heterozygous strength offset (HetSO) measures how far the heterozygous cluster center sits above or below the homozygous cluster centers in the Y dimension. In the equation above, M corresponds to the X, or contrast, dimension of the cluster plot and A corresponds to the Y, or size, dimension of the cluster plot. Low HetSO values are produced either by mis-clustering events or by the inclusion of samples that contain variations from the reference genome used to design the array probe. Most well-clustered diploid SNPs have positive HetSO values as shown in Figure 22-A (HetSO of 0.39).

Visually, SNPs with low HetSO show average signal value along the y-axis that is much lower for the heterozygous cluster center than for the homozygous clusters (Figure 22-B, Figure 22-C, Figure 22-D). Figure 22-B 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 re-analyzed with the OTV_Caller function. Figure 22-C shows a multi-cluster SNP with one very large homozygous cluster in blue (BB), divided into several sub-clusters. The heterozygous AB cluster sits very far below the BB cluster and has a negative HetSO value (~0.35). Figure 22-D shows a larger

Figure 21 Examples of SNPs with low FLD (A) and high FLD (B).

 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.

 SNP with high FLD (7.97). The three cluster centers are far apart from each other in the X dimension.
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. The low HetSO values of SNP clusters in Figure 22-C and Figure 22-D help flag these cases as problematic SNPs.

Figure 22  Examples of SNPs with different HetSO values.
Homozygote Ratio Offset (HomRO)

\[
\text{HomRO} = \begin{cases} 
\text{Min}(\text{Maa}, \text{abs} (\text{Mbb})) & \text{If both hom clusters are on correct side of zero} \\
-\text{Mbb} & \text{If both hom clusters are to the right of zero} \\
\text{Maa} & \text{If both hom clusters are to the left of zero}
\end{cases}
\]

Where Maa is the center of the AA cluster on the X, or contrast, axis and Mbb is the center of BB cluster on the X, or contrast, axis.

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

The heterozygous cluster center should be located approximately at 0 on the X-axis. If the clusters are shifted from their expected positions, then the heterozygous clusters will be 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 23-A 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 23-B 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.
Chapter 6 SNP QC metrics

Additional SNP metrics that may be used for SNP filtering

This section describes additional SNP metrics (Hardy-Weinberg p-value, Mendelian trio error, and Genotyping call Reproducibility) that may also be appropriate to examine as part of the SNP filtering process. Hardy-Weinberg p-values (pHW) are computed by Ps-metrics in APT and Axiom Analysis Suite. 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 may 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 (pHW) is a measure of the significance of the difference between the observed ratio of heterozygote calls in a population and the ratio expected if the population is in Hardy-Weinberg equilibrium (HWE). The test should be performed on unrelated individuals with relatively homogenous ancestry. Although genotyping artifacts may produce low pHW values, using this as a SNP QC metric can be tricky because a low p-value may be caused by true genotypic frequency deviation. Examination of cluster plots indicates that most of the extreme deviations (p-value < 10^{-10}) are due to poorly performing SNPs.

In Axiom Analysis Suite, the SNP Summary Table provides the pHW.
Mendelian Trio Error

Mendelian errors can be detected in parent-offspring trios. Mendelian trio error rate is calculated as the number of errors 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 (e.g., the offspring of AB and BB parents may 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 is 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 computed 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 compute 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 may vary considerably, and these differences in sample quality may influence the genotyping call error rates among samples. Therefore, the replicated sample sets should be comprised 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 in Axiom Analysis Suite, click the Concordance button in the Sample Table tab. Select Compare all combinations, select your desired SNPs and click OK.
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 24 shows the full workflow for 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, please 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 handling batches up to 50 plates of samples. Please see Chapter 3 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 25).

### Setup step 2, 3, 5, 6 and 8A, B: Set Sample Metrics, Plate Metrics, and SNP Metrics

All of these steps are entered at the same time in Axiom Analysis Suite. The Threshold Settings window provides single a location to enter and edit all of the metrics associated with the Best Practices Workflow (Figure 26). Three default configurations are available: human, diploid and polyploid. To run the analysis, select the appropriate default configuration. Please see Chapter 3 and Chapter 6 of this guide for detailed information on the thresholds.

![Figure 25 Import CEL File button](image-url)
Chapter 7  Executing Best Practices steps with Axiom™ Analysis Suite

Execute steps 1-8 with Axiom™ Analysis Suite

![Threshold Settings window](image)

**Figure 26**  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 27). Typically two default configurations are available for an array: <96 samples or ≥96 samples, though some arrays may 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. Please see section ("Genotyping inbred samples" on page 37) for more information on inbred samples. Hints files are not recommended to be used for most users.

