Axiom™ Copy Number Data Analysis

ANALYSIS GUIDE

Publication Number MAN0026736
Revision 3

ThermoFisher SCIENTIFIC

Affymetrix, Inc.
3450 Central Expressway
Santa Clara, CA 95051 USA

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

DISCLAIMER

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

Revision history: Pub No. MAN0026736

Revision	Date	Description
3	25 May 2022	Initial release in Thermo Fisher Scientific document control system. Supersedes legacy Affymetrix publication number 703518. Updated to include new copy number reference creation workflows. This guide is now in a Chapter-based format.
2	05 August 2020	Minor updates to script commands. Added more information about Hypervariable Regions.
1	15 January 2020	Initial release.

Important Software Licensing Information

Your installation and/or use of this Axiom Analysis Suite software is subject to the terms and conditions contained in the End User License Agreement (EULA) which is incorporated within the Axiom Analysis Suite software, and you will be bound by the EULA terms and conditions if you install and/or use the software.

NOTICE TO PURCHASER: DISCLAIMER OF LICENSE

Purchase of this software product alone does not imply any license under any process, instrument or other apparatus, system, composition, reagent or kit rights under patent claims owned or otherwise controlled by Life Technologies Corporation, either expressly, or by estoppel.

Legal entity

Affymetrix, Inc.| Santa Clara, CA 95051 USA | Toll Free in USA 1 800 955 6288

For support visit **thermofisher.com/support** or email **techsupport@lifetech.com**

TRADEMARKS

Contents

Contents

1 Introduction

Overview

This Guide provides information and instructions for performing copy number variation (CNV) data analysis on Applied Biosystems Axiom arrays. For an introduction to CNV analysis using Axiom, please see Copy Number Variation Analysis Using Applied Biosystems Axiom Arrays https://assets.thermofisher.com/TFS-Assets/GSD/Technical-Notes/tech-note-axiom-cnv.pdf.

This Guide should be used along with the Axiom Genotyping Solution Data Analysis User Guide https://assets.thermofisher.com/TFS-Assets/LSG/manuals/axiom_genotyping_solution_analysis_guide.pdf which provides a more detailed description of Axiom data analysis.

Two software systems are used for Axiom copy number analysis and described in this document:

1. Axiom Analysis Suite version v5.2 (AxAS) https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html and later
2. Analysis Power Tools version v2.11.6 (APT) https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html and later.

Copy number analysis may be performed only on Axiom arrays enabled for such analysis with the appropriate library file packages. Analysis is currently restricted to diploid species. For arrays with copy number enabled library files, users are in general able to:

- Perform CNV analysis using Fixed Region and Discovery workflows.
- Visualize results in AxAS using genome tracks in the interactive Whole Genome View, CN Region Plots and linked tables.
- Create custom CN reference file (cn_models) from template file (cn_models_template) provided in the library file package.

For arrays without copy number enabled library files, users will not be able to do copy number analysis using AxAS or APT. They may still be able to perform limited copy number analysis using the stand-alone Windows-based Axiom CNV Summary Tool (https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-cnv-summary-tools-software.html)

For help or questions, contact your local Thermo Fisher Scientific Field Application Scientist or thermofisher.com/support (https://www.thermofisher.com/us/en/home/technical-resources.html?CID=fl-support).

Analysis software

Axiom Analysis Suite (AxAS) is a software package that integrates all the tools necessary to perform genotyping and copy number analysis into one easy-to-use graphical interface. The software is designed to allow users to set the desired settings and to process through all steps with minimal interaction required. Axiom Analysis Suite is the recommended software system for most Axiom users.

Analysis Power Tools (APT) is a set of cross-platform command line programs that implement algorithms for analyzing and working with data from Axiom arrays. APT programs are for users who prefer programs that can be used in scripting environments and are advanced enough to handle the complexity of extra features and functionality.

Library files for CNV analysis

The following files must be present within the library folder (Analysis Directory) for a copy number enabled array:

```
<array name>.r<>.cn_models_template
<array name>.r<>.cn_models
<array name>.r<>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml
```

Arrays that support the Initial and Best Practices Copy Number Reference Creation workflows also require:

```
<array name>.r<>.AxiomCNVmixUtil.apt2.xml
<array name>.r<>.signatureSNPs.refs.txt
<array name>.r<>.plate_map.txt
<array name>.r<>.cn_region_refcalls
<array name>.r<>.cn_region_refsettings
```

Arrays enabled for Fixed Region analysis will also have the following files:

```
<array name>.r<>.generic.cn_priors
<array name>.r<>.apt-copynumber-axiom-
cnvmix.AxiomCNVmix.apt2.xml
<array name>.r<>.apt-genotype-axiom.AxiomCN_GT1.apt2.xml or
<array name>.r<>.apt-genotype-axiom.AxiomCN_PS1.apt2.xml
```

Arrays enabled for Discovery analysis will also have the following files:

```
<array name>.r<>. hmm_regions.txt
<array name>.r<>.apt-copynumber-axiom-hmm.AxiomHMM.apt2.xml
<array name>.r<>.apt-genotype-axiom.AxiomCN_GT1.apt2.xml
```

Workflows run using Axiom Analysis Suite also require:

```
<array name>.r<>.analysis_settings
<array name>.r<>.v5.ax_thresholds
```

Definitions for CNV analysis

Copy number variation and use

CNVs on test samples are detected using computed log2 ratios and B allele frequencies (BAF) for probesets interrogating individual markers across the genome.

Probeset

A probeset is a group of one or more probe sequences that interrogates a specific known polymorphic or nonpolymorphic location in the genome.

Log2 ratios

The log2 ratio is the log transformed ratio of the test sample signal intensity relative to a reference total intensity for the same probeset. Log2 ratios are negative for losses and positive for gains. In a normal diploid region, median log2 ratios for a one-copy loss may vary from about -0.3 to -0.6, and median log2 ratios for a one-copy gain range vary from about 0.2 to 0.35. Differences in log2ratios between consecutive CN states become smaller with higher CN states, such as CN4 and CN5. Empirical log2 ratios based on measured intensities become more compressed and show smaller differences between consecutive CN states. For this reason, it can be difficult to distinguish between high consecutive CN gain states.

Copy number reference

A copy number reference is a set of reference intensity values probesets generated from a set of normal samples run on the array. The reference total intensity is an estimate of the total A and B allele intensities for the probeset representing the normal CN state at that location (usually copy number 2 for autosomal regions in diploid organisms). This reference value is calculated as the median total intensity for that probeset across reference samples. If most of the individual samples in a dataset are expected to have the normal CN State, the reference may be created using all samples in the dataset. Generating a reference in genomic regions where CNVs are common poses a special challenge. For example, in the human GSTM1 gene, most individuals in a population may not be diploid. In such a region, the reference for probesets must be generated from carefully selected diploid samples.

