affymetrix

Analysis Guide

Axiom® Genotyping Solution Data Analysis Guide

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

P/N 702961 Rev. 2

Contents

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 Affymetrix® Genotyping Console™ Software (GTC) or Affymetrix® Power Tools (APT) and SNPolisher R package to perform quality control analysis (QC), for samples, plates, and 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 Affymetrix 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 Affymetrix Field Application Support or send email to Support@affymetrix.com.

Analysis Software

Three analysis software systems are used for Axiom analysis and described in this document: (1) Affymetrix® Genotyping Console™ (GTC) version 4.2 and above, (2) Affymetrix® Power Tools (APT) version 1.16.1 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 14.

GTC is a software package that provides a graphical user interface for most of the algorithms contained in APT and is designed to streamline whole-genome genotyping analysis and quality control (Genotyping Console Software information). GTC includes quality control and visualization tools to easily identify and segregate sample outliers, including a cluster graph visualization tool that enables a detailed look at the performance of SNPs of interest. GTC 4.2 and above provides the capability to manually edit genotypes.

APT is a set of cross-platform command line programs that implement algorithms for analyzing and working with Affymetrix arrays (Affymetrix 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 GTC software user manual[1] and the APT help[2].

SNPolisher R functions provide SNP quality control and classification, visualization tools, and advanced genotyping methods. Many of the SNPolisher functions are available in APT or GTC. 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 Affymetrix website (www.affymetrix.com). Select **Register** at the top of the website to register your email address with Affymetrix. From the **Partners and Programs** menu, select **Developers' Network**. Click **DevNet Tools** on the left side of the menu. SNPolisher is available under the **Analysis Tools** tab. Download the zipped SNPolisher folder

[1] *Genotyping Console 4.2 User Manual: http://media.affymetrix.com/support/downloads/manuals/gtc_4_2_user_manual.pdf*

[2] *Affymetrix Power Tools User Manual: http://media.affymetrix.com/support/developer/powertools/changelog/apt-probeset-genotype.html*

(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, and readme files, a PDF with colors for use in R, and the example R code and two 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.

Both APT and GTC software tools require the files (collectively referred to as "analysis library files") listed in Table 1.1 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.affymetrix.com) or through direct download via GTC.

Table 1.1 lists the names of all analysis files used to process Axiom genotyping arrays in GTC or APT. An *annotation* file is an additional file not required for genotyping and is not listed below, but used in GTC to display SNP annotations in SNP results tables, the cluster graph visualizations, and for some export functionality. Annotation files are available for download through GTC or on the array product page or the Secure File Exchange in the same locations as the analysis library files.

Table 1.1 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-probeset-genotype.AxiomGT1.xml= Axiom_BioBank1_96orMore_Step2.r2.apt-probeset-genotype.AxiomGT1.xml.

Analysis Library Files	GTC	APT
<axiom_array>.r<#>.ps2snp_map.ps	N/A	Required
<axiom_array>_96orMore_Step1.r<#>.apt-probeset-genotype.AxiomGT1.xml	N/A	Required
<axiom_array>_96orMore_Step2.r<#>.apt-probeset-genotype.AxiomGT1.xml	N/A	Required
<axiom_array>.r<#>.cdf	Required	Required
<axiom_array>.r<#>.qca	Required	Required
<axiom_array>.r<#>.qcc	Required	Required
<axiom_array>.r<#>.step1.ps	Required	Required
<axiom_array>.r<#>.genetic_prior.txt	Required	Required
<axiom_array>.r<#>.AxiomGT1.sketch	▪ Required for human genomes ▪ Optional for non-human genomes	▪ Required for human genomes ▪ Optional for non-human genomes
<axiom_array>.r<#>.chrXprobes	▪ Required for mammalian genomes ▪ N/A for non-mammalian genomes	▪ Required for mammalian genomes ▪ N/A for non-mammalian genomes
<axiom_array>.r<#>.chrYprobes	▪ Required for mammalian genomes ▪ N/A for non-mammalian genomes	▪ Required for mammalian genomes ▪ N/A for non-mammalian genomes
<axiom_array>.r<#>.specialSNPs	▪ Required for human genomes ▪ Required for non-human genomes if gender calling is executed	▪ Required for human genomes ▪ Required for non-human genomes if gender calling is executed
<axiom_array>_LessThan96_Step1.r<#>.apt-probeset-genotype.AxiomGT1.xml	N/A	Required for small sample size
<axiom_array>_LessThan96_Step2.r<#>.apt-probeset-genotype.AxiomGT1.xml	N/A	Required for small sample size
<axiom_array>.r<#>.AxiomGT1.MODELS	Required for small sample size	Required for small sample size

Table 1.1 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-probeset-genotype.AxiomGT1.xml= Axiom_BioBank1_96orMore_Step2.r2.apt-probeset-genotype.AxiomGT1.xml. (Continued)

Analysis Library Files	GTC	APT
<axiom_array>.r<#>.apt-geno-QC.AxiomQC1.xml	N/A	Optional
<axiom_array>r<#>.step2.ps	Optional	Optional
<axiom_array>.apt.probeset-genotype.AxiomSS1.xml	N/A	▪ 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
<axiom_array>.array_set	Required	N/A
<axiom_array>.AxiomGT1.gc_analysis_parameters	Required	N/A
<axiom_array>.gc_analysis_configuration	Required	N/A
<axiom_array>.analysis	Required	N/A
<axiom_array>.qc_thresholds	Required	N/A
<axiom_array>.gt_thresholds	Required	N/A
<axiom_array>.geno_intensity_report	Required	N/A

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[1,2,3]. 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[4]. 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 and 8.
- 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, *Instructions for Executing Best Practices Steps with GTC and APT Software* provides instructions for executing the seven Best Practices Steps with GTC combined with usage of APT version 1.16.1 or higher for step 8. Instructions for visualizing SNP cluster plots and changing genotype calls with GTC SNP Cluster Graph is also provided in this chapter.
- Chapter 8, *Instructions for Executing Best Practices Steps with Command Line Software* provides instructions for executing the Best Practices steps 1-7 with APT combined with usage of SNPolisher functions for executing Step 8. Instructions for visualizing SNP cluster plots with SNPolisher function *Ps_Visualization* is provided in this chapter.

---

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

[2] *de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet. 2008 Oct 15;17(R2):R122-8.*

[3] *Baker M. Genomics: The search for association. Nature. 2010 Oct 28;467(7319):1135-8.*

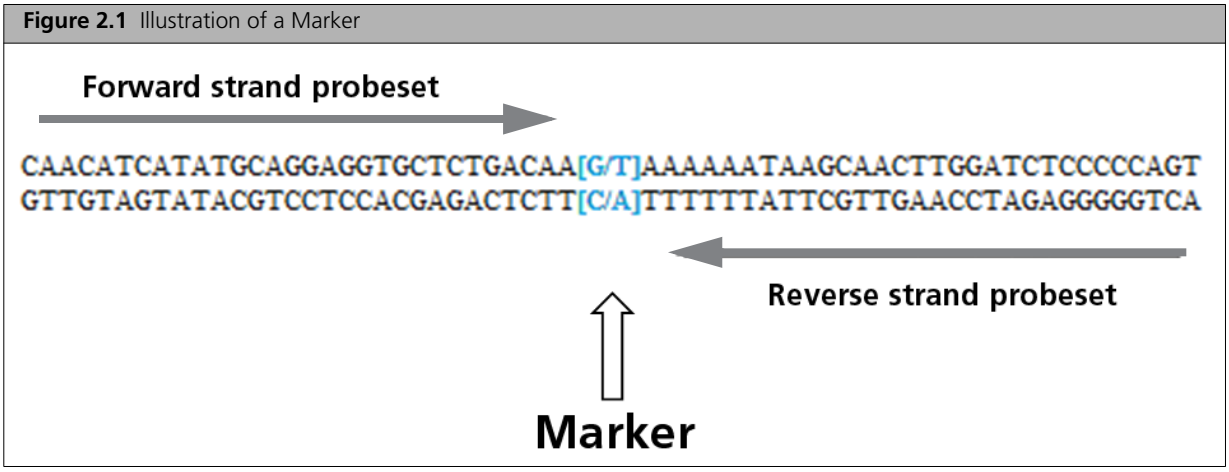
[4] *Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS; GENEVA Investigators. Quality control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

Background

Axiom® Array Terminology

Marker

Figure 2.1 Illustration of a 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 Affymetrix® 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”.

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) the SNPolisher package, via the *Ps_Visualization* function (example shown in Figure 2.2) and (2) GTC, via the *SNP Cluster Graph* function (example shown in Figure 2.3). Instructions for the *Cluster Graph* and *Ps_Visualization* function usages are provided in Chapter 7 and Chapter 2; respectively.

AxiomGT1, is a tuned version of the BRLMM-P[1] 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

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

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 [Log2(A_signal) + Log2(B_signal)]/2 and dimension X is calculated as Log2(A_signal/B_signal). X carries the main information for distinguishing genotype clusters. The X dimension is called *Contrast* in cluster plots produced by SNPolisher and *log ratio* in cluster plots produced by GTC. The Y dimension is called *Size* in cluster plots produced SNPolisher and *Strength* in cluster plots produced GTC.

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 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. The GTC SNP Cluster Graph defaults are set to have BB calls as red triangles, AB calls as blue squares, AA calls as green circles. Note, in GTC 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. SNPolisher *Ps_Visualization* defaults are set to have No Calls as gray squares and GTC *SNP Cluster Graph* defaults are set have No Calls as gray spades.

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 both the SNPolisher and GTC 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.

Figure 2.2 SNP Cluster Plot produced by the SNPolisher package, via the *Ps_Visualization* function.

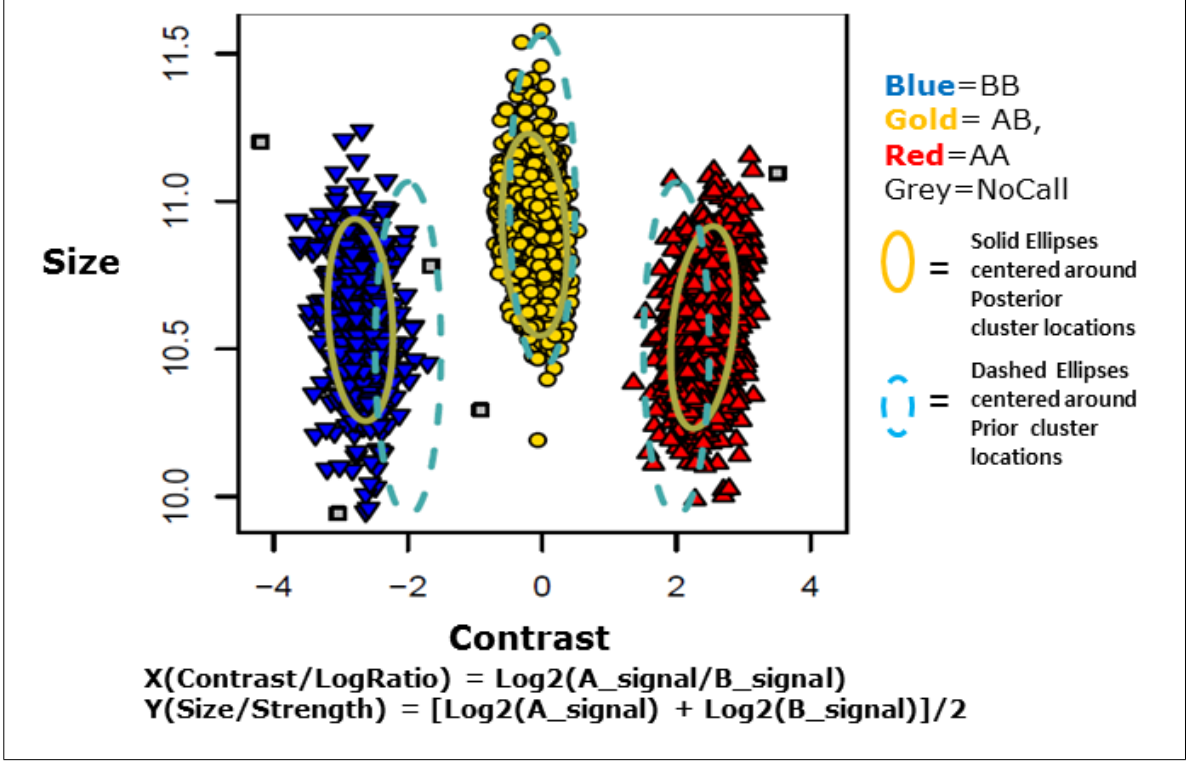
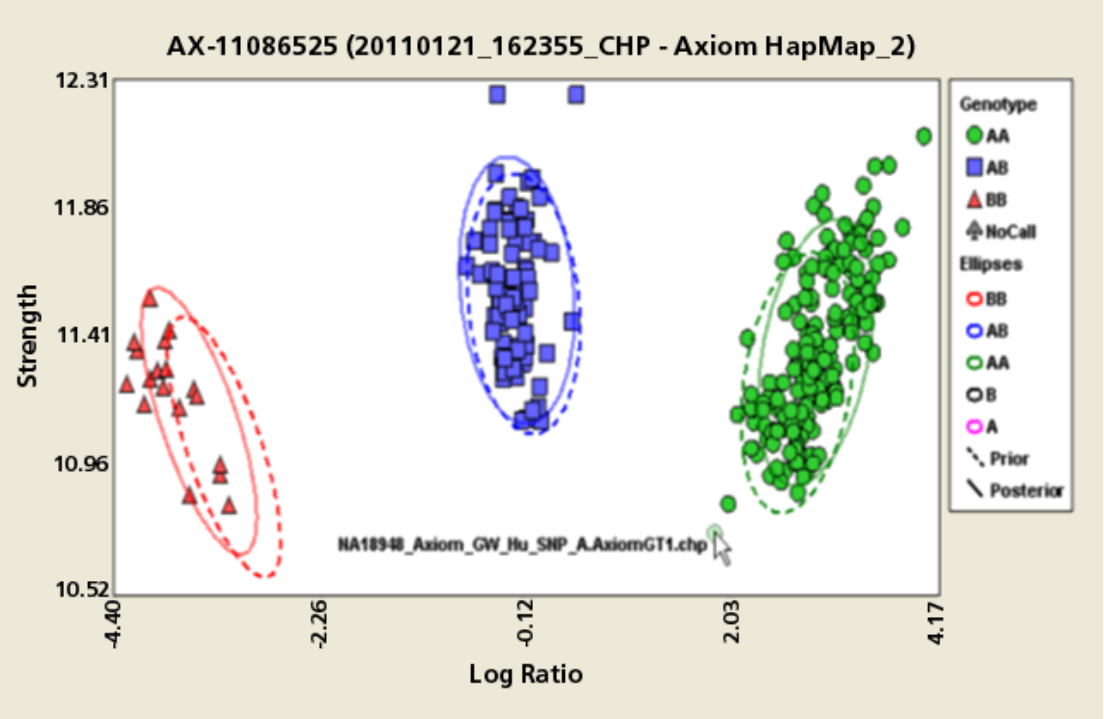


Figure 2.3 SNP Cluster Plot produced by the GTC via *SNP Cluster Graph* function.



Best Practices Genotyping Analysis Workflow

Design the Study to Avoid Experimental Artifacts

Good experimental design practices[1,2,3,4] 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.

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

[2] *Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet. 2003 Feb 15;361(9357):598-604.*

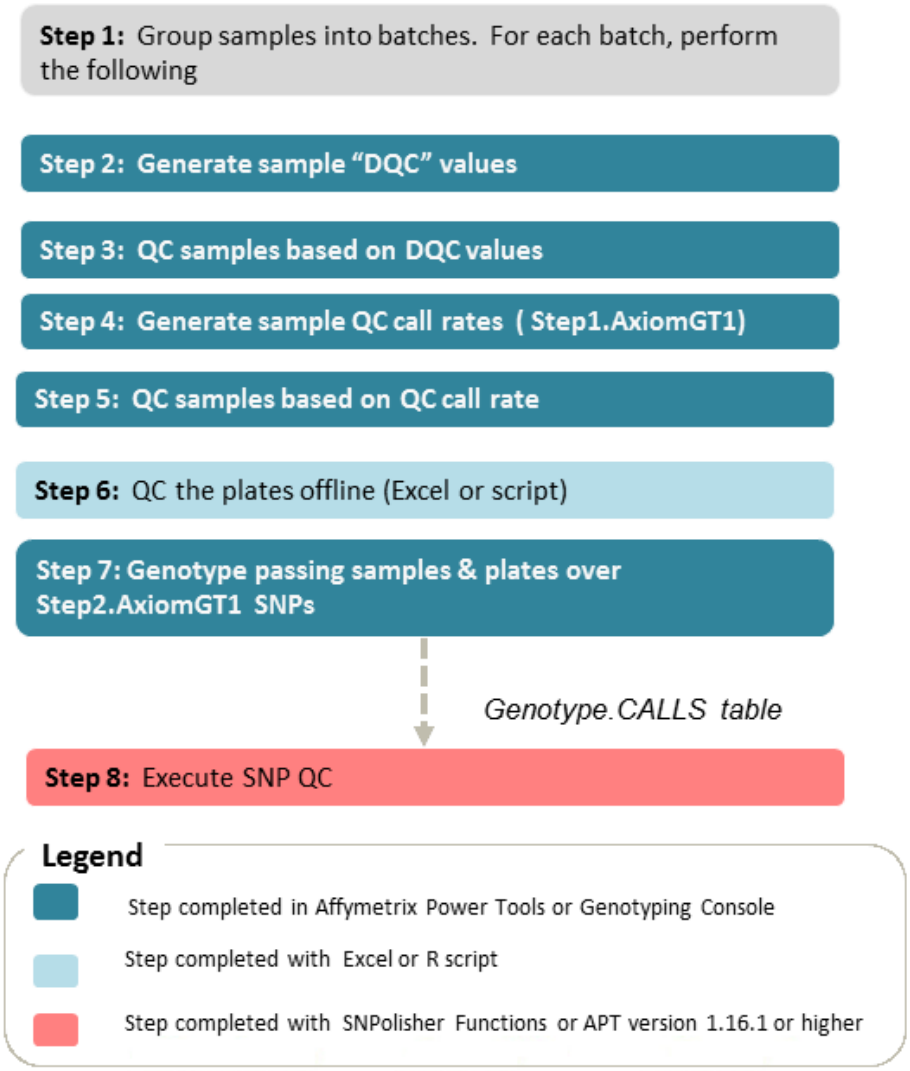
[3] *Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA. Population structure, differential bias and genomic control in a large-scale, case-control association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

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

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 3.1).

Figure 3.1 Steps for Best Practices Genotyping Analysis Workflow



The actual commands used to execute the steps differ between the software tools, APT, GTC and SNPolisher. Instructions for using GTC to execute the Best Practices Workflow Steps: 1-7, followed by APT for Step 8, are provided in Chapter 7. Instructions for using APT to execute Best Practices Workflow Steps: 1-7, followed by SNPolisher functions for Step 8, 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 50 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 both GTC and APT 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 16, and *Additional Plate QC* on page 32) 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.
- Samples with auto-polyploid and allo-polyploid 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 co-cluster only samples from the separated divergent population, identify a sub-population set of working SNPs.

Our guideline for maximum batch size is 50 Axiom® 96-Array Plates per batch. This is based on internal Affymetrix 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 SNPolisher, 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 3.1) using the apt-probe set-genotype command (*Best Practices Step 7: Genotype Passing Samples and Plates Using AxiomGT1.Step2* on page 64) 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.

Step 2. Generate Sample "DQC" Values

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

DQC is based on intensities of probe sequences for non-polymorphic genome locations (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:

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

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.

Step 3: QC the Samples, Based on DQC

Samples with a DQC value less than the default DQC threshold should be excluded from *Step 4: Generate Sample QC Call Rate Using Step1.AxiomGT1*. 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.

Step 4: Generate Sample QC Call Rate Using Step1.AxiomGT1

Not all problematic samples are detectable by the DQC metric prior to the first round of genotyping (see *Detecting Mixed (Contaminated) DNA samples* on page 30). 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), Affymetrix requests the user to provide at least a plate of Axiom data to identify probe sets that meet this criteria. Affymetrix will then provide the Axiom Analysis Library package (Table 1.1) for the array. Users should contact their local Affymetrix Field Application Support or send email to Support@affymetrix.com when such data is available.

This Best Practices Step 4 is referred to as *Step1.AxiomGT1* genotyping in the instructions provided for genotyping with GTC (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.