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 be sure to load it for both sample QC and genotyping.

![Figure 27 Analysis Settings window](image-url)
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 28). Be 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 results should be reviewed.

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 29). The Sample Table tab provides sample level information (Figure 30). We recommend reviewing all of the QC plots created. The SNP Summary Table tab provides SNP level information (Figure 31). Please see the Axiom™ Analysis Suite User Guide (Pub. No. 703307) for more information on these tabs and the default plots created.
Chapter 7 Executing Best Practices steps with Axiom™ Analysis Suite

Execute steps 1-8 with Axiom™ Analysis Suite

Figure 29  Summary tab
### Figure 30  Sample Table tab
Chapter 7 Executing Best Practices steps with Axiom™ Analysis Suite

Execute steps 1-8 with Axiom™ Analysis Suite

Table: SNP Summary Table

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 13) and produces plots that are similar to the output from the R function Ps_Visualization. However, the SNP Cluster Graph function in Axiom Analysis Suite has more functionality than the Ps_Visualization function. Since Axiom Analysis Suite executes all steps of the Best Practices Workflow, the cluster graphs will be 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 the user to adjust the shape and color of the samples. Figure 32 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 while male samples are plotted as circles. Users can select and change the calls of samples through the plotted cluster graph. See "Evaluate SNP cluster plots" on page 29 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 32 SNP cluster graph
Display a particular SNP

To display a particular SNP, click the corresponding row in the SNP Summary Table. The cluster graph will update 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 will be highlighted in the Sample Table (Figure 33).

---

**Figure 33**  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 on the plot and circling the samples with the mouse before releasing the mouse button (Figure 34). 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 34  Using 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 it. The Change Call menu appears. Select the new call (Figure 35).

Figure 35  Manually changing a call
Lasso function

The lasso function can be used in a number of different cases including cluster splits. In Figure 36 the top half shows a cluster split, and the bottom image shows the graph after setting the samples correctly to AB.

Figure 36 Using the lasso function to change calls in a cluster split.
Saving 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 hand corner of the cluster plot tab (Figure 37).

![Cluster plot Save Image button](image)

**Figure 37** Cluster plot Save Image button

- To save multiple cluster plots to a single PDF file do the following:
  1. Click the Export drop-down in the SNP Summary Table and choose Export cluster plots to PDF (Figure 38).

![Exporting multiple cluster plots to PDF file format](image)

**Figure 38** Exporting multiple cluster plots to PDF file format

  2. This opens the Report Settings window. Select if you want 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 probeset list manually. Axiom Analysis Suite creates a recommended SNP list based on the “Recommended” settings in the Threshold Analysis tab (Figure 39) as well as selecting the best probe set for a marker as indicated in the BestProbeset column of the SNP Summary tab. By default this setting matches Table 4 on page 26.

To import this SNP list into Axiom Analysis Suite:

1. Click the Manage SNP List drop-down.
2. Click the Import SNP List to Batch option.

The recommended.ps will be under the “SNPolisher” folder of the batch folder.

Figure 39 Recommended SNP lists: A. Recommended settings in Threshold Analysis tab. B. Import SNP List to Batch
If you wish to create your own recommended probe set list, do the following:

- Filter the SNP Summary Table on the **Conversion Type** by right-clicking the column header and applying a filter (Figure 40), additionally, SNPs should be filtered on the **Best Probeset** column.
- Click the **Manage SNP List** drop-down and click **Create SNP List from Table**.
- Name the SNP list and click **OK**.
- Under the **Manage SNP List** drop-down, click **Export Saved SNP List to Text File**.
- Select the previous saved SNP list from the drop-down and click **OK**.

---

**Figure 40** Filter Conversion Type column

---

**Running OTV Caller or Classification Supplemental**

Axiom Analysis Suite includes the OTV Caller and Classification Supplemental function as part of the software. To access these features, click the **Reanalyze** drop-down in the SNP Summary Table (Figure 41). Please see the SNPolisher User Guide for more information on these functions.

---

**Figure 41** Reanalyze drop-down menu

---

**Note:** Please be sure to create a new recommended probeset list after running any addition analysis.
Exporting 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 67).
2. In the SNP summary Table, click the Export drop-down and select Export Genotyping Data.
   The Export Genotype Data window appears (Figure 42).
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 drop-down next to SNP List Filter and select the created recommended SNP list.
6. Select the Output Location and Output Name.
7. Select SNP Identifier.
   Note: VCF file formats intended 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 begin export.

Figure 42 Export Genotyping Data window
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 4) using 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.