B allele frequency (BAF)

BAF is a normalized measure of the allelic intensity ratio of two alleles that indicates the level of heterozygosity in a genomic region. It can be used to identify mosaic chromosomal aneuploidies and loss or absence of heterozygosity (LOH). BAF track in the normal CN2 region will have three bands. A region with increased copy number shows more bands while loss of heterozygosity results in two bands.

Figure 1 BAF and log2 ratio tracks from AxAS WGV

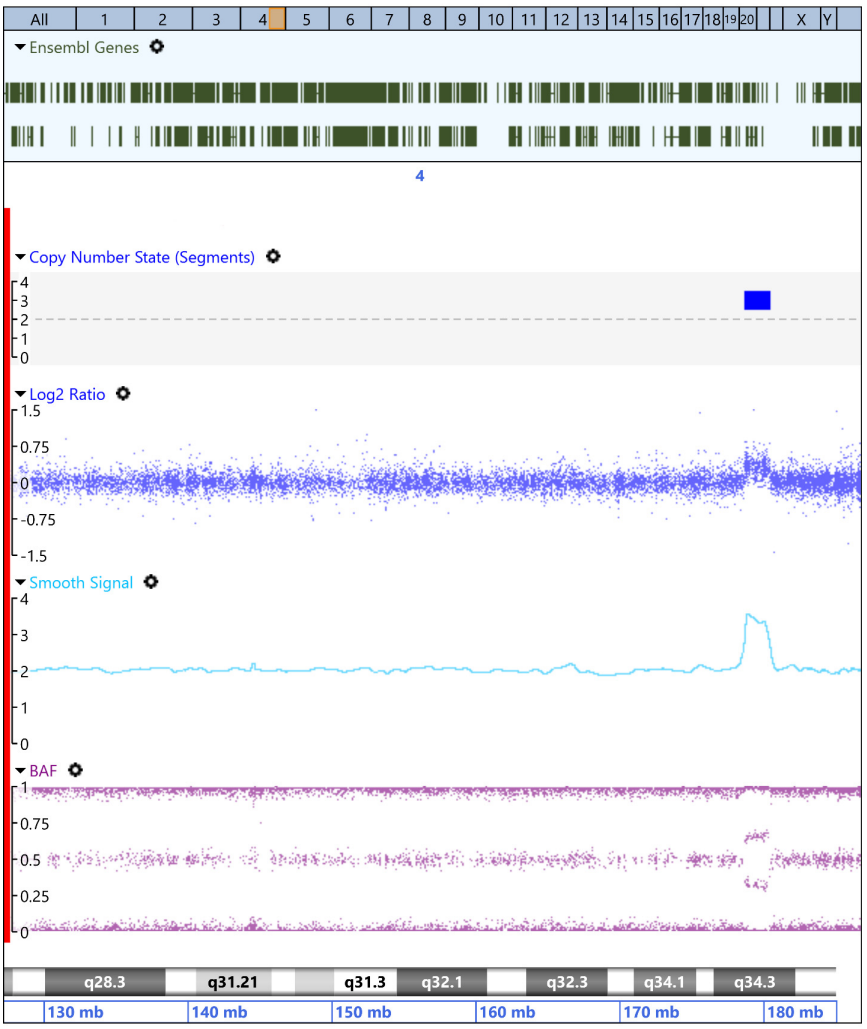


Figure 1 is an example of a single copy gain on chr4:177,197,000-178,966,00 in a sample assayed using the Axiom Precision Medicine Diversity Research Array as visualized using Whole Genome View in Axiom Analysis Suite 5.2. Points represent values measured by probesets. Log2 ratio values in the normal diploid region hover around 0.0, and show elevation from normal state in the duplication region. BAF tracks show three bands in the normal diploid region corresponding to the AA, AB and BB alleles, and four tracks in the duplication representing AAA, AAB, ABB and BBB alleles.

Hypervariable regions

On Axiom Human Genotyping arrays, there are genomic regions that show copy number changes across many samples. These are usually regions of common polymorphisms and are determined empirically. Some examples of hypervariable regions are 14q32.33, 17q21.31, and 22q11.22. Large scale analysis of copy number results may lead to the discovery of additional such regions. Copy number calls may be less accurate in these regions as it is more difficult to create a diploid reference in these regions.

Sample QC for CNV analysis

Samples that fail genotyping QC should be removed before performing copy number analysis (*Chapters 3 and 5 of the Axiom Genotyping Solution Data Analysis Guide*). Additionally, there are two copy number QC metrics used to identify poor quality samples for copy number analysis: MAPD and WavinessSD.

Median of the Absolute values of Pairwise Differences (MAPD)

MAPD is a global measure of variation in log2ratio measured across the genome. It represents the median of the distribution of differences in log2 ratio between adjacent markers. MAPD is a measure of short-range noise in the data. Higher values indicate more noise in the data, possibly due to poor quality of the DNA material or issues in sample processing. MAPD is dependent on the reference used because it is calculated from log2 ratios. A drift in data away from the reference can cause MAPD values to increase. High MAPD, especially when systemically observed, can be an indication that the reference is not appropriate (i.e. the sample data are very different from the data used to create the reference). The threshold for MAPD is typically set at 0.35. Samples with higher values should be excluded from copy number analysis. Examples of log2ratio tracks for different MAPD values are shown in Figure 2.