Step 5: QC the Samples Based on QC Call Rate

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 30 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 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 23), 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 = $\frac{\textit{Samples passing DQC and 97\% QC call rate}}{\textit{Total samples on the plate}}$ X 100
- Average QC call rate of passing samples on the plate= MEAN (QC call rates of samples passing DQC and 97% QC call rate thresholds)

Guideline for High-quality Plates

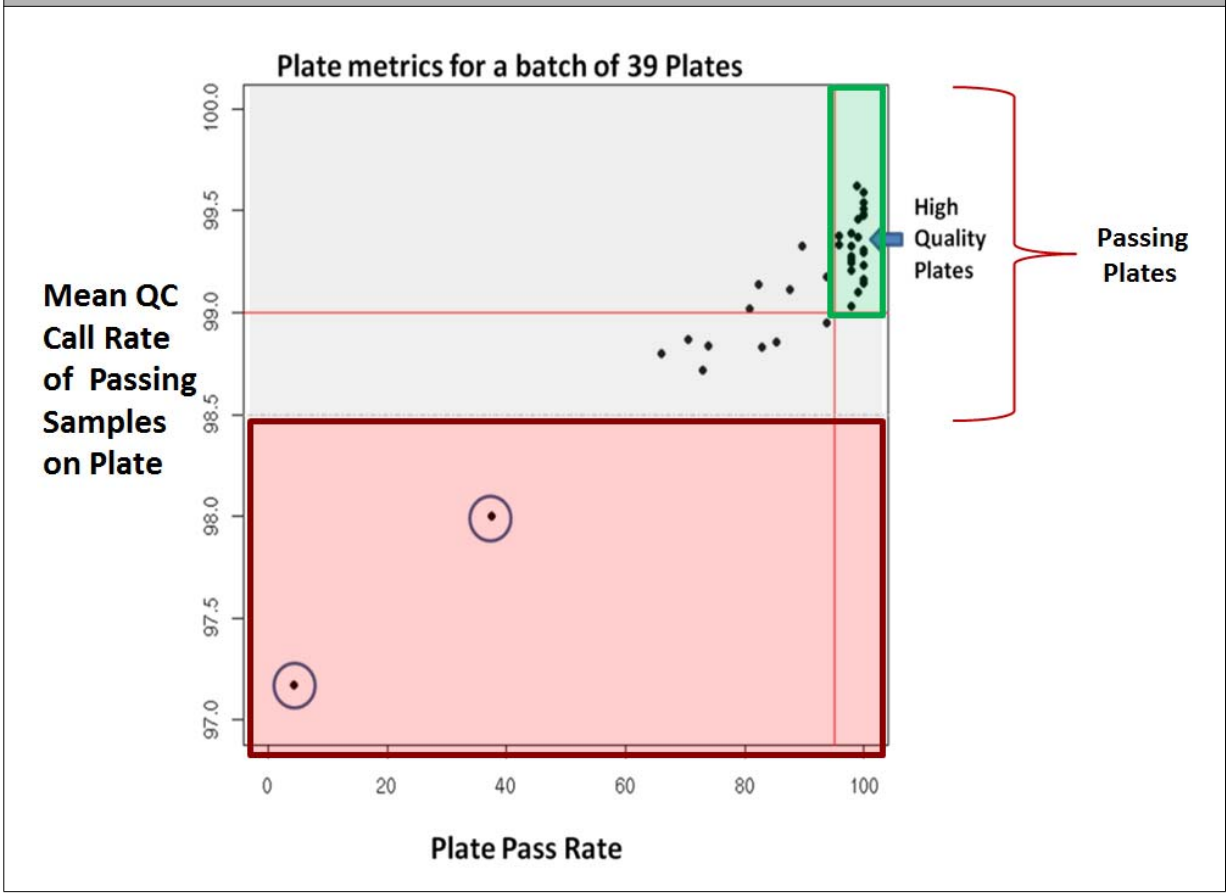
- Plate pass rate ≥ 95% for samples derived from tissue, blood or cell line, and ≥ 93% if sample source is saliva
- Average QC call rate of passing samples ≥ 99%

Guideline for Passing Plates

- Average QC call rate of passing samples ≥ 98.5%

The minimum guideline for passing plates is an average QC call rate of passing samples that is greater than or equal to 98.5% (gray and green zones in Figure 3.2). Ideally all plates in the batch will pass the guidelines for high-quality plates (green zone Figure 3.2). 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 3.2) 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 Affymetrix Field Application Support person.

Figure 3.2 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 32 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 GTC (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 10, 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 1.1 on page 7) 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 such as allo-polyploid genomes (discussed below), arrays with a large fraction of SNPs that are monomorphic in the population, and arrays whose intended usage is genomic selection. 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) often run 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 23) and whose genotypes are recommended for statistical tests in the downstream study. 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.

Steps 8 uses the *Ps_Metrics* and *Ps_Classification* functions. These functions are available in the SNPolisher R package and APT software version 1.16.1 or greater. Instructions for SNPolisher R package usage are provided in *Execute Best Practice Step 8 with SNPolisher Functions* on page 65, and for APT usage in *Execute Step 8 with APT Version 1.16.1 or Higher* on page 49. The *Ps_Metrics* and *Ps_Classification* functions are not currently available with the GTC software.

Step 8A: Create SNP QC Metrics

The *Ps_Metrics* function is used on the output files from the Best Practices Step 7 genotyping run (also referred to 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 Metrics Used in the *Ps_Classification* Step (Step 8C).

Additional SNP QC tests used in the field are discussed in *Additional SNP Metrics that may be Used for SNP Filtering* on page 39.

Step 8B: Classify SNPs Using QC Metrics

The *Ps_Classification* function is used to sort the "best" probe set per SNP into seven classes based on five SNP QC metrics generated by the *Ps_Metrics* function. The classes are described in Figure 3.3. The seven classifications are based on default QC thresholds shown in Table 3.1 for different genome types. Note, the user can change the thresholds if desired.

The best probe set is determined by the classification priority order: PolyHighRes, NoMinorHom, OTV, MonoHighRes, and CallRateBelowThreshold. For a SNP with two probe sets, where one probe set is NoMinorHom and one probe set is MonoHighRes, 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 *priority.order* argument for *Ps_Classification*. The file <axiom_array>.r<#>.ps2snp_map.ps in the Analysis Library File package (Table 1.1 on page 7) contains the list of matched probe sets and SNPs.

Figure 3.3 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 28.

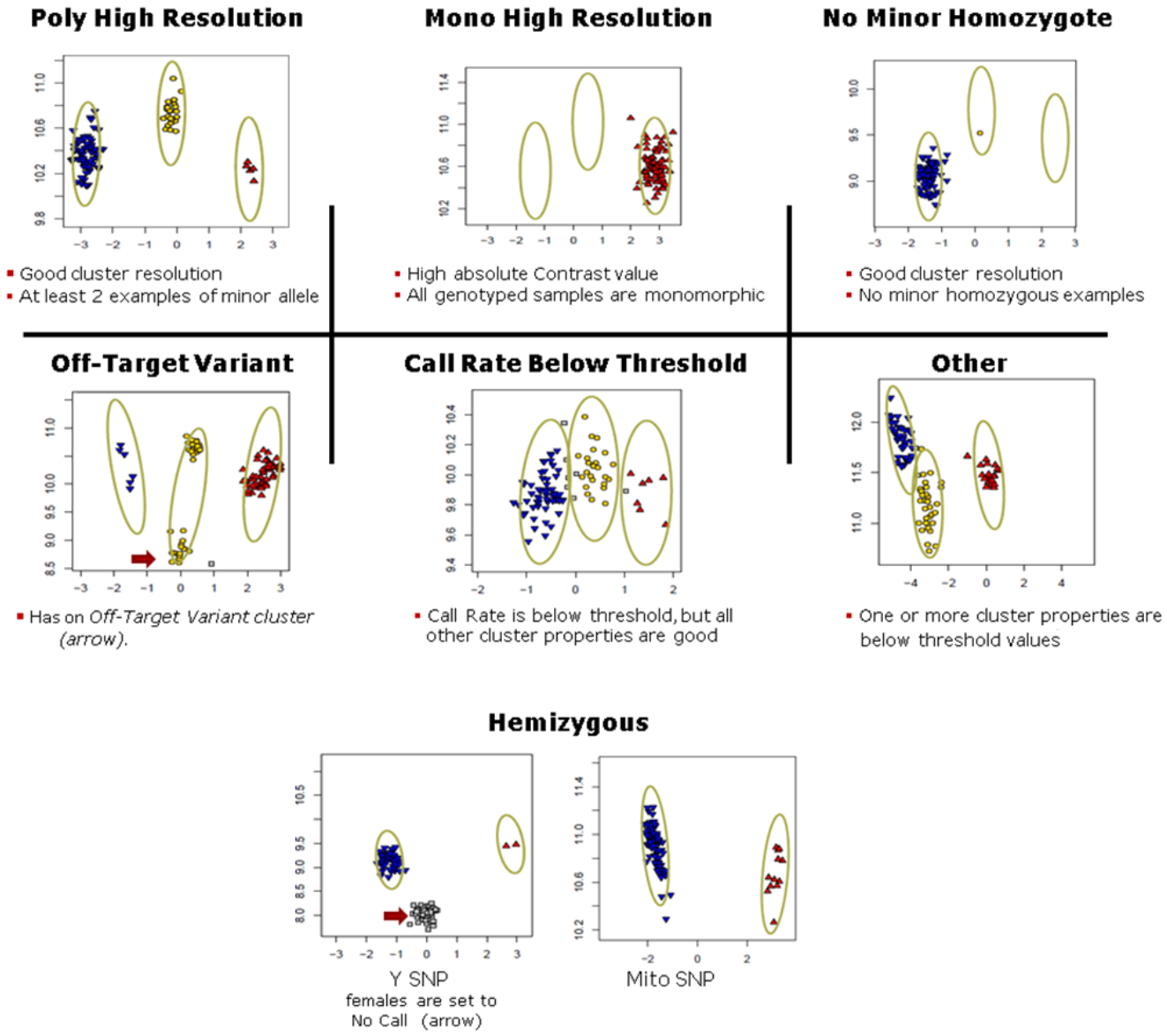


Table 3.1 Default QC Thresholds for CR, FLD, HetSO HomRO (metrics defined in Chapter 6) and nMinorAllele (number of minor alleles in the batch). Metric values must be greater than or equal to the threshold in order to be considered passing. HetSO.OTV is the HetSO threshold for OTV detection (see *Adjust Genotype Calls for OTV SNPs* on page 28). 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.

Metric	Human	Diploid	Polyploid
CR	95	97	97
FLD	3.6	3.6	3.6
HetSO	–0.1	–0.1	–0.1
HetSO.OTV	–0.3	–0.3	–0.3
HomRO1	0.6	0.6	N/A
HomRO2	0.3	0.3	N/A
HomRO3	–0.9	–0.9	N/A
nMinorAllele	2	2	2

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 3.3) 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 3.2).

Table 3.2 *Ps.performance* Column Names and Examples

Column Name	Example
probeset_id	AX-11481545
affy_snp_id	Affx-27771153
CR	99.232012934519
FLD	8.19369622294609
HomFLD	17.9734932365231
HetSO	0.450052256708187
HomRO	2.57169
nMinorAllele	5123
Nclus	3
n_AA	3112
n_AB	3383
n_BB	870
n_NC	57
hemizygous	0
HomHet	0
ConversionType	PolyHighResolution
BestProbeset	1

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 *OTV.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 an Affymetrix 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 3.3 shows which classes are recommended for the given genome type. SNPs in recommended classes are also referred to as *converted* in this document.

Table 3.3 Recommended SNP Classes Based on Genome Type and SNP Class (Figure 3.3).

	SNP Class determined by *Ps.Classification* Function				
Genome Type	**PolyHighRes**	**NoMinorHom**	**MonoHighRes**	**Hemizygous**	**OTV**
Human	Recommended	Recommended	Recommended	Recommended, visual inspection is advised	Recommended after genotyping with *OTV_Caller* function
Diploid-inbred only	Recommended	Not Recommended	Recommended	Recommended, visual inspection is advised	Recommended after genotyping with *OTV_Caller* function
Diploid-outbred or mixture of inbred and outbred	Recommended	Recommended	Recommended	Recommended, visual inspection is advised	Recommended after genotyping with *OTV_Caller* function
Polyploid	Recommended	Requires additional genetic knowledge	Not Recommended	N/A	Recommended after genotyping with *OTV_Caller* function

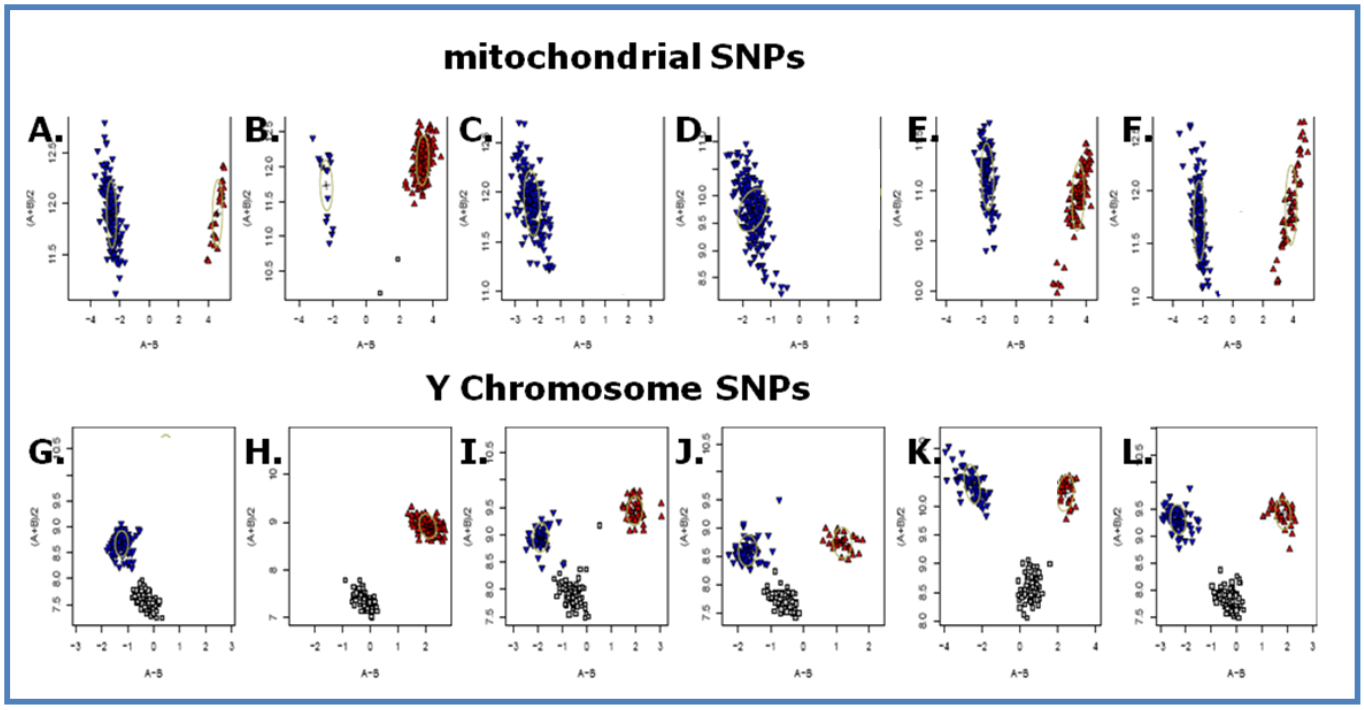
MonoHighRes SNPs are recommended with caution, especially if the best probe set for the SNP site has never been tested. An additional test for recommending MonoHighRes SNPs is to require that both probe sets (if available on the array) for the SNP site are classified as MonoHighRes and that the genotypes agree. Hemizygous SNPs are recommended by default, but visually inspection is advised (see below). SNPs that are classified as OTV may also be considered converted after the *OTV_Caller* function has been used to re-label the genotype calls (see *Adjust Genotype Calls for OTV SNPs* on page 28) 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 PolyHighRes, NoMinorHom, MonoHighRes, 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. Affymetrix still 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.

Figure 3.4 Cluster plots of mitochondrial and Y chromosome SNPs.



Panels A through F of Figure 3.4 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 3.4 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.

Evaluate SNP Cluster Plots

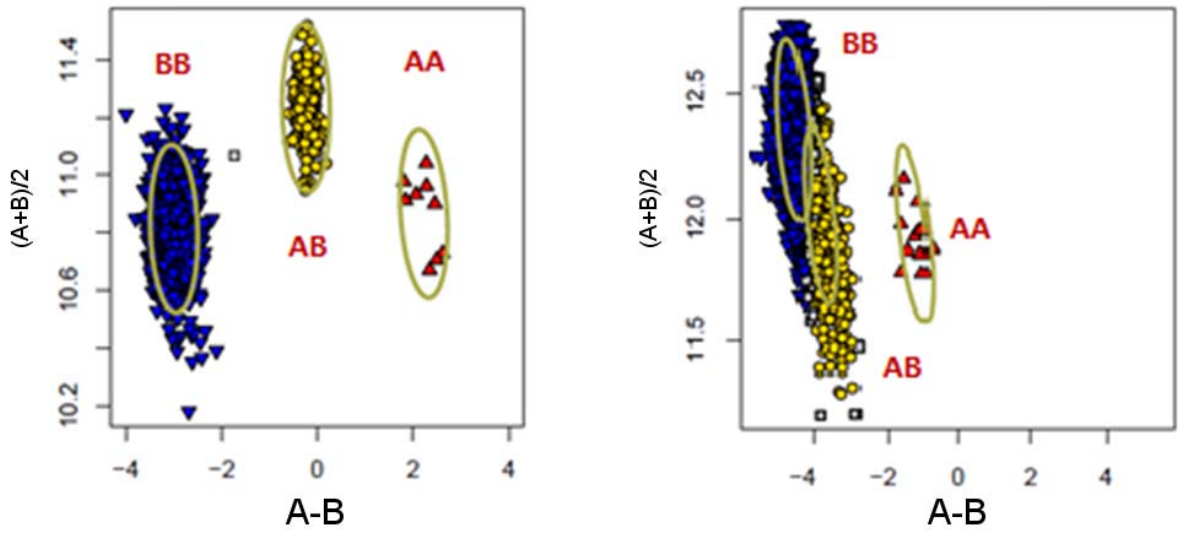
Visualization and understanding of SNP cluster plots (introduced in *What is a SNP Cluster Plot for AxiomGT1 Genotypes?* on page 10) is a key component the Best Practices workflow. Users should view a small number (~200) of cluster plots of randomly selected SNPs from each of the *Ps.Classification* function categories (Figure 3.3) 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 3.3) should have clusters are 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) GTC, via the SNP Cluster Graph function and (2) the SNPolisher package, via the *Ps_Visualization* 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 3.5 shows an example of a probe set for 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 GTC (Chapter 7).

Figure 3.5 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 3.6). 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.

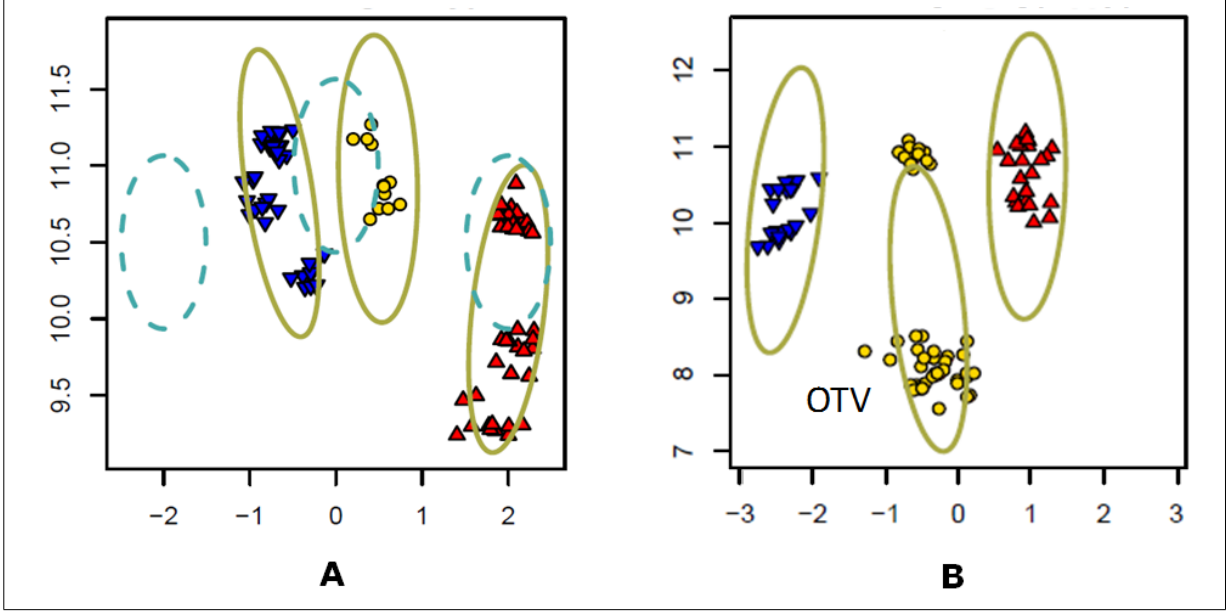
Possible genetic differences may be due to inclusion of subpopulations with copy number variations at the given SNP site, or inclusion of subpopulations whose genomes have diverged from the reference population whose genome sequence was used to design the probes for the array. Genomes of divergent subpopulations may have interfering SNPs and indels relative to the array probe sequences that decrease the genotype intensities. OTV SNP sites (Didion *et al.*, 2012)[1] 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 3.6-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.