To execute the commands/scripts in Windows:
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
```
Chapter 8 Executing Best Practices steps with command line software

Best Practices Step 2: Generate the sample “DQC” values using 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), the user must write a script to remove .CELs from the <OUTDIR>/cel_list1.txt with DQC values that are < 0.82. We will refer to filtered .CEL list from this step as cel_list2.txt.

Best Practices Step 4: Generate sample QC call rates using 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:** Choose `<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 less 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), the user must write a script to remove .CELs from the <OUTDIR>/step1/cel_list2.txt whose call rate values are less than 97%. We will refer to this .CEL list as cel_list3.txt. Note that the AxiomGT1.report.txt file will have a number of header lines beginning with #. The file can be read directly into a table (data.frame) using the R “read.table” function, which ignores lines beginning with #.

Best Practices Step 6: QC the plates

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

Note that the user must write a script or use EXCEL to compute the plate QC metrics.

- Group the .CEL files by plate, then for each plate:
  - Compute plate pass rate
    - Plate Pass Rate = \( \frac{\text{# of samples passing DQC and 97% QC call rate}}{\text{Total # of samples on the plate}} \times 100 \)
  - Compute 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
  - Compute 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 may 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 re-processing.
Best Practices Step 7: Genotype passing samples and plates using AxiomGT1.Step2

Step 7 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 7
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 \
--write-models \
--batch-folder <OUTDIR>/suitefiles \
--cel-files <OUTDIR>/cel_list3.txt

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 a number of files to <OUTDIR>/step2.

The default files are:

- AxiomGT1.calls.txt which contains the genotype calls (coded into 0, 1, 2 and -1) for each probe set and sample.
- AxiomGT1.confidences.txt, which contains the confidence score (described in "What is a SNP cluster plot for AxiomGT1 genotypes?” on page 13) for each genotype call in the AxiomGT1.calls.txt file.
- AxiomGT1.report.txt, which contains information about each sample.

Note: The output is per probe set, not per SNP. A probe set is a set of probe sequences interrogating a SNP site. Although most SNP sites are interrogated by only one probe set (and therefore there is usually a one-to-one correspondence between probe set and SNP site), some SNP sites are interrogated by more than one probe set.
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:** Choose <axiom_array>_LessThan96_Step2.r<#>.apt-genotype-axiom.AxiomGT1.apt2.xml if available to perform genotyping with SNP-specific models if batch size is less than 96 samples.

### Best Practices Step 8A: Run *Ps-Metrics*

*Ps-metrics* uses two output files from Best Practices Step 7 (AxiomGT1.Step2 above) as inputs: AxiomGT1.posteriors.txt and AxiomGT1.calls.txt. Additionally, *ps-metrics* will calculate the metrics on only a subset of probe sets if a list of desired probe sets is supplied. See the SNPolisher User Guide for a more detailed description of *Ps-metrics*.

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
```

The output from *Ps-Metrics* (the default name is “metrics.txt”) is a text file containing the SNP QC metrics. Each row is a SNP and each column is a QC metric. The output should look similar to Figure 43. This output file will be one of the input files for other SNPolisher functions, so the user must know the file’s name and location on the computer.

![Figure 43](image-url) An example of the output file from *Ps-Metrics*.
Best Practices Step 8B: Run *Ps_Classification*

Once the *Ps-Metrics* function has been run and the SNP QC metrics generated and output to *metrics.txt*, the SNPs can be classified using *Ps-Classification*. *Ps-Classification* has three required arguments and 15 optional arguments. The three required arguments are:

1. the name and location of the output metrics file from *Ps-Metrics*,
2. the location of the preferred output directory, and
3. the species (or genome) type: human, non-human diploid, or polyploid.

A 4th argument: `ps2snpFile` is needed for arrays that includes 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 on page 9). If this file has not been provided, users should contact their 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 *metrics.txt*,
2. stores classification results in the folder `<OUTDIR>/SNPolisher`,
3. the genotype data is human
4. `ps2snp.txt` file = `<ANALYSIS_FILES_DIR>/Axiom_BioBank1.r2.ps2snp_map.ps`.  

   `<ANALYSIS_FILES_DIR>` means the full path to the Analysis Library file directory.

   ```
   ps-classification \
   --species-type human \
   --metrics-file --metrics-file <OUTDIR>/SNPolisher/metrics.txt \
   --output-dir <OUTDIR>/SNPolisher \
   --ps2snpfile <ANALYSIS_FILES_DIR>/Axiom_BioBank1.r2.ps2snp_map.ps
   ```

Eight of the optional 15 arguments are 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 2 on page 25.). SNPs classified as *PolyHighResolution* must have SNP QC values that pass all of the thresholds.