Figure 2 Examples of log2ratio tracks for different MAPD values

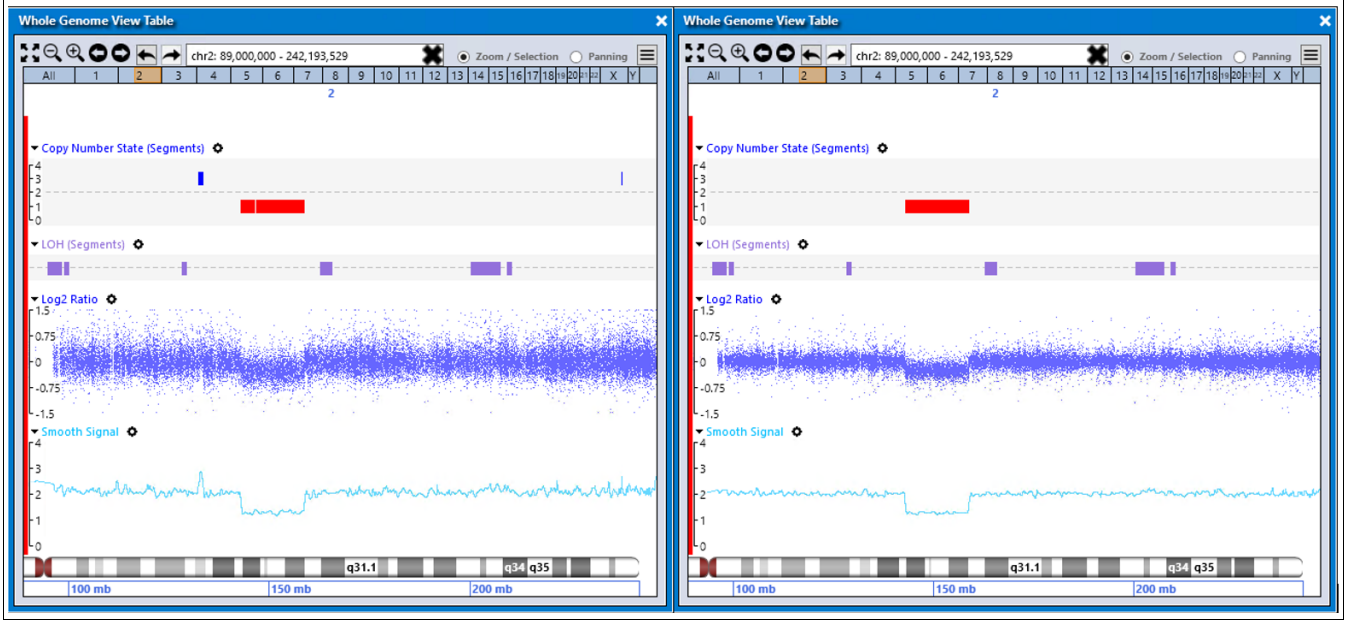


Figure 2: Log2Ratio track (in blue) for the same sample using two different copy number references. Smooth Signal (in cyan) is based on Log2Ratio. The MAPD value for the left plot is 0.27, and for the right plot is 0.15. True CN changes are easier to detect with lower MAPD, and the number of false CN calls is also reduced. This figure also demonstrates the value of an appropriate copy number reference.

WavinessSD

WavinessSD is a global measure of variation of log2 ratio that is insensitive to short range variation and focuses on long-range variation. High WavinessSD usually indicates too much noise in the data and implies either sample or processing batch effects that will reduce the quality of the copy number calls. However, elevated

WavinessSD with good MAPD can also occur in samples with many copy number changes or very large regions of change (for example, cancer samples). In such situations, it is important to inspect the data more carefully. The threshold for WavinessSD is typically set at 0.1. Samples with higher values should be excluded from copy number analysis.

Plate-corrected QC metrics

Copy number QC metrics for arrays run with internal plate controls are calculated using plate-corrected log2ratios. They are called MAPDc and WavinessSDc in the analysis software. Thresholds for these metrics are different and set in the analysis files.

Steps in CNV analysis

Before starting CNV data analysis, users should complete assessment of site readiness and ensure good study design as described in Chapter 2.

The key to a successful CNV analysis is a good copy number reference. A reference file (cn_models) is required for CNV analysis on Axiom arrays. If the reference file is present in the library folder (Analysis Directory), it may be used for CNV analysis. However, the reference file is not generated at array design time and may not be present in the library folder of some copy number enabled custom arrays. If the reference file is not present or if it needs to be updated, follow the steps in Chapter 3 to create a reference using Axiom Analysis Suite. APT users may use the command line options and manual steps described in Chapter 4. The required steps include selection of normal samples which is described in Chapter 5.

The analysis software offers two CNV analysis methods:

1. Fixed Region analysis when breakpoints of CNV regions of interest are known a priori and there is little breakpoint variability from sample to sample (Chapter 6).
2. Discovery analysis for screening the whole genome for CNVs (Chapter 7).

Copy number aware genotyping

On some arrays such as the Axiom Precision Medicine Diversity Research Array, copy number analysis is used to guide genotyping of probesets in variable copy number regions. These probesets use the Copy Number Aware Genotyping (CNAG) algorithm which reports double deletion, haploid and diploid calls. For more details, see the *Axiom Genotyping Solution Data Analysis User Guide* (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/axiom_genotyping_solution_analysis_guide.pdf).

Genome build and genomic locations

Copy number analysis requires the genomic location of markers corresponding to probesets. Therefore, the genome build used for marker locations is important and mixing genome builds is likely to cause errors. Copy number states are determined, not at any single marker location, but over a genomic region often specified by chromosome:start-end and covered by several probesets. In Discovery analysis, the exact location of probesets is important as the algorithm determines changes in copy number state from one location to the next in genomic order. In Fixed Region analysis, probesets must be assigned to copy number regions *a priori*. Genomic location is usually required for assigning probesets to copy number regions.

CNV reference generation

Reference files

The library folders of all copy number enabled arrays will have a reference template file (cn_models_template) for creating a reference file. The cn_models_template file is generated at array design time for all copy number enabled arrays. It is an HDF5 file with multiple datasets including ProbeData, Probe-covariates and VCFAnnotations. These datasets have the list of probesets used in copy number analysis along with associated covariates used to calculate adjustments to intensities and log2ratios. They also show fixed regions, if there are any, and process flags for each probeset that inform the software whether the probeset is to be used for calculating covariates, QC metrics, loss of heterozygosity, and Discovery analysis. The ReportedPS and RemotePS datasets are used for genotyping algorithms based on copy number. Not all datasets are present in all reference template files.

When a reference file (cn_models) is generated, the Reference dataset gets populated with the computed reference median signal values and mean BAFs for all probesets. The Header dataset is added. The Probeset-covariates and WaveCorrection datasets have additional probeset specific correction values that are computed from the data to account for batch effects.

The reference file (cn_models) may be used as a template file for generating a new reference file with different data.

Generating copy number reference on a new human genome array with known fixed regions

Custom Axiom arrays based on the Axiom Precision Medicine Diversity Research Array are enabled for Fixed Region copy number analysis in genes of interest in pharmacogenomics and blood typing. Library files of new arrays will have not a copy number reference file. An Initial Reference must first be generated using Axiom Training Plate, and then a Best Practices Reference must be generated using user samples.

Generating copy number reference on established human genome array with known fixed regions

Established custom Axiom arrays based on the Axiom PMD Research array should have a reference file in the library folder. This reference may be used for CNV analysis. If there is evidence that current data has drifted away from the reference based on QC metrics or performance, then the reference should be updated. The Best Practices Reference generation process should be used to generate an updated reference. The current reference in the library folder may be used as an Initial Reference instead of generating a new one with the Axiom Training Plate. In Axiom Analysis Suite, the CN

Reference File option in the Best Practices Copy Number Reference Creation workflow should point to the current reference. In APT, command line options should point to the current reference.

Generating copy number reference on array with novel fixed regions

If an array has novel fixed regions, a reference must be created from known normal samples in the fixed regions.

If copy number changes in the novel fixed regions of interest are rare, then the reference may be generated from all genotyping and CN QC passing samples. A small fraction of non-normal samples for any fixed region should not affect the reference significantly. If the array is a derivative of the Axiom Precision Medicine Diversity Research Array that is already enabled for Fixed Region copy number analysis in pharmacogenomics and blood typing genes, then the Initial and Best Practices Workflows should generate an appropriate reference for all regions.