[1] *Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, de Villena FP, Churchill GA. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics. 2012 Jan 19;13:34.*

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 3.3), Supplemental filters can be applied. (see the SNPolisher User Guide Section on *Ps_Classification_Supplemental*, note that 64-bit Perl should be installed when using *Ps_Classification_Supplemental*). 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 28).

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 both GTC 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.

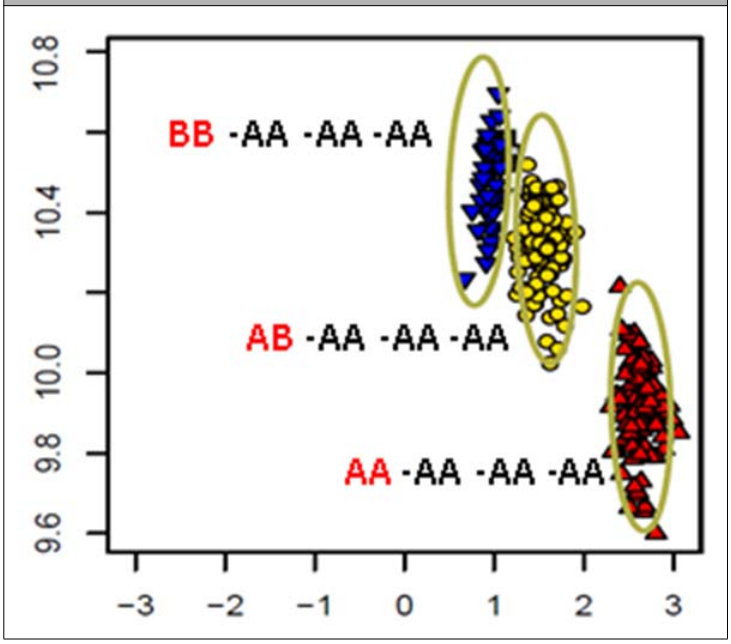
Figure 3.6 Examples of multi-cluster SNPs. A. Diploid plant SNP with 7-8 intensity clusters. **B.** Diploid plant SNP with 4-5 intensity clusters; one is an OTV.



Allo-polyploid SNP Cluster Plot Pattern

Allo-polyploid 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 allo-polyploid SNP sites usually segregate in just one sub-genome, while remaining fixed in the homeologous sites in the other sub-genomes. Allo-polyploid genomes occur in some plant and fish species and produce expected differences in SNP cluster patterns (Figure 3.7), relative to diploid genomes (Figure 3.5 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 allo-polyploid 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).

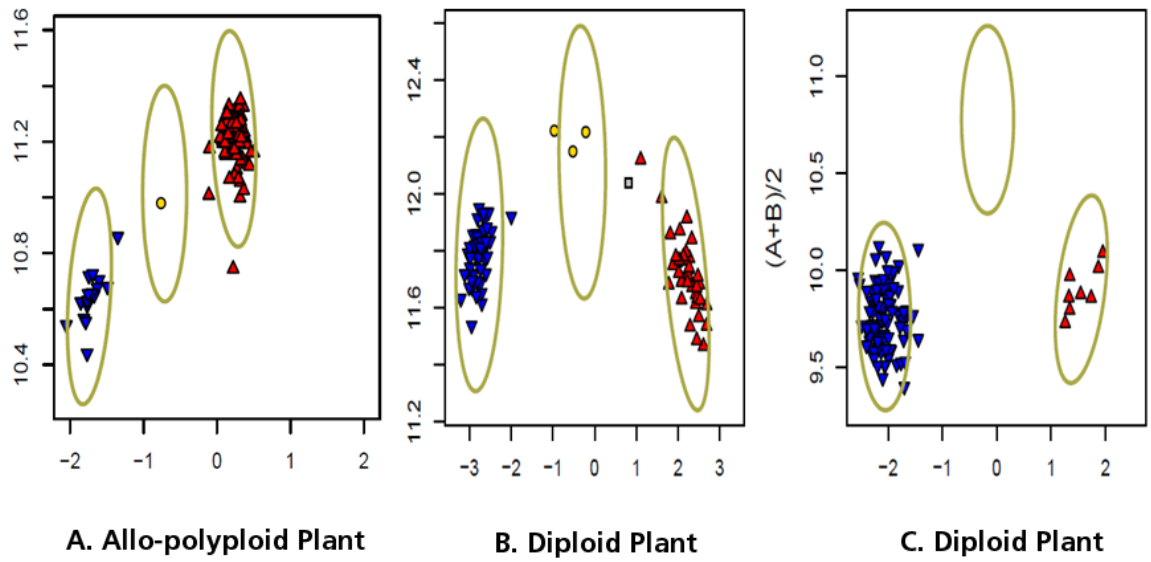
Figure 3.7 Cluster Plot for an allo-octoploid plant. Each sample is colored by the AxiomGT1 genotype call (blue, gold, red) for the sub-genome with the segregating allele. Each genotype cluster is labeled by the likely allo-octoploid genotype using the following notation: the genotypes of 4 sub-genomes 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.



SNP Cluster Plot Patterns for Inbred Populations

Inbred populations produce few or no heterozygous genotypes and there is often a high frequency of both of the homozygous genotypes (Figure 3.8). All cases will be classified as PolyHighResolution as long as the Polyploid or Diploid option is selected in the *Ps_Classification* step. However, the SNP producing cluster plot Figure 3.8-C (with no Heterozygous genotypes) will be classified as Other if the Human option is selected in the *Ps_Classification* step. AxiomGT1 analysis options should be set to include the inbred penalty when genotyping inbred populations. Users should contact their local Affymetrix Field Application Support or send email to Support@affymetrix.com for usage of the inbred penalty option.

Figure 3.8 Cluster Plots for Inbred Populations. A. Allo-polyploid 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 - cases include SNPs with cluster splits and some cases of multi-cluster SNPs such as OTV cluster patterns which escape the OTV classification. Instructions are provided in *Visualize SNPs and Change Calls if Desired through GTC Plotted Cluster Graph* on page 52.

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) are SNP sites (Didion *et al*., 2012)[1], 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 low 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 4.1-A).

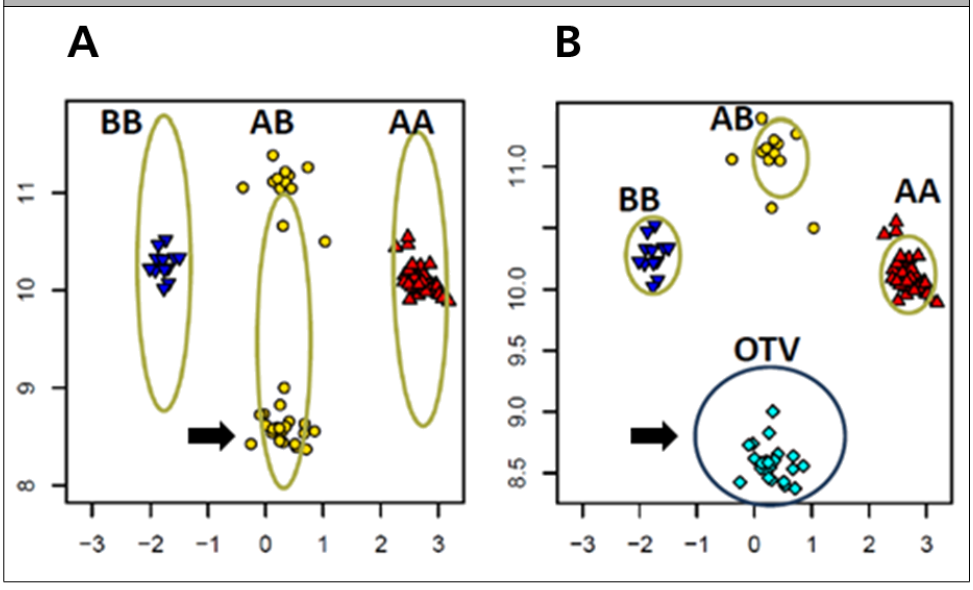
The SNPolisher *OTV_Caller* function performs post-processing analysis to identify miscalled AB clustering and identify which samples should be in the OTV cluster and which samples should remain in the AA, AB, or BB clusters. Samples in the OTV cluster are re-labelled as OTV (Figure 4.1-B).

SNPolisher *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 19).

Instructions for executing OTV calling are provided in the SNPolisher User Guide; see the SNPolisher User Guide 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*.

Figure 4.1 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).



[1] *Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, de Villena FP, Churchill GA. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics. 2012 Jan 19;13:34.*

Genotyping Auto-tetraploids

Auto-polyploids (occurring in some plant and fish species) are polyploids with two sub-genomes generally derived from different 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 allo-polyploid and diploid genomes.

The R package fitTetra (http://cran.r-project.org/web/packages/fitTetra/index.html) produces genotypes for auto-tetraploids 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 *et al.*, 2011)[1].

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.

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 10. 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.

[1] *Voorrips RE, Gort G, Vosman B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics. 2011 May 19;12:172. - ISSN 1471-2105 - p. 11.*

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 GTC conveniently produces a signature SNP report in the pre-genotyping QC process. Affymetrix 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 Laurie *et al*.[1]

Detecting Mixed (Contaminated) DNA samples

This section discuss 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.

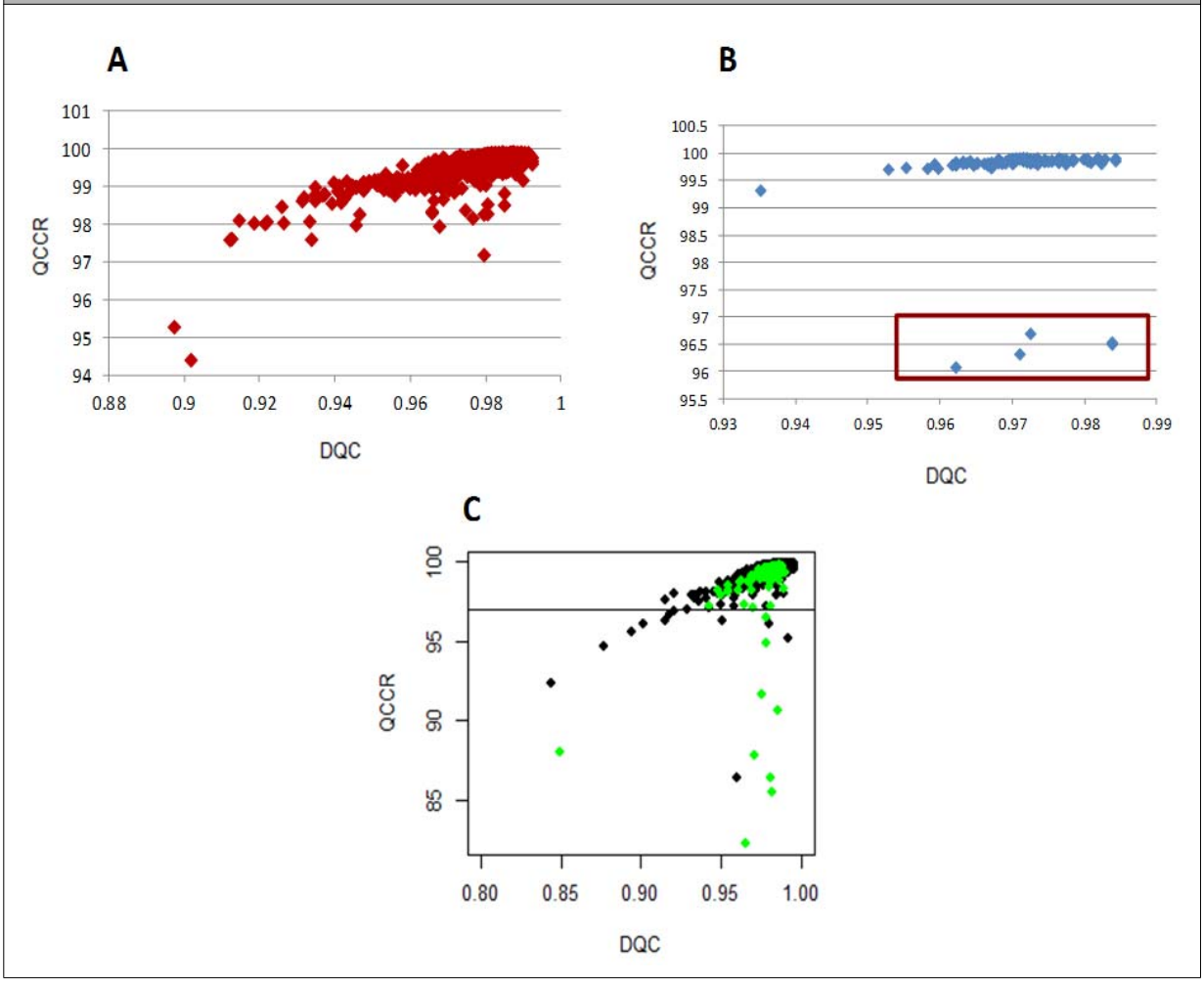
Samples Have Relatively High DQC and Low QC Call Rate (QCCR) Values

In general, higher DQC values correlate with higher sample call rates (see Figure 5.1-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 5.1-B shows the effect of deliberately mixing 4 samples (enclosed in box). Figure 5.1-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 5.1-C, but it has also been observed that large image artifacts on the array surface can produce this pattern as well.

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

Figure 5.1 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.



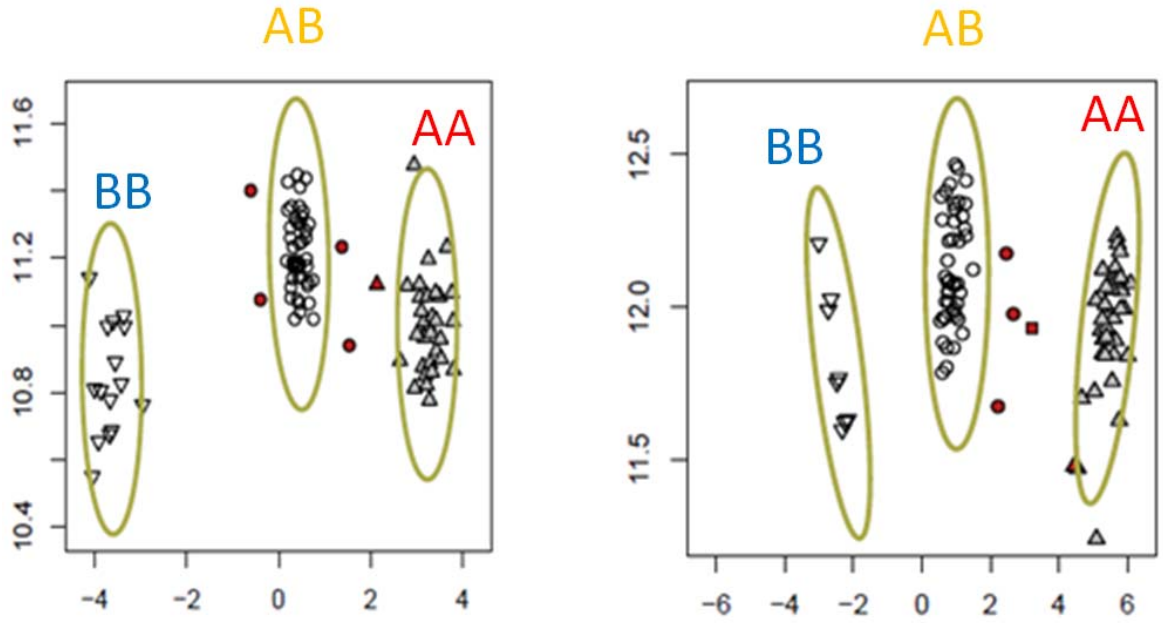
Samples Have a High Percentage of Unknown Gender Calls

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 5.2 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 5.2 Cluster plots for two SNPs. Deliberately mixed samples are colored red. Uncontaminated samples are colored gray.



Unusual Patterns of Relatedness

Cross-contamination of samples can cause samples to appear to be related to each other when examining their genotypes. Depending on the extent of the cross-contamination, it can be just a pair of samples or entire sections of the plate that show increased relatedness. Relatedness can be examined using the method described in the "Relatedness" section of Laurie *et al.*[1]

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 CHP summary table in GTC provides % of AB calls for a sample under the "het_rate" column. "het_rate" is displayed together with sample call rate (call_rate) in the CHP summary table

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.

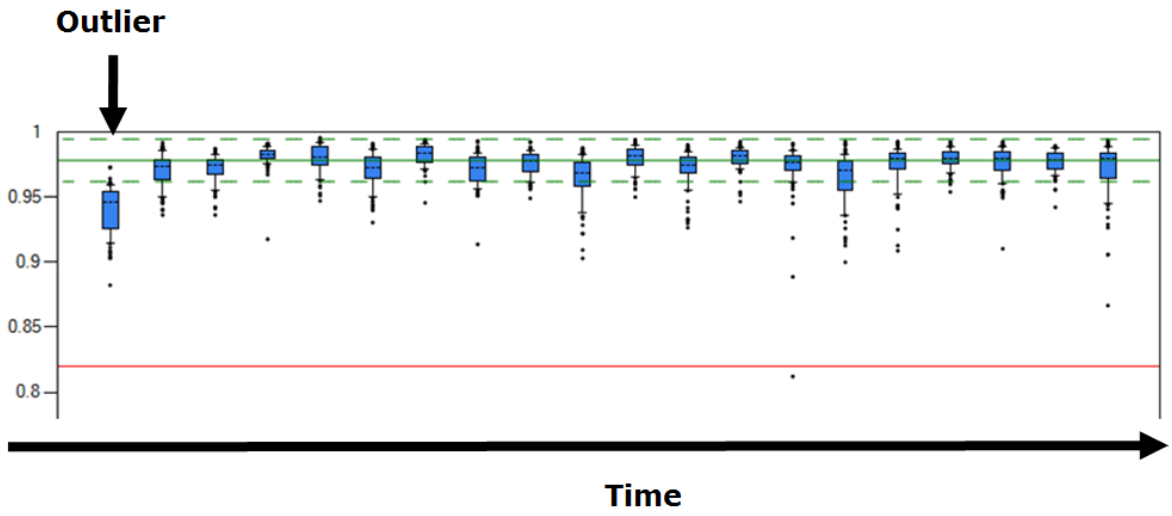
Evaluate Pre-genotyping Performance with DQC Box Plots

Monitoring DQC plate box plots (Figure 5.3) 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 15).

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

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 boxplot 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 16).

Figure 5.3 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.

Check for Platewise MAF Differences

Assuming a randomized study design, the SNP minor allele frequency (MAF) values on a given plate should not systematically differ from the MAF values for the same SNPs on the remainder of the plates. Such a shift in MAFs 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, *et al*., 2008)[1]. A description of this method as described in Laurie *et al*., 2010[2] and summarized here.

To detect batch effects on allelic frequencies, we use a homogeneity test suggested by N. J. Cox (Pluzhnikov *et al*., 2008)[3]. 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 samples), then a 1 degree of freedom chi-squared test statistic is given by $\frac{n_i n_{(i)} (\tilde{p}_i - \bar{p}_{(i)})^2}{[n\bar{p}(1-\bar{p})]}$ 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.