There are two logical indicators: `hom-ro` indicates if HomRO thresholds should be used (default is TRUE), and `hom-het` indicates if the HomHet metric should be used (default is TRUE). Polyploid genotypes do not use either of the HomRO thresholds so `hom-ro` flag should be set to FALSE.

When the HomHet metric is set to TRUE (default), *Ps-Classification* will classify two-cluster SNPs with one homozygote and one heterozygote cluster as NoMinorHom. If set to FALSE, SNPs will be classified as PolyHighResolution. Missing the minor homozygote cluster is unreasonable for highly inbred species (e.g., wheat). This metric should be turned on when classifying probe sets in highly inbred species.

The two optional arguments that deal with conversion are `converted` and `priority-order`. `converted` is a logical indicator for outputting a list of converted/recommended SNPs to the file `converted.ps` (default is FALSE).

`priority-order` is used when performing probe set selection: the best probe set is selected according to the priority order of probe set conversion types. These are based on the default category order: PolyHighResolution, NoMinorHom, OTV,
MonoHighResolution, and CallRateBelowThreshold. The priority-order argument allows the user to change the order of categories when determining which probe sets are selected as the best probe set for a SNP. All five of the listed categories must appear in priority-order, where the user specifies the order.

Ps-Classification accepts a list of probe sets, and will categorize the SNPs in this file only. The first row of this file should always be “probeset_id”.

See the SNPolisher User Guide for more details of the arguments for Ps-Classification.

Visualize SNP cluster plots with SNPolisher *Ps_Visualization* function

Once the *Ps-Metrics*, *Ps-Classification*, and *OTV-Caller* functions have been run, the *Ps_Visualization* function can be used to produce SNP cluster plots. Plots include the posterior information (default) and prior information (optional). Reference genotypes can also be included. The cluster plots can help quality check SNPs and diagnose underlying genotyping problems. All plots are output as PDFs. The user can adjust the colors used in the plots and highlight select samples. *Ps_Visualization* must be run in R and is not available in APT. For more details on installing R and the various options with *Ps_Visualization*, see the SNPolisher User Guide.

*Ps_Visualization* takes four required arguments and many optional arguments. The four required arguments are the pidFile with the probeset_ids for plotting, the name and location of the output PDF file, the name and location of the summary file, the name and location of the calls file, the name and location of the confidences file, and the name and location of the posteriors file.

The list of probeset_ids (pidFile) can either be the list of probeset_ids output by *Ps_Classification* for one category files or a list of probeset_ids selected by the user. The first line of the pidFile should always be “probeset_id”.

The user should also give *Ps_Visualization* the name of a temporary directory which will be used for putting intermediate files (*temp.dir*). If no directory is given, the default used is “Temp/”.

*Ps_Visualization* has many optional arguments. Please refer to the SNPolisher User Guide for a complete description of all options. A few examples of some useful options are shown here. Reference genotypes may be visualized as shown in Figure 44 where samples are plotted and colored by both called genotypes and reference genotypes. Samples may be highlighted in the cluster plots as shown in Figure 45 and Figure 46.
Chapter 8 Executing Best Practices steps with command line software
Visualize SNP cluster plots with SNPolisher Ps_Visualization function

Figure 44 One SNP plotted with and without reference genotype

Chapter 8 Executing Best Practices steps with command line software
Visualize SNP cluster plots with SNPolisher Ps_Visualization function

**Figure 45** One SNP plotted with color and sample highlighting options

- **A**: One SNP plotted with default color options and highlighted samples.
- **B**: The same SNP plotted with different colors and the same highlighted samples.

**Figure 46** One SNP with samples highlighted in two colors

- **A**: One SNP plotted with default color options and samples highlighted in green and turquoise.
- **B**: The same SNP with all non-highlighted samples plotted in light gray.
Documentation and support

Related software documentation

<table>
<thead>
<tr>
<th>Document</th>
<th>Publication number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axiom™ Analysis Suite User Guide</td>
<td>703307</td>
<td>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.</td>
</tr>
<tr>
<td>SNPolisher User Guide</td>
<td>N/A</td>
<td>SNPolisher is an R package that provides advanced SNP quality control (QC), genotyping, and visualization tools.</td>
</tr>
</tbody>
</table>

Customer and technical support

Visit thermofisher.com/support for the latest in services and support, including:

- Worldwide contact telephone numbers
- Product support, including:
  - Product FAQs
  - Software, patches, and updates
- Order and web support
- Product documentation, including:
  - 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.
References