In other situations, such as if the novel fixed regions of interest are highly variable or if the array has only novel fixed regions, Axiom Analysis Suite workflows and APT command pipeline may still be used but they require additional advanced steps. Please contact your local Thermo Fisher Scientific Field Application Scientist or thermofisher.com/support (https://www.thermofisher.com/us/en/home/technical-resources.html?CID=fl-support).

Generating copy number reference on array for Discovery analysis

If a Best Practices Reference can be generated for an array, it may be used for Discovery analysis as well. Best Practice Reference generation may not be enabled on some arrays such as those that are designed for Discovery analysis only. In such situations, a basic Copy Number Reference may be created from all genotyping and CN QC passing samples.

Generating copy number reference on array for each plate of data

CNV analysis results depend on the reference used and can change if different references are used. For this reason, it may be best to generate and use a single robust reference for the entire study or project. Such a reference maybe generated using recommendations and workflows described in Chapters 2, 3 and 4 of this Guide.

In the rare circumstance that these recommendations and workflows are inappropriate, a copy number reference may be generated for each plate separately and used for analysis of that plate. Good experimental design practices should be followed such as ensuring that the plate is not mostly cases from a case-control study and that most individual samples on the plate have the normal CN state in regions of interest.

2 Site readiness for copy number reference creation

Overview

Copy number calling in Axiom is not with respect to a published reference genomic sequence. Rather, it is with respect to reference intensity values for each probeset generated from a set of normal samples run on the array. Variability in sample preparation, assay processing, or lab-to-lab variance all are influencers of the performance of a reference. As such, creating a lab-specific reference mitigates some of these variables.

Before generating a Copy Number Reference, laboratories should undergo a Site Assessment. The Site Assessment reviews lab preparedness and QC performance to ensure that the new reference will be appropriate and robust. Lab stability and proficient operation of the assay are critical for Copy Number Reference Creation workflows, especially workflows that utilize the lab's own samples versus workflows that use high-quality samples from the Axiom Training Plate. This chapter describes considerations for site stability and QC performance assessments as well as sample selection guidance for Copy Number Reference Creation.

Site assessment

1. **Lab Preparedness:** The customer should have recently completed the Axiom onboarding training plan or have completed an SOP Review with the local support team. Equipment and processing should adhere to site preparation guides and user guides. Any deviations from methods and equipment substitutions should be approved and documented, following the *Axiom Onboarding Plan (Pub. No. MAN0018543*). Key laboratory preparedness actions described in that document include the following:
    a. Confirm that the laboratory is adhering to the methods and laboratory best practices outlined in the user guide.
    b. Confirm that the GeneTitan MC instrumentation is working within specifications.
    c. Confirm that any liquid handlers are working within specifications and any manual or automated pipettes are appropriately calibrated.
    d. Confirm that all ancillary instrumentation (e.g. incubators, freezers, and thermocyclers) are calibrated to specifications.
    e. Confirm that all operators have demonstrated proficiency with the assay.
2. **Performance Review:** Any poor performance (defined as a plate failure of Genotyping QC) in recent runs should undergo troubleshooting. All issues from the troubleshoot should be addressed. The default thresholds for plate failures include two metrics 1) The Average QC Call Rate for Passing Samples should be ≥ 98.5%. 2) The percent of samples passing the default QC thresholds should be ≥ 95% (≥ 93% for sample types buccal and saliva). These metrics assume the use of the default

sample genotyping QC thresholds which are (DQC ≥ 0.82, QC Call Rate ≥ 97%). If troubleshooting is needed, contact your local Field Application Scientist. Further description of the QC parameters used in the Best Practices Workflow can be found in the *Axiom Genotyping Solution Data Analysis Guide (Pub. No. MAN0018363)*.

QC assessment for CN reference generation using samples from an Axiom training plate

The Axiom Training Plate consists of well-characterized samples. These samples, when processed on a Copy Number capable array, can be leveraged for the creation of an initial CN Reference.

After completing the site assessment and performance review, the processing of an Axiom Training Plate should be performed on the desired array design by operators that have demonstrated proficiency in performing the Axiom Assay.

After processing, analyze the QC using the Best Practices Workflow, and ensure that the plate passes QC before executing CN Reference generation.

QC assessment and sample selection for CN reference generation using customer samples

A more robust Copy Number Reference can be developed using samples processed as part of the lab's normal operations. The reference should be developed using multiple plates, with six plates as the minimal recommendation. The laboratory should be demonstrating lab stability concerning laboratory preparedness and plate performance (outlined above).

The set of samples or CEL files used to generate the reference file can impact the accuracy of copy number calling. Therefore, this set of samples should be chosen carefully. Below are some requirements for sample input as well as recommendations that can improve the robustness or reduce bias within a copy number reference.

Requirements

- Plates should pass genotyping QC from the Best Practices Workflow.
  - Percent Passing Samples: ≥ 95% (≥ 93% for saliva or buccal samples) on a plate should pass the default QC thresholds defined in the Best Practices Workflow (DQC ≥ 0.82, QC Call Rate ≥ 97%).
  - Average QC Call Rate For Passing Samples: Each plate used should meet the criteria of Average QC Call Rate For Passing Samples ≥ 98.5%.
- Samples should pass CN QC criteria specified in the analysis files. CN QC metrics are MAPD and WavinessSD defined in **Chapter 1**.

Recommendations

- Add numerous samples. A minimum of six plates of data is an estimation for a good dataset for building a reference.
- The samples should be representative of the sample type (e.g. blood, buccal, etc.) used for the study.
- If the array contains probesets on the sex chromosomes, using a minimum of 40 samples of each gender is recommended to build references for probesets on the sex chromosomes.
- Include diversity for variables expected during the study. These may include:
  - Plates from different processing periods.
  - Plates processed on different array lots or reagent lots.
  - Plates processed by different operators, liquid handlers, or GeneTitan MC instruments.
- Good experimental design practices include randomizing as many processing variables as possible, 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. (*Chapter 3 of Axiom Genotyping Solution Data Analysis Guide (Pub. No. MAN0018363)*.
- Inclusion of numerous samples in the study with the same aberration (for example large aberrations such as trisomy 21) may lead to a biased reference.
- Monitor MAPD over time. Increasing trends in MAPD or an increased number of calls indicate shifts in process or that a new reference needs to be generated.
- Consider regenerating a reference after major shifts in processing (changing collection or extraction methods, shifting from manual to automated target prep).
- In Agrigenomics applications, a dataset may contain samples from multiple populations that are different from each other in their copy number profiles. In such a situation, the choice of samples to generate a reference depends on study objectives. If the study is to compare populations, a reference built from samples belonging to the base population would work well. A reference built from samples belonging to multiple populations may make interpretation of results challenging. Depending on selection of samples, populations, and their copy number profiles, the reference may represent non-diploid non-integer copy number states at various genomic locations.