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

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

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

SNP QC Metrics

SNP Metrics Used in the *Ps_Classification* Step (Step 8C)

SNP Call Rate (CR)

SNP Call Rate = #Samples Called/N

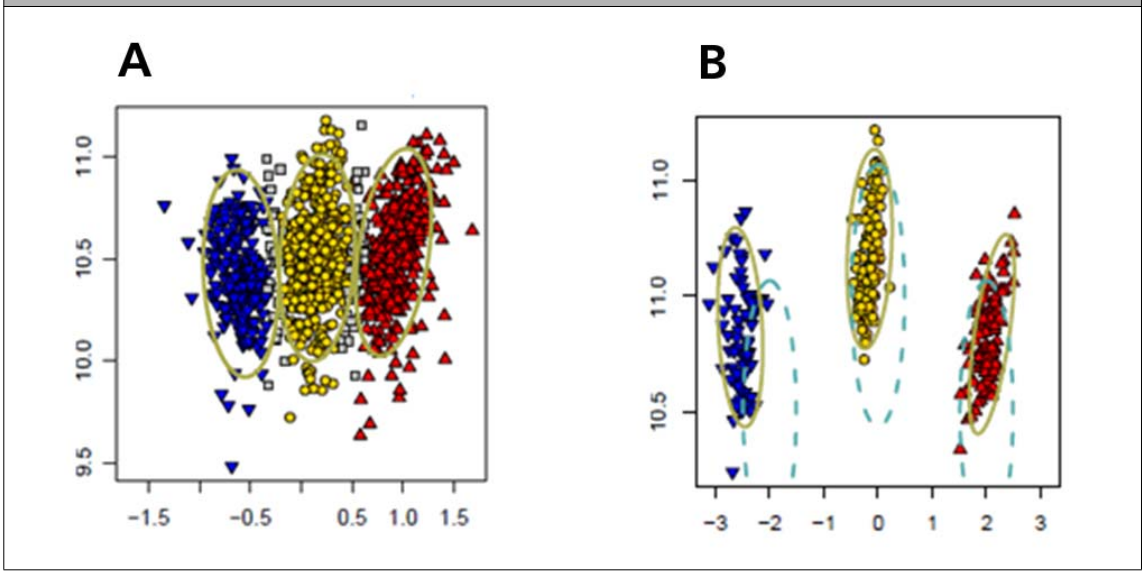
#Samples Called = the number of samples assigned a genotype call of either AA, BB or AB at the SNP locus. That is the number of samples that do not have a "No Call" assignment.

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 6.1-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 6.1 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 3.1 on page 21 for default CR thresholds used in the *Ps_Classification* step.

Fisher's Linear Discriminant (FLD)

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

Where: M_{ab} = center of het cluster in log ratio dimension;
M_{aa}, M_{bb} = center of hom a,b cluster in log ratio dimension; sd = square root of variance pooled across all three distributions. FLD is undefined when there is only one genotype cluster.

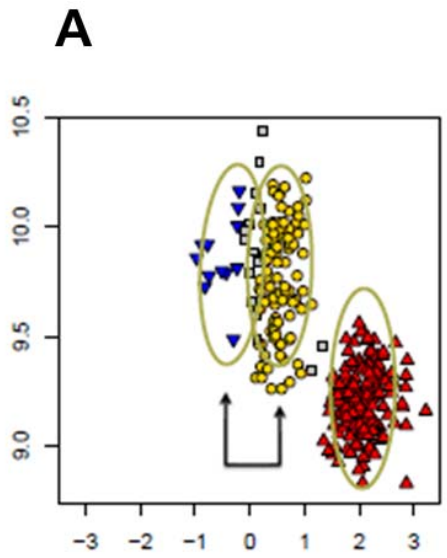
FLD is a measurement of the cluster quality of a SNP. High-quality SNP clusters have well-separated centers, and the clusters are narrow. High-quality clusters can be identified by examining the shape and separation of the SNP posteriors that are produced during genotyping.

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.

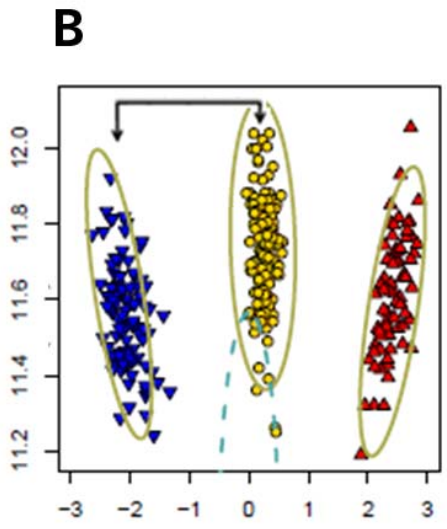
HomFLD is a version of FLD computed for the homozygous genotype clusters. HomFLD is undefined for SNPs without two homozygous clusters.

Figure 6.2-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. In contrast, the well-clustered SNP in Figure 6.2-B has a high CR and separated cluster centers, producing an FLD of 7.97.

Figure 6.2 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 cluster centers are far apart from each other in the X dimension.

Heterozygous Strength Offset (HetSO)

SNP HetSO (Het Strength Offset)

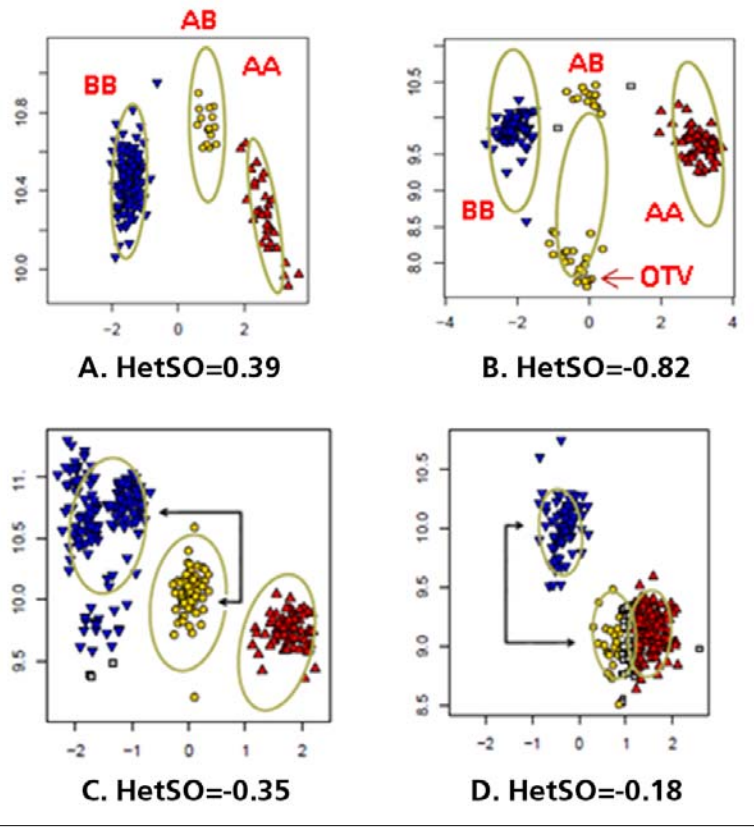
$$\text{HetSo} = A_{ab} - A_{bb} - (A_{aa} - A_{bb}) \times \left(\frac{M_{ab} - M_{bb}}{M_{aa} - M_{bb}} \right)$$

Where $(M_{aa},\ A_{bb})$ = center of aa cluster, etc.

Heterozygous strength offset (HetSO) measures how far the heterozygous cluster center sits above or below the homozygous cluster centers in the Y dimension. 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 6.3-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 than for the homozygous clusters (Figure 6.3-B, Figure 6.3-C, Figure 6.3-D). Figure 6.3-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 6.3-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 6.3-D shows a larger homozygous cluster in blue (BB) and a large cluster that has been split between heterozygous AB calls (yellow) and homozygous AA calls (red). This cluster split has caused the true heterozygous cluster to be called as the homozygous cluster. This produces a HetSO value of –0.18. The low HetSO values of SNP clusters in Figure 6.3-C and Figure 6.3-D help flag these cases as problematic SNPs.

Figure 6.3 Examples of SNPs with different HetSO values.



Homozygote Ratio Offset (HomRO)

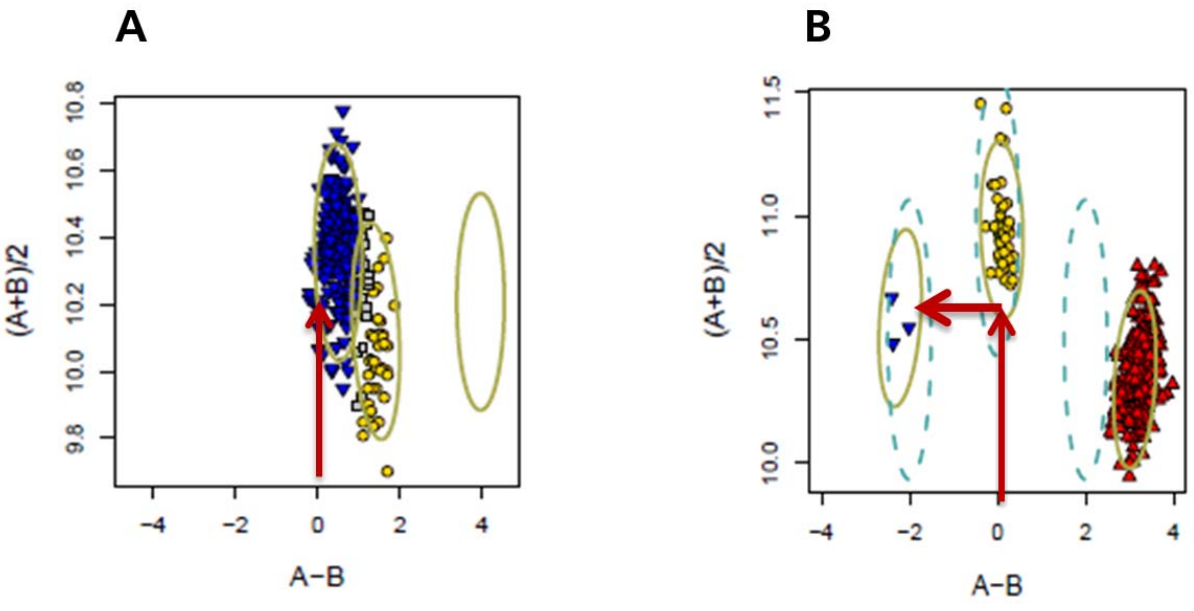
$$\text{HomRO} = \begin{cases} \text{Min}(Xaa, \text{abs}(Xbb)) & \text{If both hom clusters are on correct side of zero} \\ -Xbb & \text{If both hom clusters are to the right of zero} \\ Xaa & \text{If both hom clusters are to the left of zero} \end{cases}$$

Where Xaa is the center of the AA cluster on the X-axis (Size Dimension) and Xbb is the center of BB cluster on the X-axis (Size)

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 6.4-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 6.4-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.

Figure 6.4 Examples of SNPs with low HomRO (A) and high HomRO (B).



SNP with low HomRO (–0.51).

SNP with high HomRO (2.21).

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 GTC. No Axiom software is provided for reproducibility calculation and 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 GTC, the SNP summary statistics (including SNP call rate and HWE values) can be viewed in the SNP summary table. Prior to viewing the SNP summary table in GTC, create a SNP list (i.e., a list of the SNPs of interest that have been retained after SNP QC filtering) using the **Create SNP List…** menu option in GTC. Next, select **Show SNP Summary Table** from the same menu to display the SNP call rate and other summary metrics.

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)[1] 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

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

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.

Instructions for Executing Best Practices Steps with GTC and APT Software

This section provides an overview of the QC and genotyping workflows to be used in Affymetrix® Genotyping Console™ (GTC) version 4.2 or higher and Affymetrix® Power Tools (APT) version 1.16.1 or higher.

Execute Steps 1-7 with GTC Software

GTC Setup

Analysis library files and annotation files can be directly downloaded from within GTC 4.2, or they can be manually downloaded from www.affymetrix.com and unzipped into the current GTC library folder.

For more detailed instructions on how to install GTC and obtain analysis library and annotation files, please consult the *Genotyping Console™ User Guide* (P/N 702982), available at the support section of the Affymetrix website (www.affymetrix.com)

Step 1: Group Samples into Batches in GTC

Samples can be grouped into batches either within or outside GTC. In order to group samples into different batches within GTC:

- import all CEL files from the study into a dataset, and
- then add the selected CEL files to individual custom intensity data groups (batches) in the Intensity QC table.

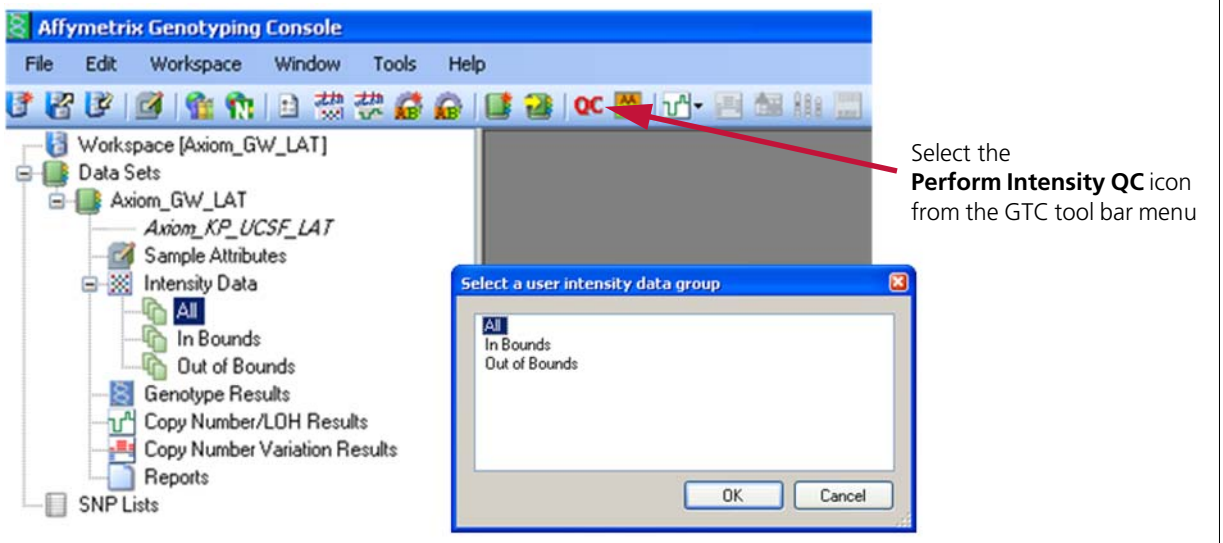
Alternatively, samples can be grouped into batches outside GTC, and only the CEL files of interest should be imported into a dataset during GTC data import.

Steps 2 and 3: Generate DQC Values and QC the Samples Based on DQC in GTC

To perform a sample quality check in GTC using DQC values, import the CEL files of interest into a dataset created for the specific Axiom® array. For instructions on creating a dataset in GTC and for general instructions on working with GTC, consult the *Genotyping Console™ User Guide* (P/N 702982). GTC will calculate the quality control metrics at the time of import (or at any other time) by performing any one of the following operations:

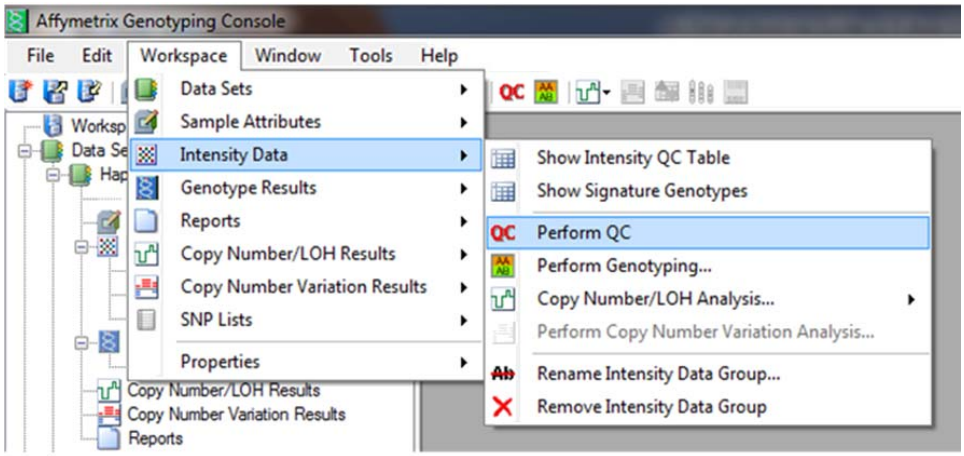
- Select the QC icon  from the tool bar menu (Figure 7.1).

Figure 7.1 How to select the 'Perform QC' command from the GTC tool bar menu.



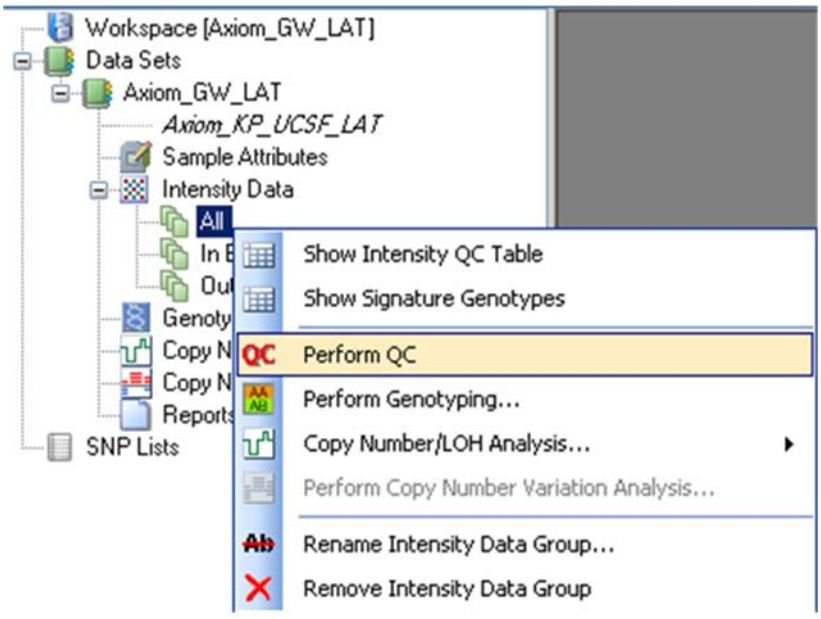
- Select the **Intensity Data** option from the **Workspace** drop-down menu, and then select **Perform QC** (Figure 7.2)

Figure 7.2 How to select the **Perform QC** command from the Workspace drop-down menu.



- Right-click the **All** data group displayed in the data tree on the left and select the **Perform QC** option from the popup menu that appears (Figure 7.3).

Figure 7.3 How to select the **Perform QC** command from the data tree popup menu.



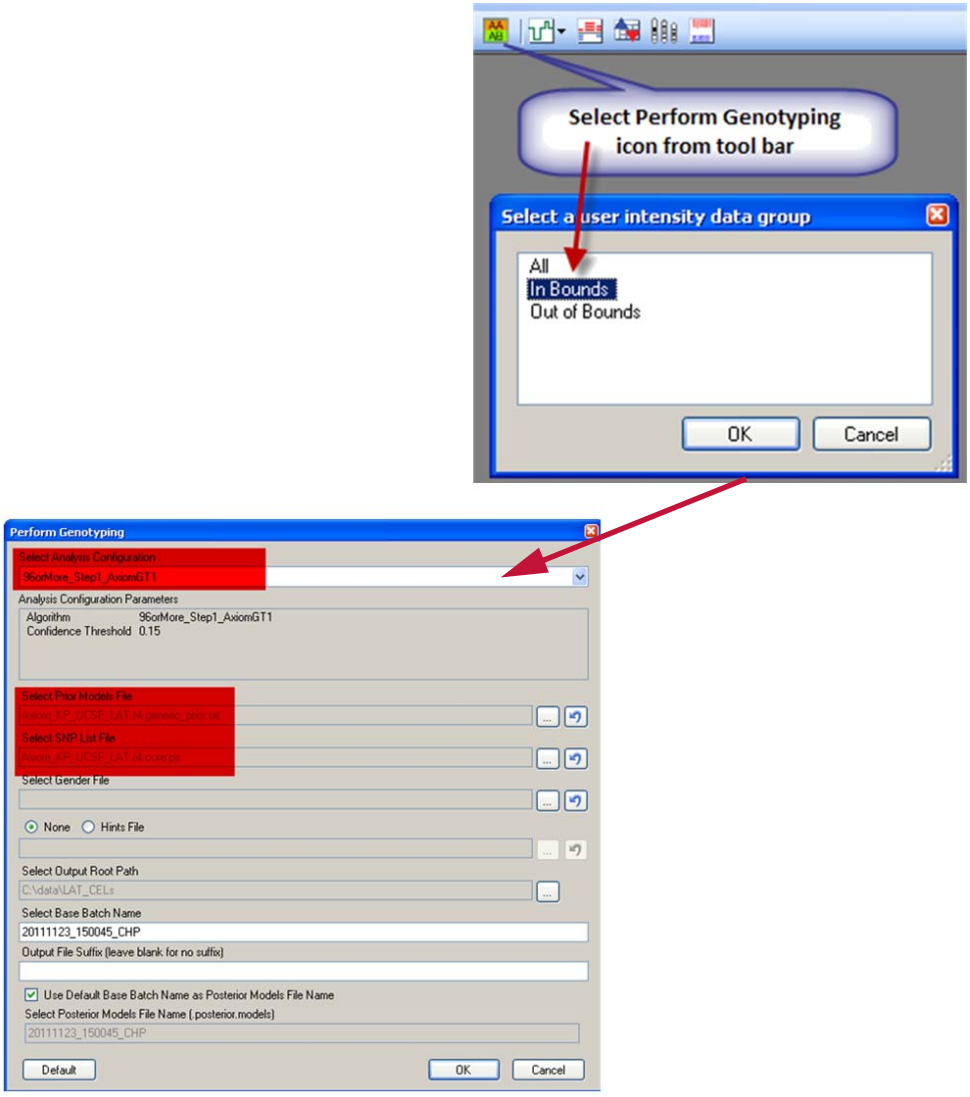
After the QC process has been completed, the resulting DQC values and other quality metrics are displayed by GTC in an Intensity QC table. Any samples with a DQC value < 0.82 are highlighted and grouped into the **Out of Bounds** group to be excluded from the genotyping analysis. For further information about the QC thresholds and the other metrics, please refer to the *Genotyping Console™ User Guide* (P/N 702982).

Steps 4, 5, and 6: Generate QC Sample Call Rates, QC the Samples, and QC the Plates

Perform QC genotyping for the group of CEL files that have passed the DQC threshold (DQC ≥ 0.82, i.e., those CEL files in the **In Bounds** data group) by performing any one of the following operations:

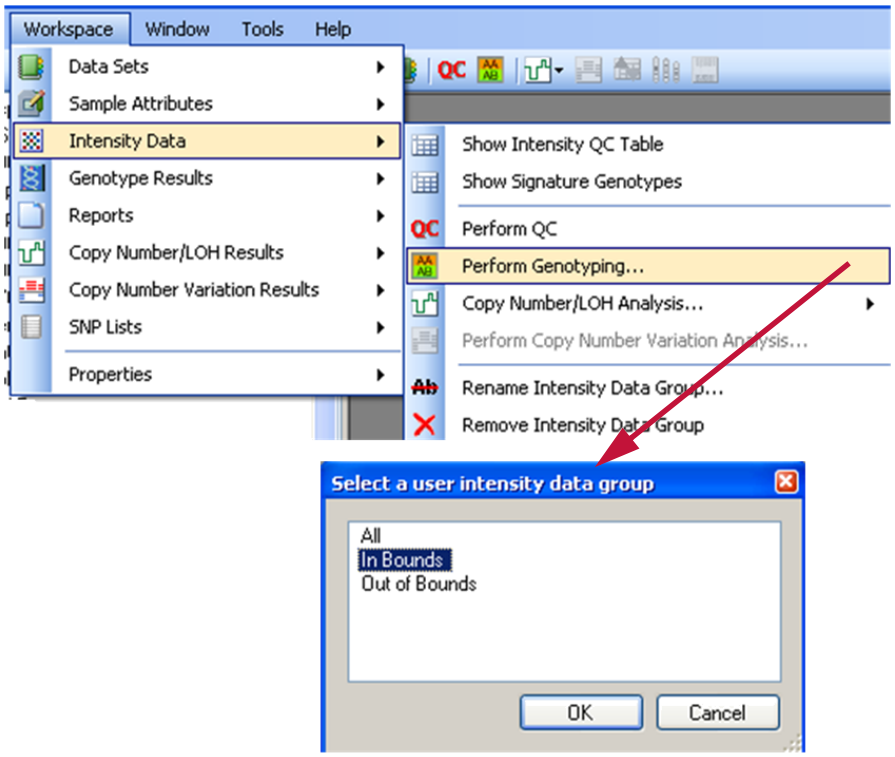
- Select the Genotyping icon on the tool bar menu (Figure 7.4).
  - Choose the **96orMore_Step1_AxiomGT1** as the analysis configuration from the drop-down menu to perform QC genotyping with generic priors model file if the batch size is equal or more than 96 samples (Figure 7.4).
  - Choose the **LessThan96_Step1_AxiomGT1** as the analysis configuration from the drop-down menu to perform QC genotyping with SNP-specific models file if the batch size is less than 96 samples.

Figure 7.4 Perform genotyping from the GTC tool bar menu.



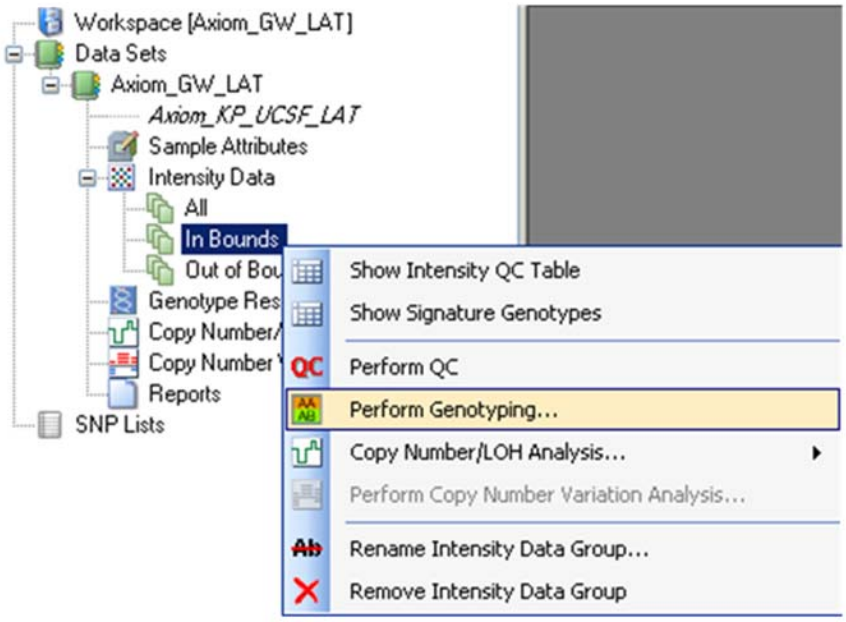
- Select the **Intensity Data** option from the **Workspace** drop-down menu, and then select **Perform Genotyping…**, select **In Bounds**, and click **OK** (Figure 7.5).
  - Choose the **96orMore_Step1_AxiomGT1** as the analysis configuration from the drop-down menu if the batch size is equal to or more than 96 samples. (Refer to Figure 7.4 for more details.)
  - Choose the **LessThan96_Step1_AxiomGT1** as the analysis configuration from the drop-down menu to perform QC genotyping with SNP-specific models file if the batch size is less than 96 samples.

Figure 7.5 Performing genotyping from the **Workspace** pop out menu.



- Right-click the **In Bounds** data group displayed in the data tree, and select the **Perform Genotyping…** option (Figure 7.6).
  - Choose the **96orMore_Step1_AxiomGT1** as the analysis configuration in the drop-down menu if the batch size is equal to or more than 96 samples. (Refer to Figure 7.4 for more details.)
  - Choose the **LessThan96_Step1_AxiomGT1** as the analysis configuration from the drop-down menu to perform QC genotyping with SNP specific models file if the batch size is less than 96 samples.

Figure 7.6 Performing genotyping by right-clicking the **In Bounds** data group in the data tree and selecting the **Perform Genotyping** command from the drop-down menu.



After the QC genotyping has been completed in GTC, identify outlier samples with a sample call rate lower than 97% in the CHP summary table. Next, create an intensity data group that excludes the CEL files for these outlier samples. After the sample level QC is done, it is recommended to perform plate level QC for each plate to calculate the following metrics using Table 7.1.

Table 7.1

Metric	Calculation	Minimum Criteria
Plate pass rate	$\frac{\text{Average call rate of passing samples on the plate}}{\text{Total samples on the plate}} \times 100$	None
Average call rate of passing samples on the plate	AVERAGE (call rates of samples passing DQC and 97% call rate)	≥ 98.5%

NOTE: The table gives the minimum criteria for a passing plate; please see *Step 6: QC the Plates* on page 16 for more discussion about these plate metrics. Also see *Additional Plate QC* on page 32. These metrics can be manually calculated in Excel or with a script using data from the CHP summary table.

Step 7: Genotype all Passing Samples and Plates from the Same Batch Using All SNPs

The purpose of this step is to perform genotyping using all SNPs on the array and only high-quality samples.

Using the intensity data group that includes only the CEL files with sample call rates > 97% after the QC genotyping, perform genotyping using all the SNPs on the array. Figure 7.7 summarizes the steps to perform genotyping:

Step 7A: Double-click the genotyping result batch in the data tree (Figure 7.7 Panel A) to open the CHP summary table (Panel B) that is generated after performing the QC genotyping on all the CEL files that pass the DQC cutoff (In Bounds).

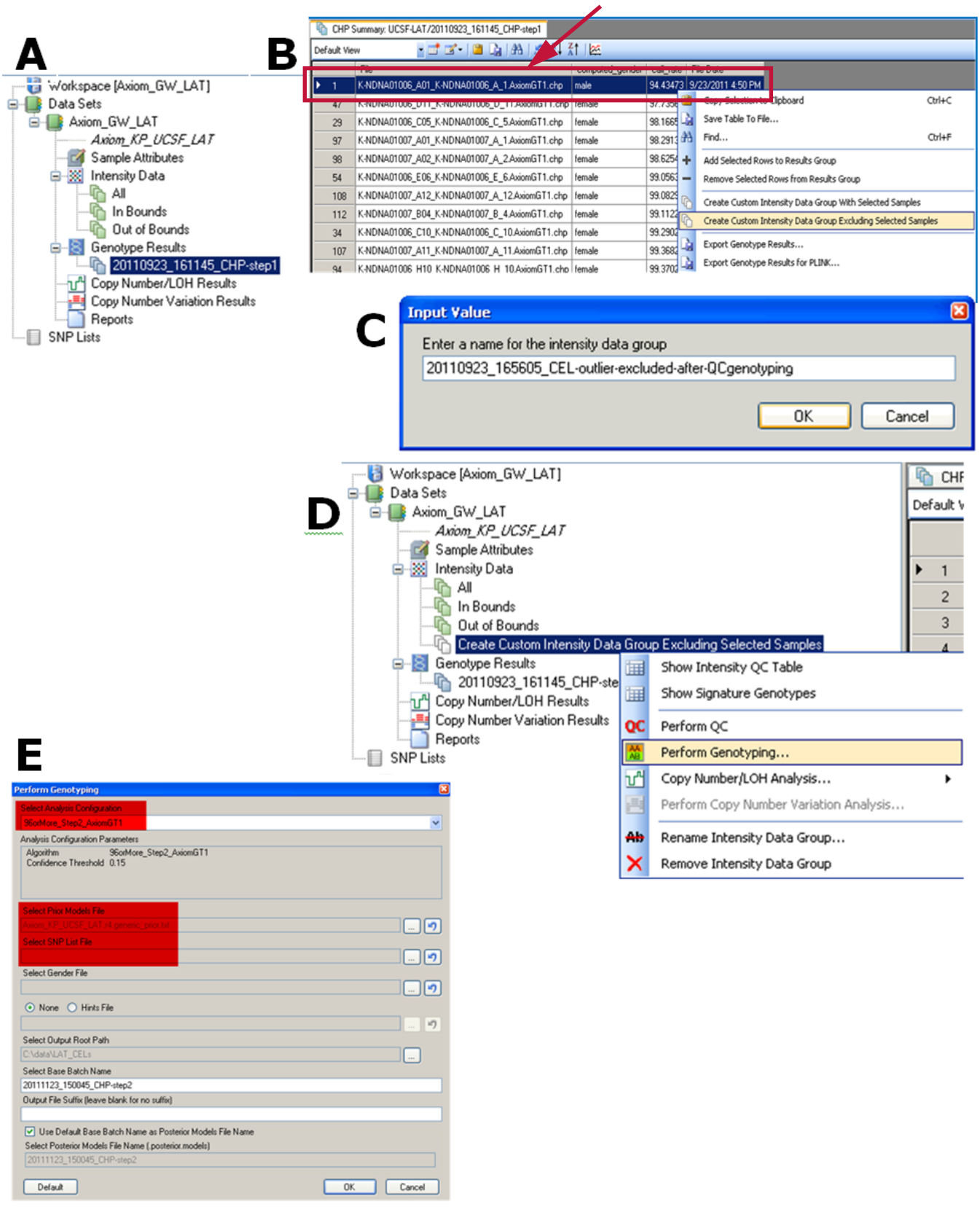
Step 7B: Identify the outlier samples (i.e., those samples with a sample call rate < 97%) by sorting the CHP summary table by sample call rate. (Select the **call_rate** column and click the **sort ascending** icon in the tool bar [red arrow in Figure 7.7 Panel B].) In Figure 7.7, there is one highlighted outlier sample that has a sample call rate below the 97% cutoff. To exclude outlier samples, select all of the samples with a call rate < 97%, and then right-click in the CHP summary table and choose the **Create Custom Intensity Data Group Excluding Selected Samples** menu option. Provide a name for the newly created intensity data group (Figure 7.7 Panel C); for example, "20110923_165605_CEL-outlier-excluded-after-QCgenotyping".

Step 7C: Go back to the data tree and right-click the newly created "20110923_165605_CEL-outlier-excluded-after-QCgenotyping" intensity group to perform genotyping (Figure 7.7 Panel D).

Step 7D: Select the **96orMore_Step2_AxiomGT1** analysis configuration to perform genotyping for this group of high-quality samples (≥ 96 samples) using the generic prior models file and all SNPs on the array (Figure 7.7 Panel E).

NOTE: Select the LessThan96_Step2_AxiomGT1 analysis configuration to perform genotyping for smaller groups of high-quality samples (< 96 samples) using SNP-specific models file and all SNPs on the array.

Figure 7.7 Overview of steps to perform genotyping on passing samples only using all SNPs and generic priors.



Execute Step 8 with APT Version 1.16.1 or Higher

In this section, we provide instructions for executing steps 8 of the best practice analysis workflow using *Ps_Metrics* and *Ps_Classification* programs from APT Version 1.16.1 or higher combined with some simple scripts (to be written by the user). See Affymetrix Power Tools information for download instructions.

Note that the same Step 8 results are produced when executing SNPolisher functions *Ps_Metrics* and *Ps_Classification*. If SNPolisher usage is preferred please go to *Execute Best Practice Step 8 with SNPolisher Functions* on page 65.

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.

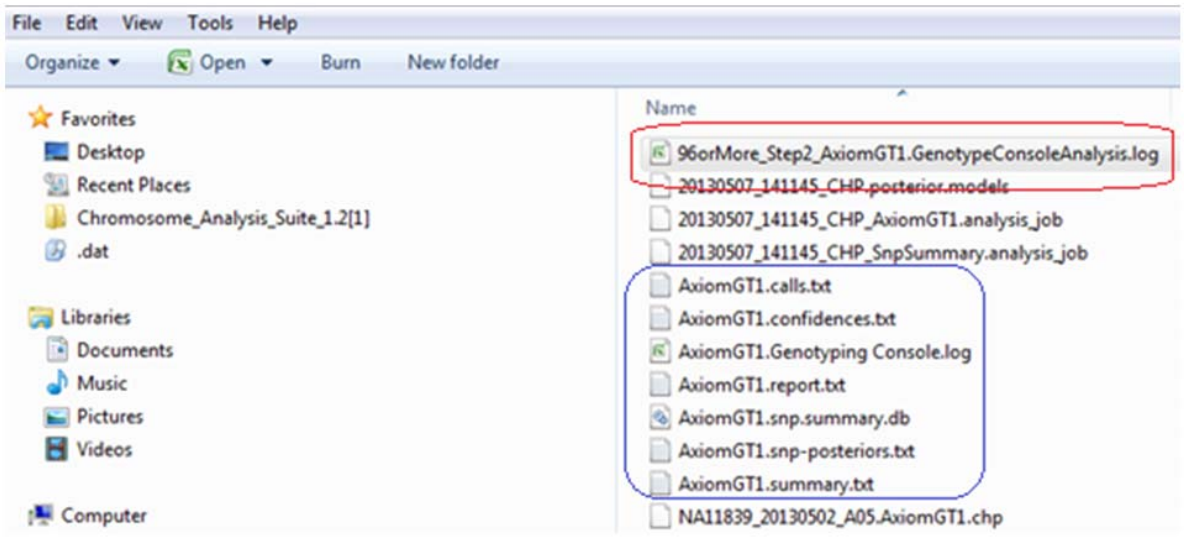
To execute the commands/scripts Windows users should:

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.

Locate GTC Step2_AxiomGT1 (Step 7) Output Files

The files used to run the Step 8A Ps_Metric software are generated from running Step2_AxiomGT1 (*Step 7: Genotype all Passing Samples and Plates from the Same Batch Using All SNPs* on page 47) genotyping in GTC as discussed above (also see chapter 7, Genotyping Analysis, in the *Genotyping Console™ User Guide* (P/N 702982)). The user must take care to use the *Step2_AxiomGT1* output files as input for Step 8 (enclosed in blue box in Figure 7.8). The required files are *AxiomGT1.calls.txt*, and *AxiomGT1.snp-posteriors.txt*. Note that *AxiomGT1.confidences.txt*, and *AxiomGT1.summary.txt* are used with some SNPolisher functions.

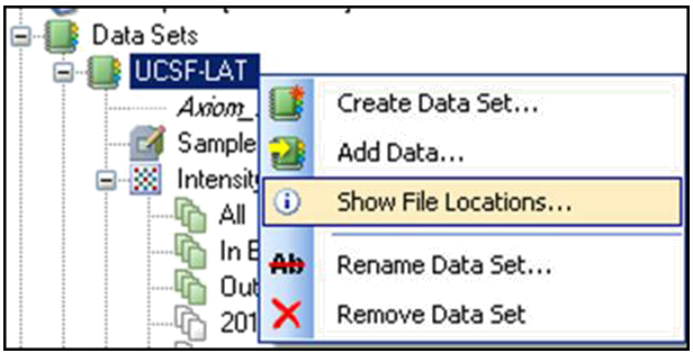
Figure 7.8 The Step2_AxiomGT1 (Step 7) output files.



The file in the red circle is the Step 2 GTC analysis log, and the files in the blue circle are the AxiomGT1 output files.

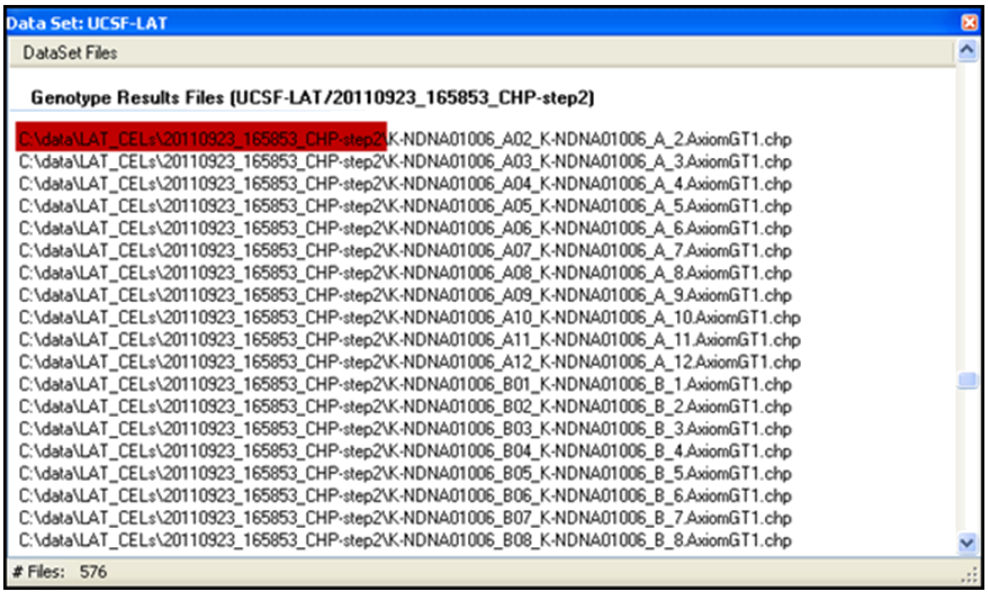
To locate the exported text files to use as input for Ps_Metric, right click the dataset and choose **Show File Locations...** option (see Figure 7.9-A); use the file location data to identify the genotype result folder (see Figure 7.9-B); then navigate to the folder and confirm that all Step 2 output files are present (see Figure 7.9-C).

Figure 7.9 How to locate the **Step2_AxiomGT1** (Step 7) output files.