3 Reference creation workflows in AxAS

Overview

Some library packages do not include a copy number reference file (cn_models), which is needed for copy number analysis. Library packages that already have a CN reference file may benefit from making a new CN reference file that is more appropriate for the samples run in the lab.

To make the first CN reference file, the library package may support either the **Copy Number Reference Creation** workflow, or the **Initial** Copy Number Reference Creation workflow.

The newer **Initial** workflow requires a training plate of samples available from Thermo Fisher Scientific. The library package may also support the **Best Practices Copy Number Reference Creation** workflow. This workflow has additional checks to improve the quality of the CN reference.

Initial copy number reference creation

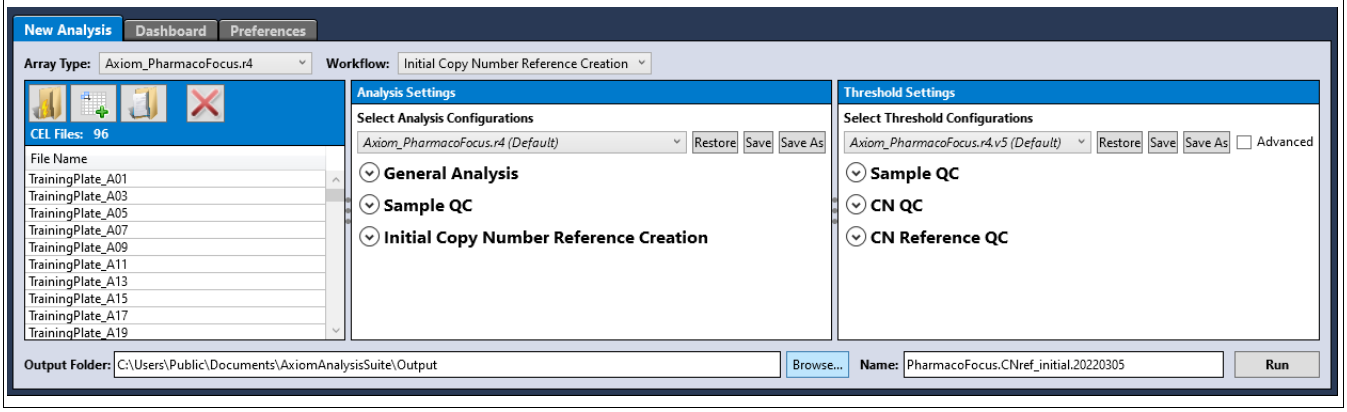
Library files that do not include a CN reference file may support the Initial Copy Number Reference Creation workflow. This workflow has software checks to ensure that CEL files from supported plates, like the Axiom DNA Training Plate 96F, are used. CEL files are matched to recognized samples based on SignatureSNP genotypes (*Chapter 5 of Axiom Genotyping Solutions Data Analysis Guide*). This set of samples with known copy number in fixed regions of interest is used to generate the Initial Copy Number Reference.

Steps to create a reference using the Initial Copy Number Reference Creation workflow are shown below. The example shown is a human genotyping array with 10 fixed regions on CYP2A6, CYP2D6, GSTT1, GSTM1, SULT1A1 and UGT2B17. One Axiom Training Plate data is used to create the Initial Reference.

1. Open Axiom Analysis Suite.

   The New Analysis tab window appears. (Figure 3)

Figure 3 AxAS New Analysis tab window

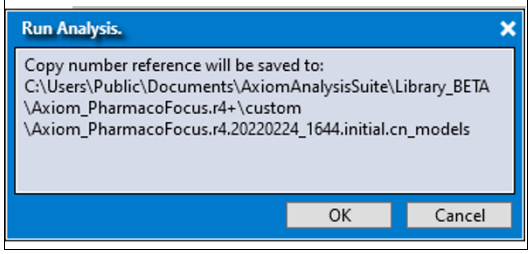


2. Select your Array Type, then Analysis Configuration, then the Workflow **Initial Copy Number Reference Creation**.
3. Select a Threshold Settings option, if not already selected.
4. Add the CEL files from a single training plate run.

   This workflow will create a CN reference using samples that pass sample QC. The analysis batch that is created manages the results of testing the CN reference. The desired name and destination of the new batch should be automatically filled but can be edited.
5. Click **Run**.

   A dialog informs you where the Initial Copy Number Reference will be saved. (Figure 4)

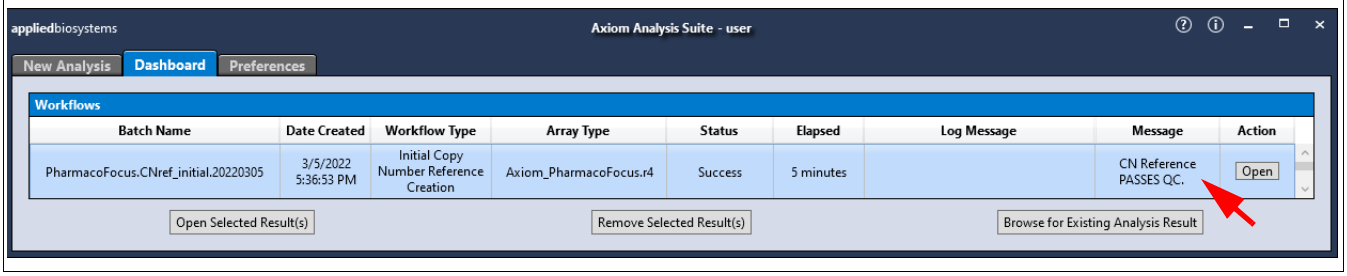
Figure 4 Dialog window



6. Click **OK**.

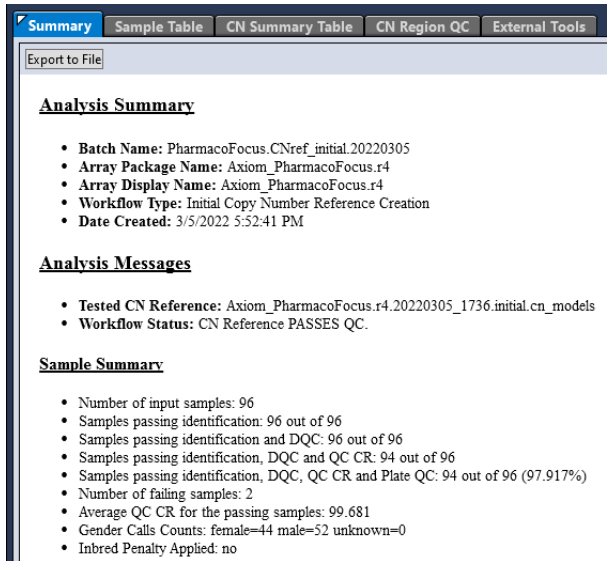
   Once the workflow completes, confirm the Message in the Dashboard is **CN Reference PASSES QC**. If this is what is reported, there is no need to open the batch. The CN reference has been created, tested, and can be selected for genotyping the first few plates, until you have enough plates to make a more robust CN reference.