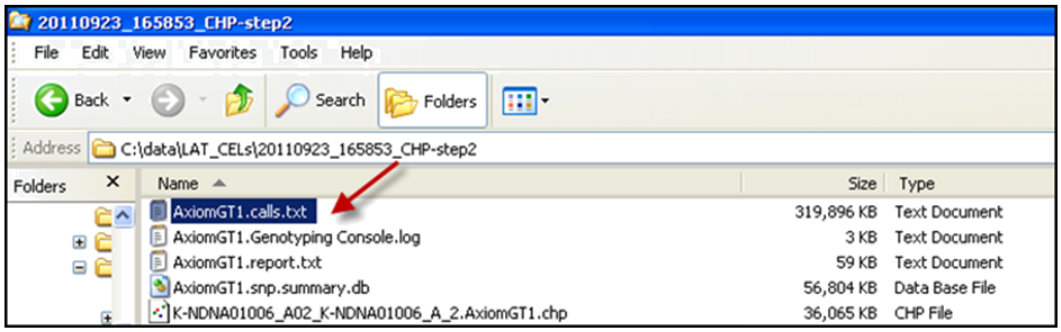


A

Right-click the dataset and select **Show File Locations...**



B Use the file location data to identify the folder where the files are stored.



C Navigate to the folder and confirm that the *Step2_AxiomGT1* (Step 7) files are all there.

Step 8A: Run *Ps_Metrics*

Example *Ps_Metrics* Script

In this example the working directory is "./input". And the output is written to "./input/metrics.txt". The Ps_Metric program is available in APT versions 1.16.1 and higher.

```
Ps_Metrics \
--posterior-file ./input/AxiomGT1.snp-posteriors.txt \
--call-file ./input/AxiomGT1.calls.txt \
--metrics-file metrics.txt
```

The following parameters are required to run *Ps_Metrics*:

- **posterior-file** - The file output by GTC/APT which contains posterior summaries of clusters for SNPs (usually named "AxiomGT1.snp-posteriors.txt") produced by Step 7 genotyping.
- **metrics-file** - The name of the output file. The user must supply a file name in order to save the output as a separate file, which is read in to other SNPolisher functions such as *Ps_Classification*.

In addition, one of the following two parameters is required:

- **call-file** - The file output by GTC/APT which contains genotype calls for SNPs (usually named "AxiomGT1.calls.txt") produced by Step 7 genotyping. The call file may use either the labels "AA", "AB", "BB" and "NoCall", or the numeric labels 0 (AA), 1 (AB), 2 (BB) and -1 (NoCall).
- **chp-files** - A text file containing a list of CHP files from which to take calls (one file path per line). CHP files are produced by GTC.

NOTE: Either the --call-file or the --chp-files parameter must be provided in order to run *Ps_Metrics*. Computational performance is much better using the call.txt file rather than the chp file. So if it is an option we recommend using the calls.txt file.

Step 8B: Run *Ps_Classification*

Example *Ps_Classification* Script

To run *Ps_Classification* on the metrics file metrics.txt for human samples, which will generate the output files in the folder named "output" in the working directory. The *Ps_Classification* program is available in APT versions 1.16.1 and higher.

```
Ps_Classification \
--species-type Human \
--metrics-file metrics.txt \
--output-dir ./output \
-- ps2snp-file <ANALYSIS_FILES_DIR>/<axiom_array>.r<#>.ps2snp_map.ps
```

<ANALYSIS_FILES_DIR> is full path to the analysis lib file directory (Table 1.1 on page 7).

The following 8 files will be written to ./output directory

- *PolyHighResolution.txt* (the table with probe set performance and classification results)

and seven category files:

- *CallRateBelowThreshold.txt*
- *Hemizygous.txt*
- *MonoHighResolution.txt*
- *NoMinorHom.txt*
- *OffTargetVariant.txt*
- *Other.txt*
- *PolyHighResolution.txt.*

The output is discussed in *Step 8B: Classify SNPs Using QC Metrics* on page 19.

The main parameters accepted by *Ps_Classification* are:

- **metrics-file** - The name of the file containing output generated by *Ps_Metrics*.
- **ps2snp-file**- The name of a file containing the mapping of probe set IDs to SNP IDs. The column names should be "probeset_id" and "snpid". The user must supply this file when classifying SNPs that have 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 1.1 on page 7). If this file has not been provided, users should contact their local Affymetrix Field Application Support or send email to Support@affymetrix.com
- **output-dir** - Output directory for results files.
- **species-type** - The type of organism genotyped. The current options are "Human", non-human "Diploid", and "Polyploid".
- **output-converted** - Flag indicating whether to output a list of converted SNPs to a file converted.ps: default is FALSE.
- **cr-cutoff** - Threshold for call rate. If not specified, the default for human is 95 and for diploid and polyploid is 97.
- **fld-cutoff** - Threshold for FLD: default=3.6.
- **het_so_cutoff** - Threshold for HetSO: default=-0.1.
- **het-so-otv-cutoff** - Threshold for OTV detection: default=-0.3.
- **hom-ro-1-cutoff** - Threshold for HomRO when a SNP has one genotypes: default=0.6.
- **hom-ro-2-cutoff** - Threshold for HomRO when a SNP has two genotypes: default=0.3.
- **hom-ro-3-cutoff** - Threshold for HomRO when a SNP has three genotypes: default=-0.9.
- **hom-ro** - Flag indicating whether the metric HomRO is used in classification: default is TRUE.
- **hom-het** - Flag indicating whether the metric HomHet is used in classification. The HomHet metric identifies two-cluster SNPs/probe sets with one homozygote cluster and one heterozygote cluster. This checks if the minor homozygote cluster is missing, which is unreasonable for highly inbred species (e.g., wheat). This metric should be turned on when classifying SNPs/probe sets in highly inbred species: default is TRUE.
- **num_minor-allele-cutoff** - Threshold for the number of minor alleles: default=2.
- **priority-order** - When performing probe set selection, the best probe set is selected according to the priority order of probe set conversion types. Valid values are sequences of the strings "PolyHighResolution", "NoMinorHom", "OTV", "MonoHighResolution" and "CallRateBelowThreshold", separated by commas, with each string occurring exactly once: default is 'PolyHighResolution,NoMinorHom,OTV,MonoHighResolution,CallRateBelowThreshold"

Ps_Classification reads in the SNP QC metrics table (usually named metrics.txt) generated by *Ps_Metrics* and performs SNP classification based on the customizable criteria. It classifies SNPs/probe sets into seven major categories: "PolyHighResolution", "NoMinorHom", "MonoHighResolution", "OTV", "CallRateBelowThreshold", "Other", and "Hemizygous". The metric "hom-ro" should be turned off for organisms where the homozygote clusters may center at zero in the contrast space, particularly polyploid species. If a ps2snp file is provided, then the best performance probe set will be selected to represent SNPs with multiple probe sets. There are seven output files, one for each category, containing lists of probe set IDs belonging to each category. There is also a summary output file.

Visualize SNPs and Change Calls if Desired through GTC Plotted Cluster Graph

GTC contains a function for plotting SNP cluster graphs (*What is a SNP Cluster Plot for AxiomGT1 Genotypes?*) that produces plots that are similar to the output from the R function *Ps_Visualization*. However, the SNP Cluster Graph function in GTC 4.2 has more functionality than the *Ps_Visualization* function. We strongly suggest using the GTC 4.2 *SNP Cluster Graph* function with the output generated by *Ps_Metrics* and *Ps_Classification*.

For a more detailed introduction to the *SNP Cluster Graph* function, see the *Genotyping Console™ User Guide 4.2* (P/N 702982), chapter 9, Using the SNP Cluster Graph.

SNP Cluster Graph allows the user to adjust the shape and color of the samples and posterior and prior distributions, and users can select and change the calls of samples through the plotted cluster graph. See *Evaluate SNP Cluster Plots* on page 23 for interpretation of cluster graphs.

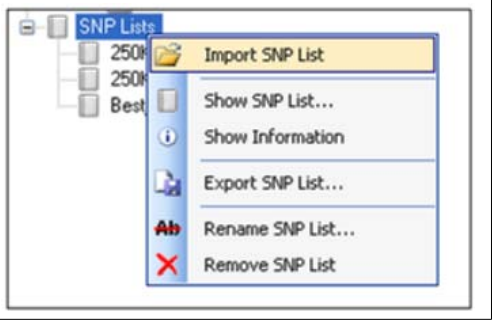
In order to visualize probe sets listed in the category files from APT *Ps_Classification*: *PolyHighResolution.txt, NoMinorHom.txt, Hemizygous.txt, MonoHighResolution.txt, CallRateBelowThreshold.txt, Other.txt*, and *OTV.txt* (described in *Step 8B: Classify SNPs Using QC Metrics* on page 19), these files must be imported into GTC as a SNP list. The GTC SNP list file must be a text file with a .txt or .tsv extension and contain a column labeled "Probe Set ID". The category files from APT *Ps_Classification* software (described above) follows these conventions and can be imported directly into GTC.

The SNPolisher *Ps_Classification* function also outputs the 7 category files named: *PolyHighResolution.ps, NoMinorHom.ps, Hemizygous.ps, MonoHighResolution.ps, CallRateBelowThreshold.ps, Other.ps*, and *OTV.ps*. These SNPolisher files require manual changes before they can by imported into GTC. The default column header for *Ps_Classification* category files is "probeset_id", and the default file extension is .ps. If the GTC flag in SNPolisher *Ps_Classification* was not set to TRUE, the column header must be updated by hand. Open the category file in a text editor such as Notepad or Word and change *probeset_id* to Probe Set ID then save with a .txt extension. If the GTC flag in *Ps_Classification* was set to TRUE, the category files still must be renamed with a .txt or .tsv extension.

To Import a SNP List

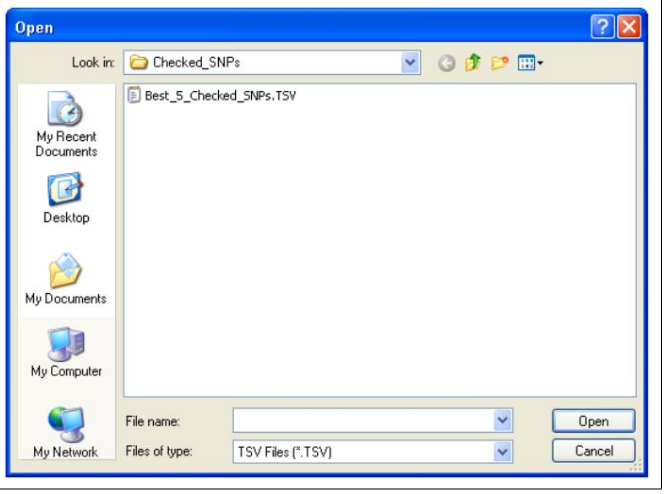
1. Right-click **SNP Lists** in the data tree and select **Import SNP List** (see Figure 7.10).

Figure 7.10 Select SNP List



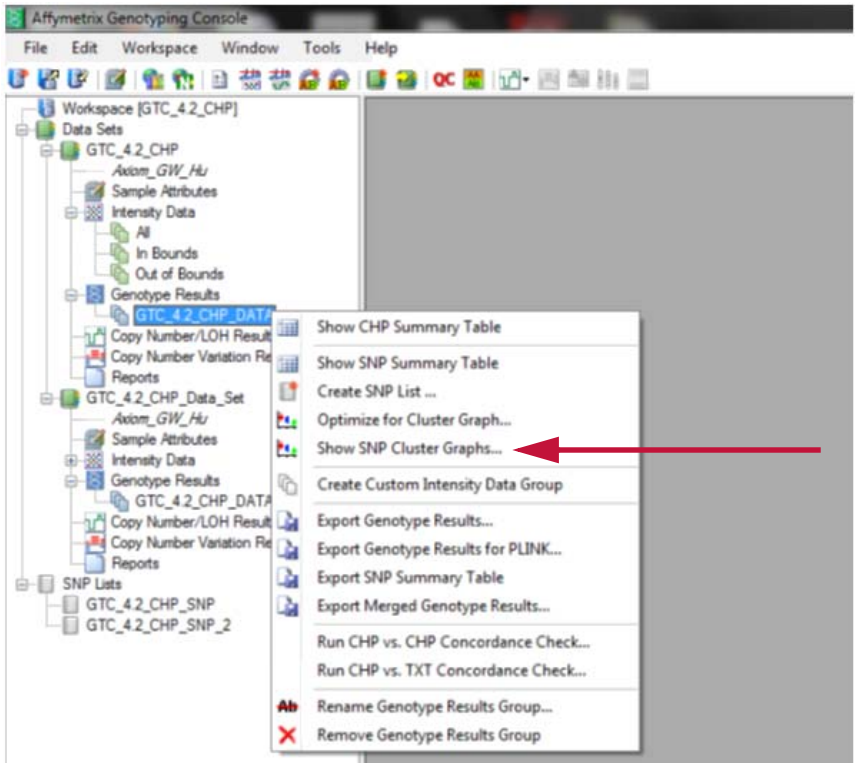
2. The Open dialog box appears. Navigate to the location of the SNP List and select a list. Click **Open** (see Figure 7.11).

Figure 7.11 The Open dialog box.



3. Enter a name for the SNP list and click **OK**. The SNP list will be displayed in the data tree.
4. After a SNP list has been imported, right-click a **Genotype Results** batch and select **Show SNP Cluster Graphs** from the shortcut menu that appears (Figure 7.12).

Figure 7.12 The Shortcut Menu from the Genotype Results

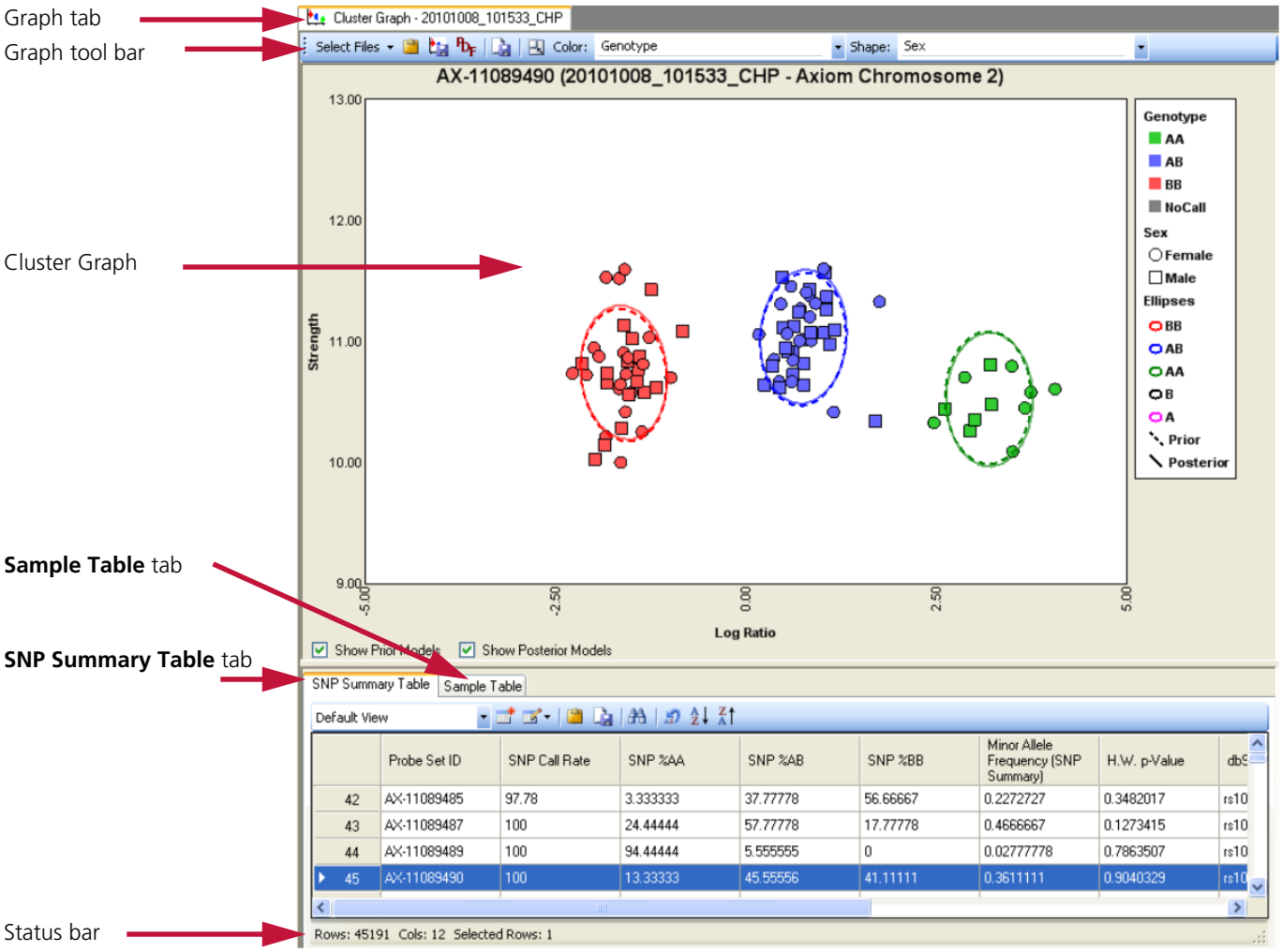


5. If none of the CHP files have matching sample files or if all of the CHP files have matching sample files, no warning appears and the cluster graph is generated. If some of the CHP files are missing matching sample files (ARR), a warning will appear that some CHP files do not have matching sample files. Click **Yes** to produce a cluster plot with a gray spade for samples without the attributes selected from the **Color** or **Shape** drop-down lists.
6. Next, a SNP list dialog box appears. Select a SNP list and click **Okay**.
   If there are no SNPs in common between the SNP list and the array probes, a warning will appear and no cluster plot will be generated. If some but not all of the SNPs on the SNP list are in common with the array probes, a warning notice will appear, stating that some of the SNPs are invalid. A new SNP list containing only the SNPs in common with the array probes will be created in the genotyping results output folder. Import this SNP list to produce SNP cluster plots.

   If the SNP list has no invalid SNPs, the “Select an Annotation File” dialog box opens if an annotation file has not already been selected. If an annotation file is not available on the computer, the user will be prompted to download one.

   The SNP cluster plot is displayed (Figure 7.13).

Figure 7.13 The SNP Cluster Plot Generated by the *GTC SNP Cluster Graph* Function



The *SNP Cluster Graph* function has several drop-down menus will allow the user to change the genotype color, the posterior and prior distribution colors, and the shapes used for the samples. Figure 7.13 shows a cluster plot where the AA cluster is red, the AB cluster is blue, the BB cluster is green, and female samples are plotted as circles while male samples are plotted as squares.

For more information on adjusting color and shape, see the *Genotyping Console™ User Guide 4.2* (P/N 702982), chapter 9, Using the SNP Cluster Graph.

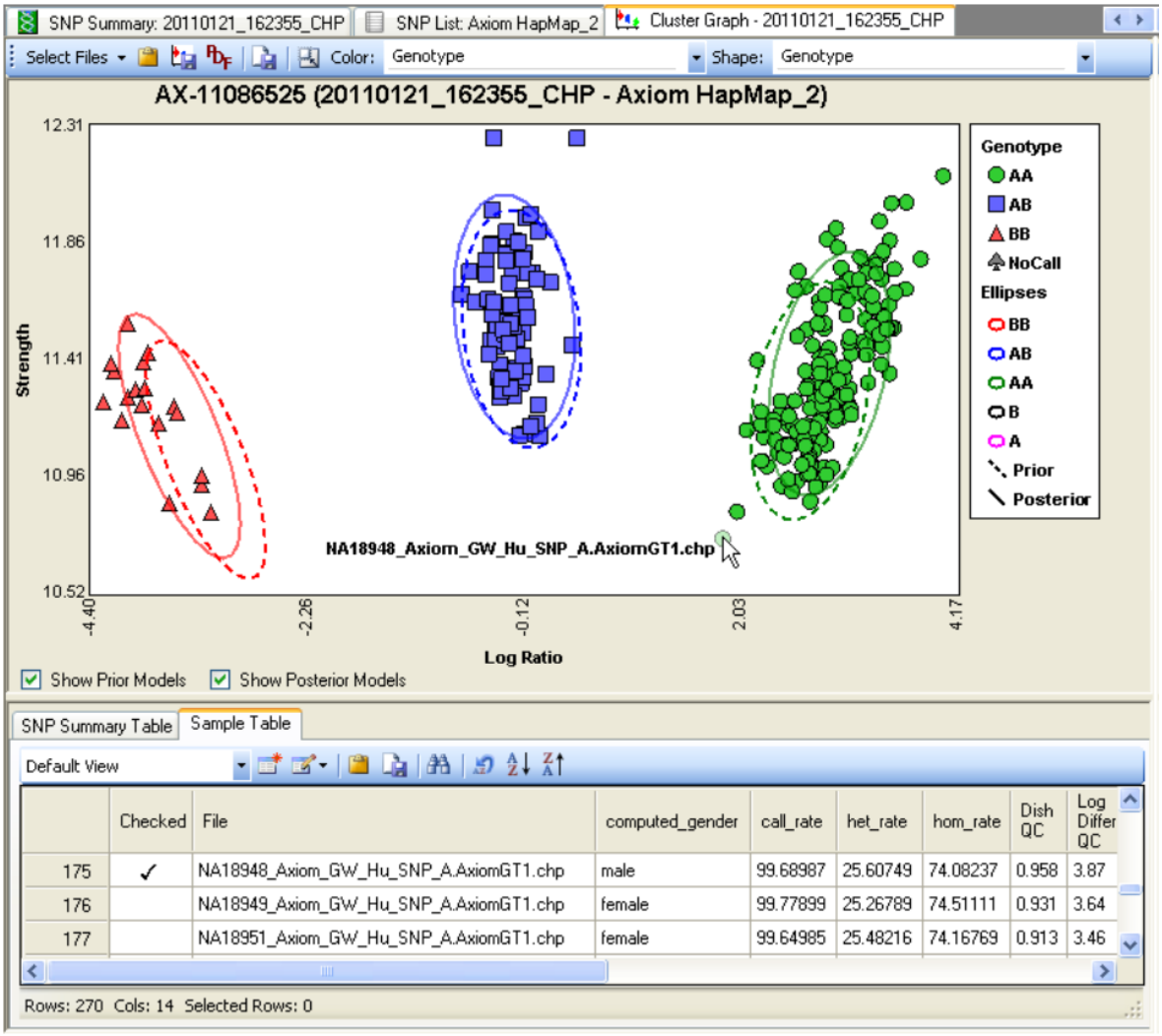
Display a Particular SNP

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 checked in the Sample Table and the **Checked** column in the table will display a check mark (Figure 7.14).

Figure 7.14 Selecting a Single Sample

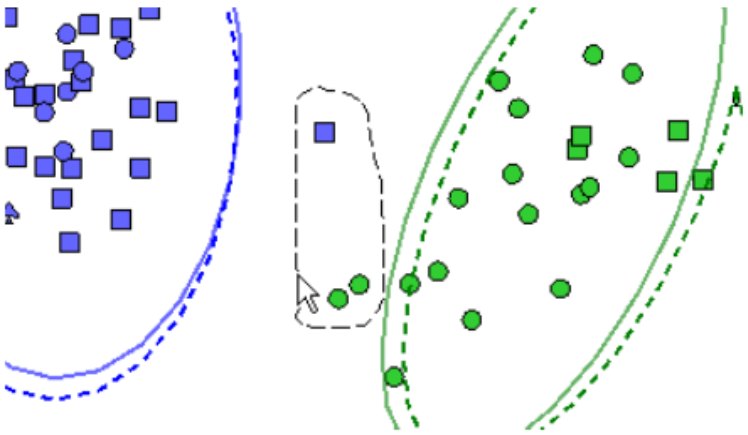


The mouse is hovering over the selected sample. The cluster plot shows the name of the CHP file for the sample. The corresponding row of the Sample Table has a check mark in the **Checked** column for the sample.

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 (see Figure 7.15). 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. If the Sample Table is sorted by the **Checked** column, the selected samples will be moved to the top of the table.

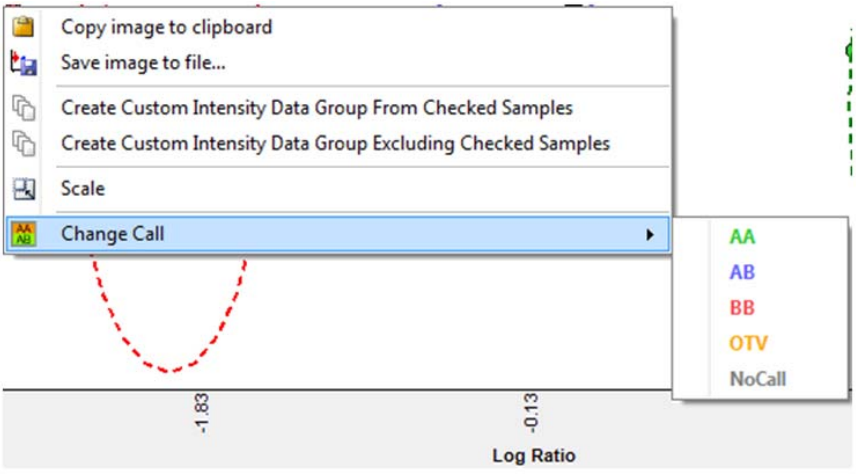
Figure 7.15 Selecting Multiple Samples



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 7.16).

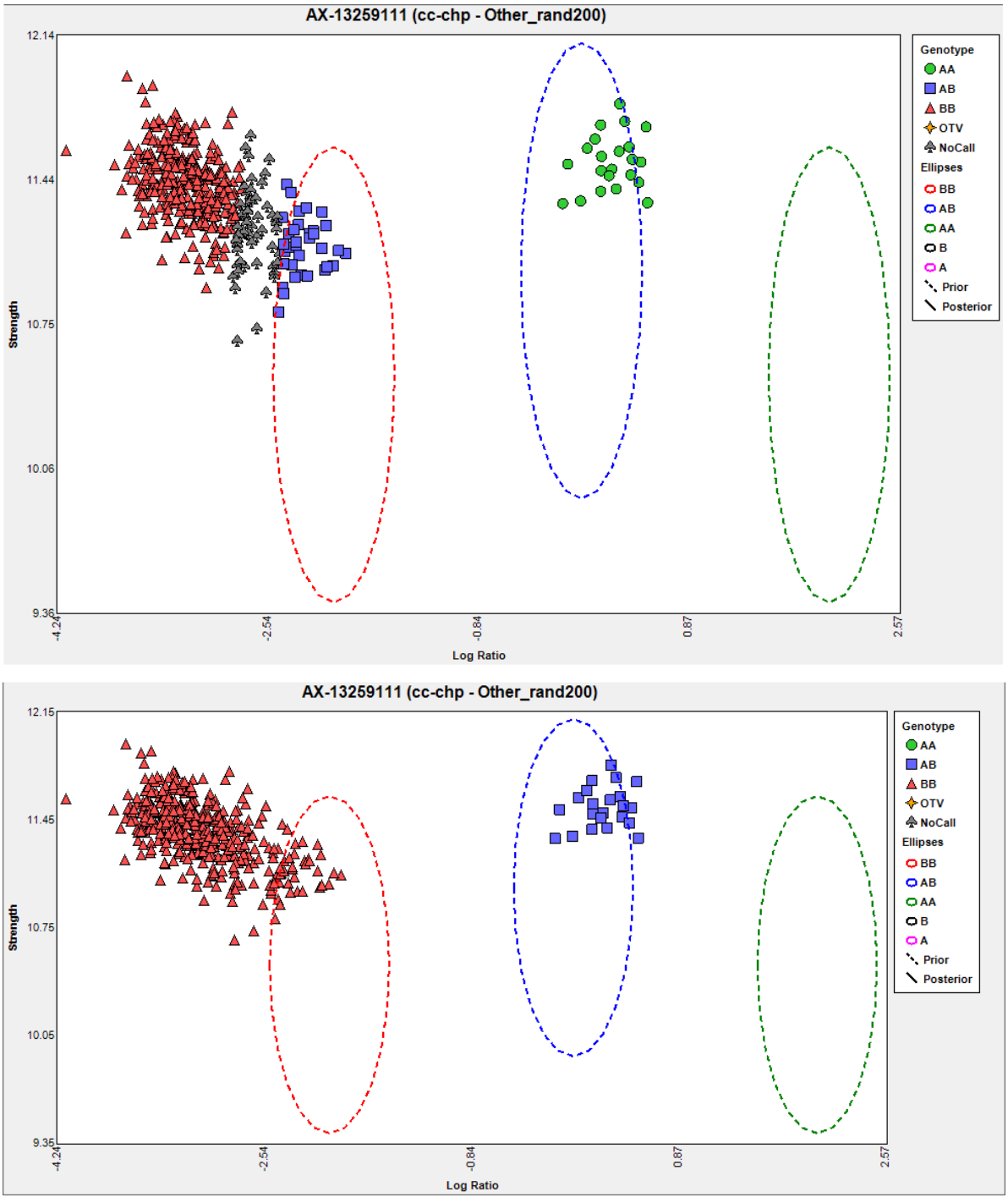
Figure 7.16 Manually Changing a Genotype Call



Lasso Function

The lasso function can be used in a number of different cases including cluster splits. In Figure 7.17 the top half shows a cluster split, and the bottom image shows the graph after setting the samples correctly to AB.

Figure 7.17 Correcting a Cluster Split

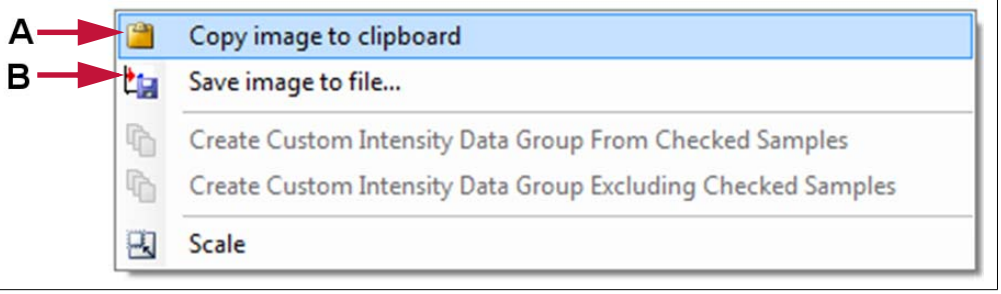


Saving a Cluster Plot

There are several different methods for saving a cluster plot.

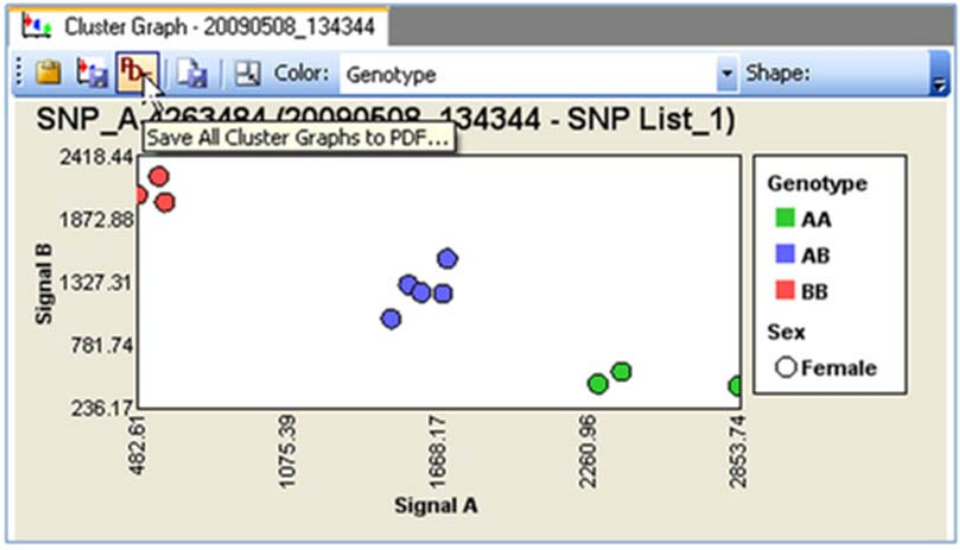
- To copy the cluster plot to the clipboard: right-click the plot and select **Copy image to clipboard** (Figure 7.18-A).
- To save the cluster plot as a PNG file, right-click the plot and select **Save image to file…** (Figure 7.18-B).

Figure 7.18 Saving a Cluster Plot



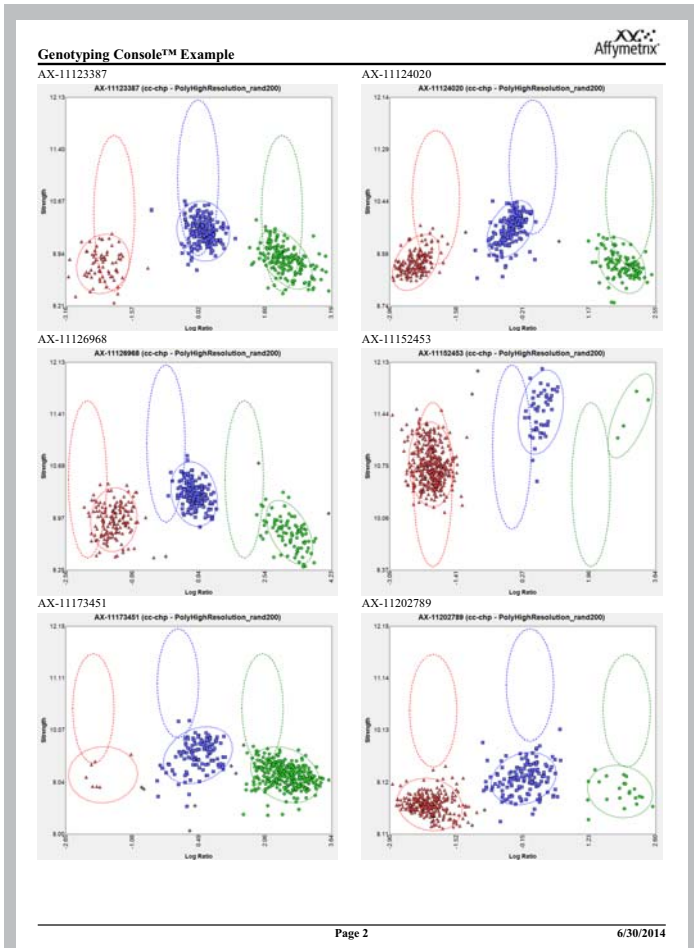
- To save the cluster plots for all SNPs in a SNP List to a single PDF file, click the **Save All Cluster Graphs to PDF** shortcut on the SNP Cluster Graph tool bar (Figure 7.19).

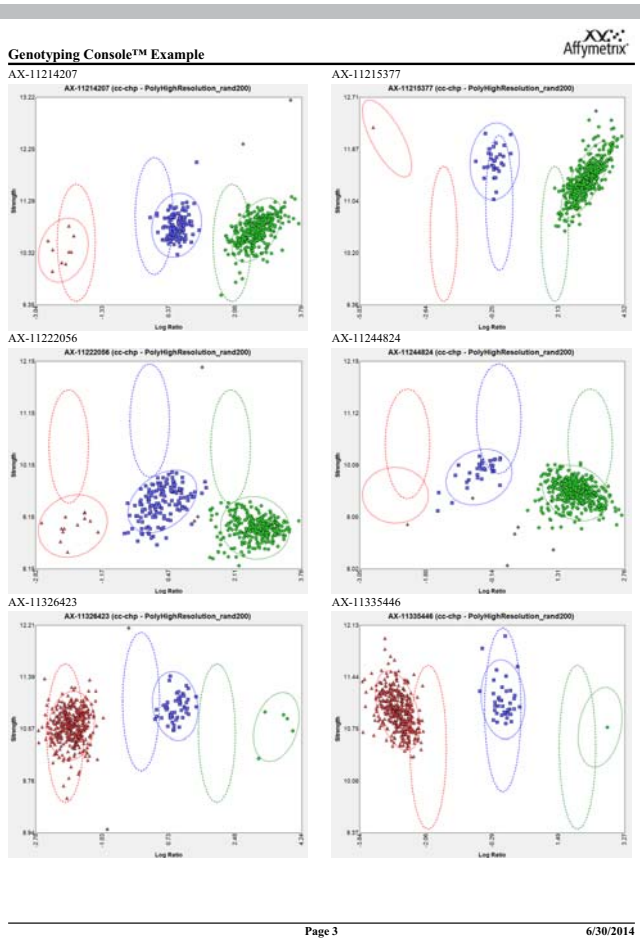
Figure 7.19 The **Save All Cluster Graphs to PDF** Shortcut



The first page of the PDF displays the common legend for all of the cluster plots. The remaining pages display six graphs per page (Figure 7.20).

Figure 7.20 SNP Cluster Plots Saved as a PDF File

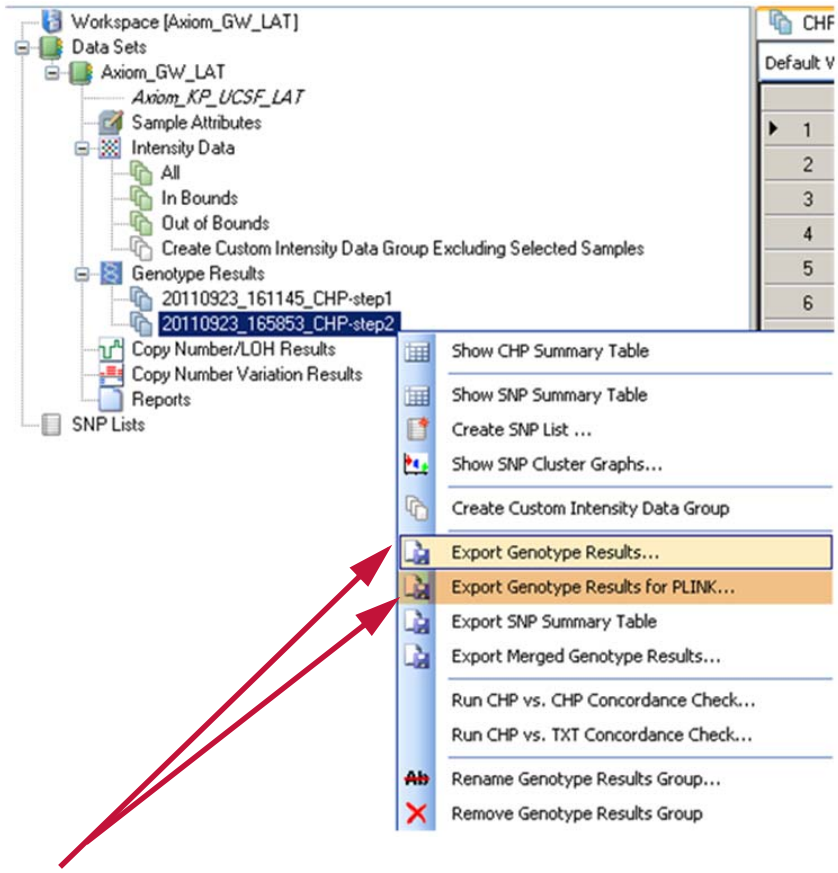




Export Genotypes with GTC Software

The genotype calls for passing samples and recommended SNPs can be exported from GTC for downstream analysis with third-party software. One option is to directly export genotype calls to text files through GTC (Figure 7.21). This export can be limited to a desired SNP list that has already been imported into GTC. Once output text file(s) are generated, downstream analysis can be performed using third-party software. It is also possible to export genotype calls in the PLINK-compatible formats and directly analyze the output files using PLINK software.

Figure 7.21 Exporting genotype results from GTC for third-party analysis.



The **Export Genotype Results…** and **Export Genotype Results for PLINK** commands are found in the pop out menu that appears after right-clicking a genotyping results batch in the data tree.

Instructions for Executing Best Practices Steps with Command Line Software

Execute Best Practice Steps 1-7 with APT Software

In this section, we provide instructions for executing steps 1-8 of the best practice analysis workflow (see Figure 3.1) 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 the Sample DQC values' step), .CEL files corresponding to each batch must be collected into a file (we will refer to the files within each array batch as the 'cel_list') with the full path to each .CEL file in each row and with a header line = "cel_files". We will refer to this list as "cel_list1.txt". Below is a useful Linux one-liner for making cel_lists.

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

Best Practices Step 2: Generate the Sample "DQC" Values 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

```
../bin/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-dir <OUTDIR> \
  --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 (GTC 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-probe set-genotype. 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-probe set-genotype script for step 4 of the best practice analysis workflow using APT

```
../bin/apt-probe set-genotype \
--log-file  <OUTDIR>/apt-probe set-genotype.log \
--xml-file  <ANALYSIS_FILES_DIR>/<axiom_array>_96orMore_Step1.r<#>.apt-probe set-genotype.AxiomGT1.xml \
--analysis-files-path <ANALYSIS_FILES_DIR> \
--out-dir <OUTDIR> \
--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-probe set-genotype.AxiomGT1.xml to perform QC genotyping with *SNP specific models* if batch size is less than 96 samples.