Figure 5 Dashboard tab window



7. Click **Open** to review the results. The batch will open in the Viewer. In the Summary Tab, the Analysis Messages will list the Tested CN Reference as well as the Workflow Status or an error message.

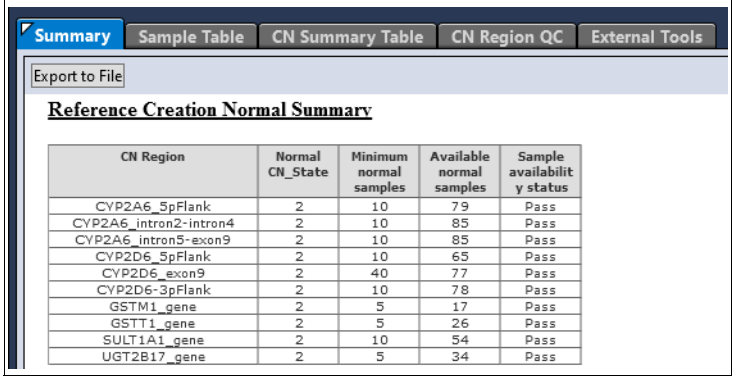
Figure 6 Summary tab window



The Sample Summary section (Figure 6) includes a **Samples passing identification** metric. CEL files are matched to known samples before DishQC is calculated. Samples pass the identification check if they match to known samples AND the samples are arranged on the plate in one of a few allowed arrangements. These checks ensure that expected samples are used when making the important initial reference.

In the Summary tab, the Reference Creation Normal Summary table reports how many normal copy number samples pass QC, and so are available to make a reference for each CN Region. A passing CN reference needs to measure signals from a minimum number of normal copy number samples for each region.

Figure 7 Summary tab window (Reference Creation Normal Summary

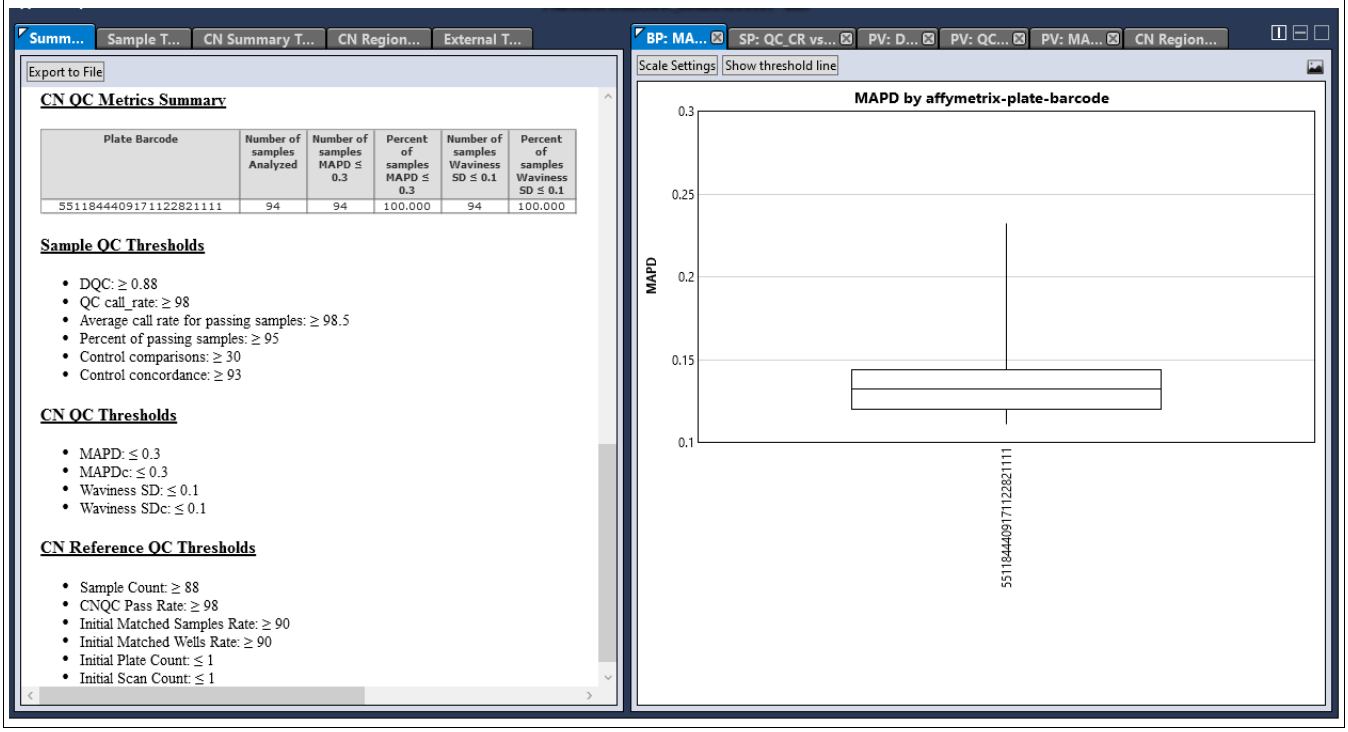


CN Region	Normal CN_State	Minimum normal samples	Available normal samples	Sample availability status
CYP2A6_5pFlank	2	10	79	Pass
CYP2A6_intron2-intron4	2	10	85	Pass
CYP2A6_intron5-exon9	2	10	85	Pass
CYP2D6_5pFlank	2	10	65	Pass
CYP2D6_exon9	2	40	77	Pass
CYP2D6-3pFlank	2	10	78	Pass
GSTM1_gene	2	5	17	Pass
GSTT1_gene	2	5	26	Pass
SULT1A1_gene	2	10	54	Pass
UGT2B17_gene	2	5	34	Pass

8. Scroll down to the CN QC Metrics Summary, CN QC Thresholds and CN Reference QC Thresholds sections (Figure 8)

   Make a note of the MAPD threshold and the CNQC Pass Rate. For the CN reference to pass QC, a minimum percentage of samples tested with the new initial CN reference must pass the MAPD threshold. In the CN QC Metrics Summary, note the percentage of samples that pass the MAPD and Waviness SD thresholds.