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>/ AxiomGT1.report.txt" produced by step 4. When executing the workflow with APT (GTC automates this step), the user must write a script to remove .CELs from the <OUTDIR>/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 = $\frac{\text{Samples passing DQC and 97\% call rate}}{\text{Total samples on the plate}}$ X 100
  - Compute the average 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 call rate of the remaining samples for the given plate
- Guidelines for passing plates in:
  - average 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*

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-probe set-genotype. 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-probe set-genotype script for step 7

```
../bin/apt-probe set-genotype \
--log-file  <OUTDIR>/apt-probe set-genotype.log \
--xml-file <ANALYSIS_FILES_DIR>/<axiom_array>_96orMore_Step2.r<#>.apt-probe set-genotype.AxiomGT1.xml  \
--analysis-files-path <ANALYSIS_FILES_DIR> \
--out-dir <OUTDIR> \
--summaries \
 --write-models \
 --cc-chp-output \
--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-probe set-genotype.AxiomGT1.xml

whereas the example script for step 4 executes:

<axiom_array>_96orMore_**Step1**.r<#>.apt-probe set-genotype.AxiomGT1.xml

The step 7 genotyping script includes options to write out a number of files to <OUTDIR>. 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 10) 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.
- CHP files - one for each sample - are enabled by the *-cc-chp-output* option.
  CHP files can be input into GTC and thus enable analyses (such as viewing of cluster plots and generation of files) in PLINK format.

Note: Choose <axiom_array>**_LessThan96_Step2.r**<#>.apt-probe set-genotype.AxiomGT1.xml to perform genotyping with SNP-specific models if batch size is less than 96 samples.

Execute Best Practice Step 8 with SNPolisher Functions

SNPolisher Setup

The R package files to install SNPolisher are available on the Affymetrix website (www.affymetrix.com). Select **Register** at the top of the website to register your email address with Affymetrix. Once you have registered, SNPolisher can be downloaded from the "DevNet Tools" page. From the **Partners and Programs** menu, select **Developers' Network**. Log in using your email address and click **DevNet Tools**. SNPolisher is available under the **Analysis Tools** menu. 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, and readme files; a PDF with colors for use in R; and the example R code and 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. Follow the instructions in the SNPolisher User Guide for installation instructions for R, Perl, and SNPolisher. This section assumes some familiarity with the R programming language. Please see the SNPolisher User Guide for instruction on R basics, or contact your local FAS for support, or send email to Support@affymetrix.com.

NOTE: Step 8 also be can be executed using APT version 1.16.1 or higher, using the *Ps_Metrics* and *Ps_Classification* software.

Best Practices Step 8A: Run *Ps_Metrics*

After perl and R have been installed and the SNPolisher library has been loaded (see SNPolisher User Guide), the *Ps_Metrics* function can be run. *Ps_Metrics* uses two output files from Best Practices Step 7 (AxiomGT1.Step2 above) as inputs: *AxiomGT1.posteriors.txt* and *AxiomGT1.calls.txt*. The user should specify the name and location of the AxiomGT1.posteriors.txt and AxiomGT1.calls.txt file, and the desired name and location for the new metrics output file. 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 longer example of running the SNPolisher functions.

The input files required for *Ps_Metrics* should either be located in the working directory or the user must provide the file path when typing the arguments. If the file path is wrong or no file path is given and the files are not in the working directory, R will return an error of "file not found". R working directories are discussed in the SNPolisher User Guide.

If a user has set the R working directory to the desired output folder location and the Step2_AxiomGT1 (Step 7 Best Practices) files are in the <OUTDIR> folder, then the *Ps_Metrics* command in R running on Windows would be:

```
> Ps_Metrics(posteriorFile=<OUTDIR>/AxiomGT1.snp-posteriors.txt",
                                callFile=<OUTDIR>/AxiomGT1.calls.txt",
                                output.metricsFile="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 8.1. 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 8.1 An example of the output file from *Ps_Metrics*.

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

The first 13 rows of the output file from *Ps_Metrics* ("metrics.txt"), opened with Excel.

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 1.1 on page 7). If this file has not been provided, users should contact their local Affymetrix Field Application Support or send email to Support@affymetrix.com.

Below is an R command example for *Ps_Classification* which:

1. uses output from *Ps_Metrics* metrics results in metric.txt,
2. the classification results should be stored in the folder called *Output*,
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(metricsFile="metrics.txt",
                    output.dir="Output",
                    SpeciesType="Human",
                    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 3.1 on page 21.). SNPs classified as *PolyHighRes* must have SNP QC values that pass all of the thresholds.

There are two logical indicators: *HomRO.flag* indicates if HomRO thresholds should be used (default is TRUE), and *HomHet.flag* indicates if the HomHet metric should be used (default is TRUE). Polyploid genotypes do not use either of the HomRO thresholds so *HomRO.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 *output.converted* and *priority.order*. *output.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.

GTC is a flag indicating if the category files (*PolyHighResolution.ps, NoMinorHom.ps, Hemizygous.ps, MonoHighResolution.ps, CallRateBelowThreshold.ps, Other.ps*, and *OTV.ps*) will be used with the GTC SNP Cluster Graph function. The default = FALSE. In order to visualize probe sets listed in the category files with the GTC SNP Cluster Graph function, these files must be imported into GTC as a SNP list. The GTC SNP list file must be a text file with a .txt or .tsv extension and contain a column labeled “Probe Set ID”. If the GTC flag in SNPolisher *Ps_Classification* was set to FALSE, the column header must be updated by hand. Open the category file in a text editor such as Notepad or Word and change *probeset_id* to Probe Set ID then save with a *.txt* extension. If the GTC flag in *Ps_Classification* was set to TRUE, the category files still must be renamed with a *.txt* or *.tsv* extension.

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 longer examples of running the SNPolisher functions and for more details of the arguments for *Ps_Classification*.

Note that if SNPolisher v 1.4 or less is used, the example command will be

```
> Ps_Classification(metrics.file =“metrics.txt”,
                    output.dir =“Output”,
                    SpeciesType =“Human”,
                    ps2snp.file = <ANALYSIS_FILES_DIR>/Axiom_BioBank1.r2.ps2snp_map.ps)
```

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. For more details on the various options with *Ps_Visualization*, see the SNPolisher User Guide.

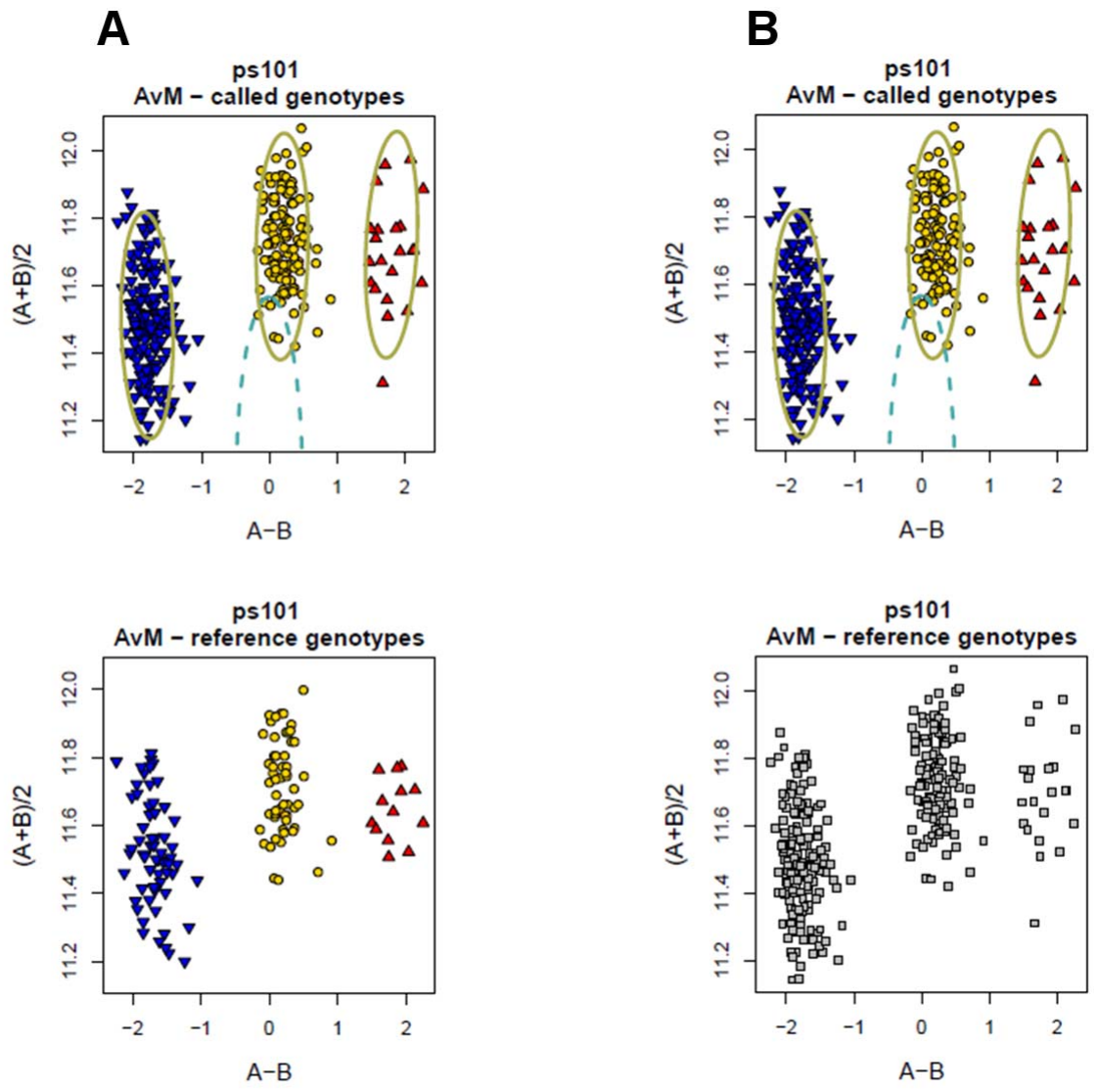
Ps_Visualization takes six required arguments and eight optional arguments. The six required arguments are the name of the ps file with the SNPs for plotting, the name and location of the output PDF file, the name and location of the summary file, the name and location of the calls file, the name and location of the confidences file, and the name and location of the posteriors file.

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

The user should also give *Ps_Visualization* the name of a temporary directory which will be used for outputting intermediate files (*temp.dir*). If no directory is given, the default used is “Temp/”. The accompanying logical operator *keep.temp.dir* indicates whether the temporary directory should be kept or deleted at the end of *Ps_Visualization* (default is FALSE). The user may wish to keep this temporary directory if the intermediate output files are needed for closer inspection.

Ps_Visualization has eight optional arguments. *sampFile* takes the name of a file containing a list of samples to be highlighted in a plot. This file has no header line, and the CEL or sample names must match the names in the summary and calls files. *refFile* takes the name of a file containing reference genotypes for plotting (default is NULL). The source of reference genotypes may be a SNP discovery project such as HapMap or the 1000 genomes project, or it may be genotypes produced in an NGS project for the user's samples. The user must create the *refFile* with the same format and genotype codes as the AxiomGT1.calls.txt file. Samples without reference genotypes have their reference plotted as a gray square for "No call". If there are very few reference genotypes, the reference plot will contain mostly gray squares. In this case, the user may wish to consider plotting only those samples with known reference genotypes (Figure 8.2).

Figure 8.2 One SNP Plotted With and Without Reference Genotype



A: A SNP plotted with reference genotypes. B: The same SNP plotted without reference genotypes.

priorFile takes the name of a file containing the prior information for the genotypes (default is NULL). This file must be in the same format as the posterior information file. The accompanying logical operator *plot.prior* indicates if the prior genotype cluster centers are plotted (default is FALSE). If *plot.prior* is FALSE, *Ps_Visualization* ignores *priorFile*. If *plot.prior* is TRUE and *priorFile* is NULL, then *Ps_Visualization* plots a generic prior.

match.cel.file.name is a logical operator indicating if sample file names in the calls file are checked against those in the confidence summary files. Input files may not have been checked against each other, so the default is TRUE. *max.num.SNP.draw* is the maximum number of SNPs that should be plotted. If the list of probe set IDs used is one of the categorical lists output by *Ps_Classification*, it is important to set *max.num.SNP.draw* (we suggest < 500) or all SNPs in the list will be plotted. *nclus* is the number of genotype clusters. The default value is 3, and needs to be set of 5 for auto-tetraploids.

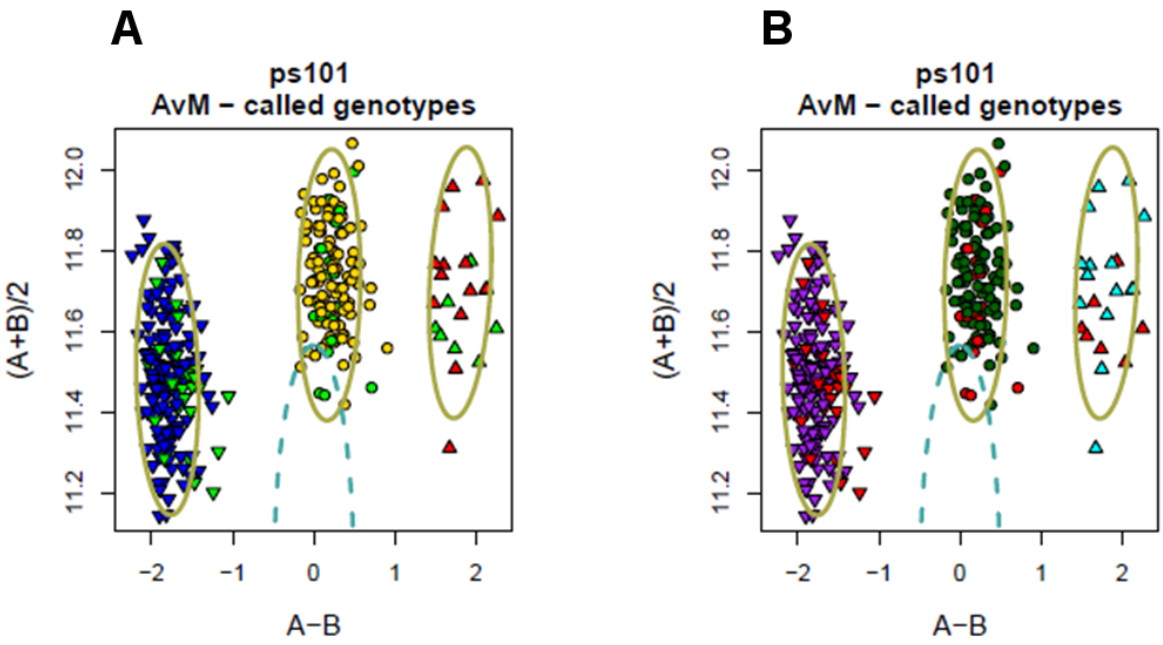
geno.col is the list of colors for plotting the genotypes. The input for *geno.col* must be given as a vector of colors. To make a vector, use the command c (for concatenate):

> c("red","yellow","orange","green","blue","purple")

[1] "red" "yellow" "orange" "green" "blue" "purple"

The user cannot change the colors of the posterior and prior ovals (yellow and light blue). The default values for *geno.col* are "red","gold","blue","gray","cyan","green","darkgreen","purple". These colors correspond to the AA cluster, the AB cluster, the BB cluster, null calls, the OTV cluster, highlighted samples, a fourth cluster (optional), and a fifth cluster (optional). See Figure 8.3 for one SNP plotted with different colors.

Figure 8.3 One SNP Plotted With Color and Sample Highlighting Options



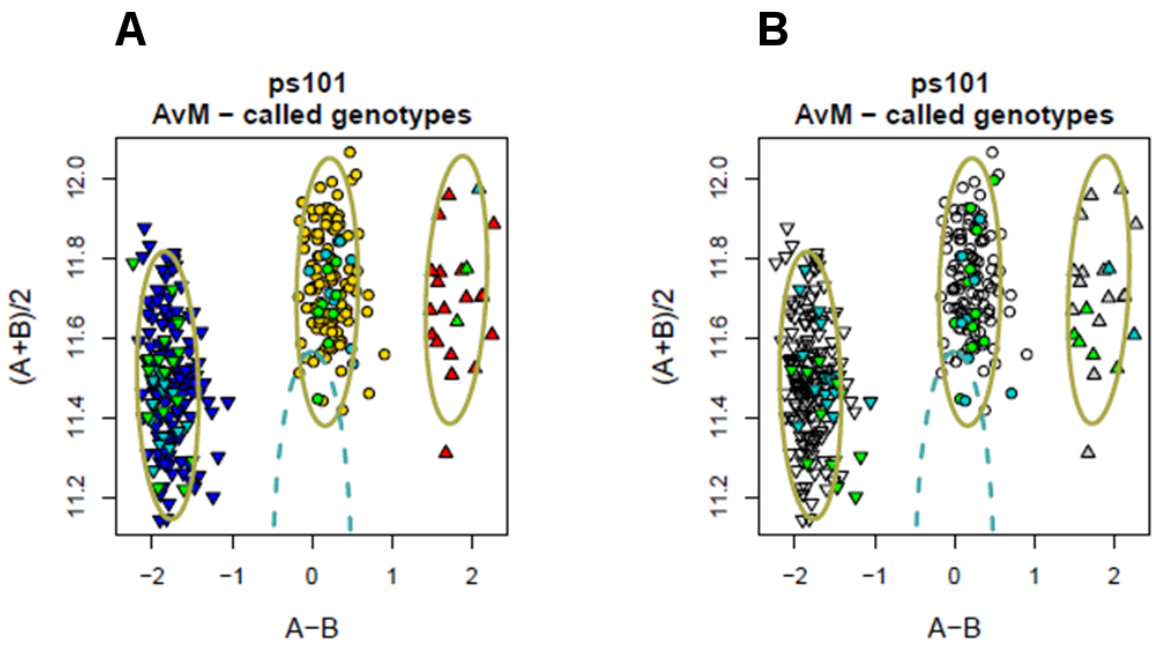
A: One SNP plotted with default color options and highlighted samples.

B: The same SNP plotted with different colors and the same highlighted samples.

The encoding for colors in R is quite complicated. For more information on selecting colors in R, see the SNPolisher User Guide.

In addition to setting colors through the *geno.col* argument, highlighted samples can be set to multiple colors using the text file given for *sampFile* (see Figure 8.4-A). In this case, the text file is a two-column tab-deliminated file. The first column is the sample name and the second column is the desired color. The first line should read "sample color", separated by a tab. To keep the highlighted samples colored and change the color of all other samples to a single color, set *geno.col* to be only one color: *geno.col=c("lightgray")* (see Figure 8.4-B). When using multiple colors for highlighted samples, be sure to check that the colors are clearly visible against the colors used for other samples.

Figure 8.4 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.

If the user in the previous examples wishes to produce cluster plots after running *OTV_Caller*, the command in R should be:

> Ps_Visualization(pidFile="PolyHighResolution.ps",
output.pdfFile="Cluster_PolyHighResolution.pdf",
summaryFile="C:\data\AxiomGT1.summary.txt",
callFile="C:\data\AxiomGT1.calls.txt", confidenceFile="C:\data\AxiomGT1.confidences.txt",
posteriorFile="C:\data\AxiomGT1.snp-posteriors.txt", temp.dir="Temp/", keep.temp.dir=FALSE,
refFile=NULL,plot.prior=T, priorFile=NULL, atch.cel.file.name=TRUE,max.num.SNP.draw=6,
geno.col=c("red","gold","blue","gray","cyan","green","darkgreen","purple"),nclus=3)

In this example, the working directory contains the output from *Ps_Classification*, including the PolyHighResolution SNP list. The output PDF file will be named "Cluster PolyHighResolution.pdf" and will be made in the working directory. There is no reference genotype file. The prior distributions will be plotted but no prior information file is given, so it will be a generic prior. The maximum number of SNPs plotted will be 6. The genotype colors given are the default colors, and there are three clusters per SNP.

See SNPolisher User Guide for a longer example of running the SNPolisher functions and for more details of *Ps_Visualization*.

References

Affymetrix® Genotyping Console™ User Manual:

- Genotyping Console™ 4.2, supported on Windows 7 (64-bit) and Windows 8.1: http://media.affymetrix.com/support/downloads/manuals/gtc_4_2_user_manual.pdf
- Genotyping Console™ 4.1, supported on 32-bit and/or Windows XP systems: http://media.affymetrix.com/support/downloads/manuals/gtc_4_1_user_manual.pdf

Affymetrix Power Tools Manual:
Manual: apt-probeset-genotype (1.16.1):
http://media.affymetrix.com/support/developer/powertools/changelog/apt-probeset-genotype.html

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

Baker M. Genomics: The search for association. *Nature*. 2010 Oct 28;**467**(7319):1135-8.

Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet*. 2003 Feb 15;**361**(9357):598-604.

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

de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*. 2008 Oct 15;**17**(R2):R122-8.

Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, de Villena FP, Churchill GA. Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. *BMC Genomics*. 2012 Jan 19;**13**:34.

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

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

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

Voorrips RE, Gort G, Vosman B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*. 2011 May 19;**12**:172.

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