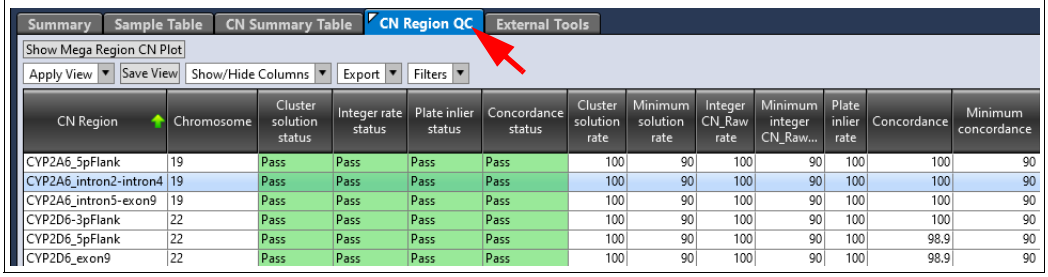
Figure 8 Summary tab window CNQC Metrics Summary and Sample QC Thresholds



Note the MAPD Box plot in right pane. We want samples to have low MAPD, below the maximum MAPD threshold. Samples used to create the CN Reference will 'look' most like the reference, and so will have the lowest MAPD. Low MAPD samples are less likely to have false segment calls in a Copy Number Discovery workflow. Samples with a MAPD above the threshold will not report a fixed region CN call.

9. Click on the **CN Region QC** tab. Confirm that the status column values are Pass for all CN Regions (green background), as shown in Figure 9.

Figure 9 Dashboard tab window

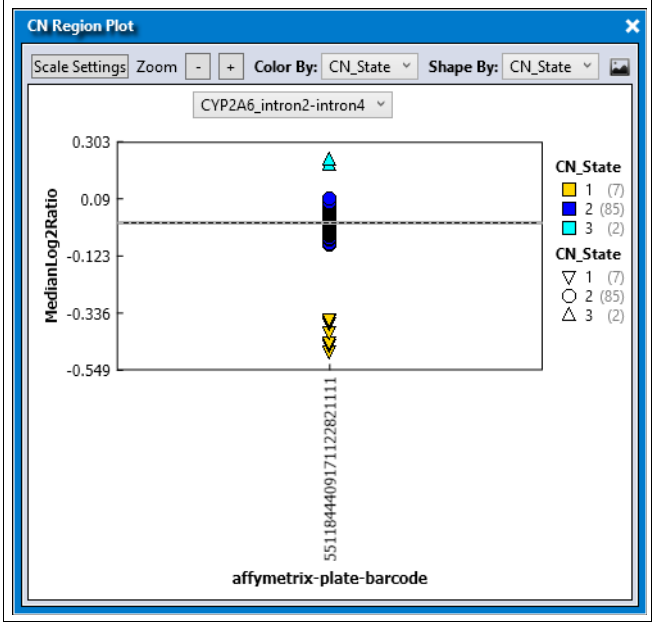


CN Regions table definitions

Column	Description
Cluster solution status	'Pass' if 'Cluster solution rate' is at least 'Minimum solution rate'; 'Fail' otherwise.
Integer rate status	'Pass' if 'Integer CN_raw' rate is at least 'Minimum integer CN_raw rate'; 'Fail' otherwise.
Plate inlier status	'Pass' if 'Plate inlier rate' is above a threshold; 'Review' otherwise.
CN passes MAPD	Yes, if the sample's MAPD value is not greater than the MAPD threshold used by CN QC.
Concordance status	'Pass' if 'Concordance' is at least 'Minimum concordance';'Fail' otherwise.
Cluster solution rate	Percentage of plates for which at least some samples are assigned a call other than NoCall.
Integer CN_Raw rat	Percentage of calls with integer CN_Raw values (i.e. 1,2,3), excluding samples that fail CN QC.
Plate inlier rate	Percentage of plates whose distribution of CN_State calls is consistent with at least one population's expected distribution of CN_State calls or is consistent with the CN_State calls of the other plates in the batch.
Concordance	Percentage of CN_State calls that agree with expected calls, for samples matched to known samples. The library files include reference CN data for some 1000Genomes samples and the Ref103 control sample.

10. Select a CN Region in the CN Region QC table. The Region is displayed in the CN Region Plot, as shown in Figure 10.

Figure 10 CN Region Plot window



11. Scroll through all the CN Regions to see if the CN2 samples are centered on MedianLog2Ratio=0 (the dashed line), as they should be. You can also select the region from the title drop-down menu, and then up/down arrow key to scroll through the plots.

    If the initial CN reference passes QC tests, it is appropriate to use until you can make a more robust multiplate reference using the Best Practices Copy Number Reference Creation workflow.

Best practices copy number reference creation

Best Practices Copy Number Reference Creation is a two-step procedure to make and test an updated CN Reference File. The Step 1 workflow determines which CEL files pass all QC tests. It then uses an existing CN Reference File to select which normal copy number samples should be used to make a new reference for each CN Region.

To start Step 2 of the procedure, you need to:

1. Open the Step 1 analysis results in the viewer.
2. Review and possibly edit the selected **Reference Normal** samples for each CN Region.
3. Click **Create CN Reference** from the Sample Table tab.

   Step 2 of the workflow will then create and test the new CN Reference File.

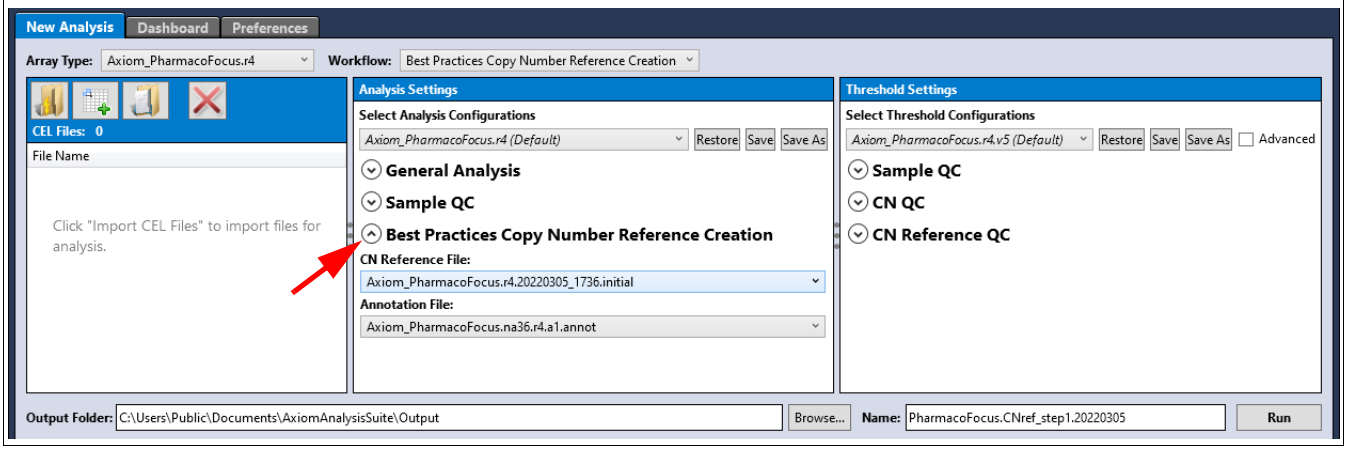
Sample recommendations

- Make the reference using at least six plates of samples
- Include a training plate from Thermo Fisher. Training plate samples and Ref103 control all have known CN states. It is a good idea to have more than a few samples in the batch known to the library package, so that reported concordance is statistically significant.

Steps to create a reference using the Best Practices Copy Number Reference Creation workflow are shown below. The workflow assumes that an Initial Reference exists. The example shown is a human genotyping array with 10 fixed regions on CYP2A6, CYP2D6, GSTT1, GSTM1, SULT1A1 and UGT2B17. Eight plates of data including the Axiom Training Plate are used in this example.

1. Open Axiom Analysis Suite
2. Select your Array Type, then Analysis Configuration, then Workflow **Best Practices Copy Number Reference Creation**.
3. Select a Threshold Settings option, if not already selected.
4. 4.In the Analysis Settings pane, click the arrow to expand the **Best Practices Copy Number Reference Creation** section. (Figure 11)
5. Click the **CN Reference QC** drop-down arrow button to select an existing copy number reference file. If one doesn't exist, you must create one using a different copy number reference creation workflow.
6. Add the CEL files from your multiple plates.

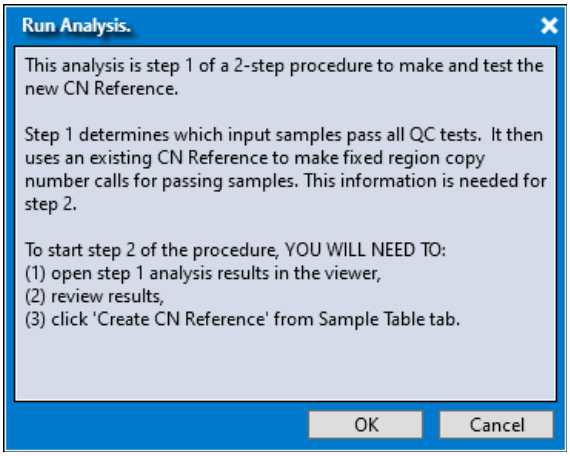
Figure 11 Best Practices Copy Number Reference Creation arrow



7. Click **Run**.

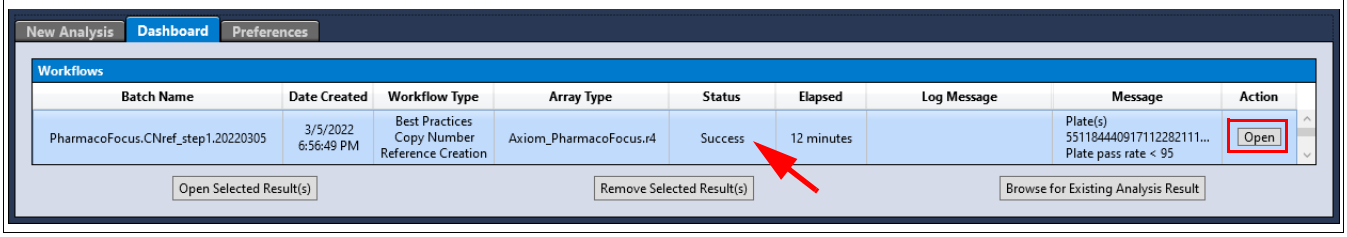
   A **Run Analysis** pop-up window appears. (Figure 12)

Figure 12 Run Analysis message



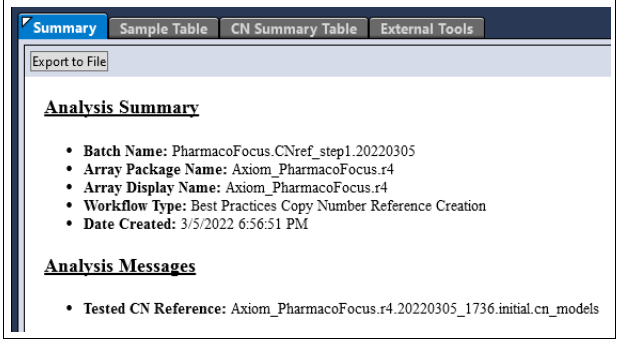
8. Read this message carefully to understand what will happen, then click **OK** to proceed.
9. After the workflow completes, confirm that the **Status** for Step 1 of the analysis is a success, then click the **Open** button. (Figure 13)

Figure 13 Dashboard: Status Success



The Summary tab Analysis Messages section describes the initial CN Reference that this workflow created and tested. (Figure 14)

Figure 14 Summary tab window Workflow created



10. In the Summary tab, check if any plate has an unusually low percent of passing samples. If it appears that a plate suffers from a significant processing issue, no samples from it are appropriate for building a CN reference, and these samples should be excluded later. (Figure 15)

Figure 15 Summary tab window Plate QC Summary

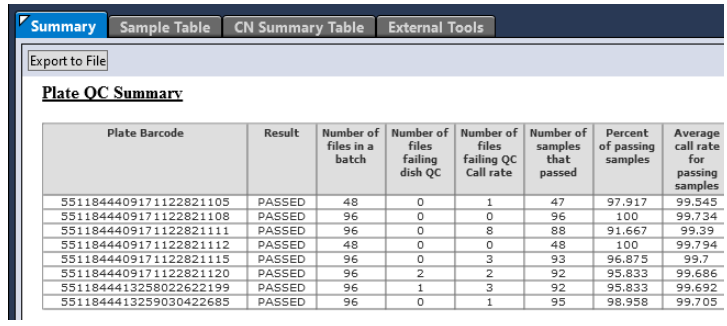


Plate Barcode	Result	Number of files in a batch	Number of files failing dish QC	Number of files failing QC Call rate	Number of samples that passed	Percent of passing samples	Average call rate for passing samples
551184440917112282110 5	PASSED	48	0	1	47	97.917	99.545
551184440917112282110 8	PASSED	96	0	0	96	100	99.734
551184440917112282111 1	PASSED	96	0	8	88	91.667	99.39
551184440917112282111 2	PASSED	48	0	0	48	100	99.794
551184440917112282111 5	PASSED	96	0	3	93	96.875	99.7
551184440917112282112 0	PASSED	96	2	2	92	95.833	99.686
551184441325802262219 9	PASSED	96	1	3	92	95.833	99.692
551184441325903042268 5	PASSED	96	0	1	95	98.958	99.705

11. In the Summary tab, review the percent of samples below the MAPD threshold for each plate. It should be a high percentage for most plates. (Figure 16)

Figure 16 Summary tab window Percentages

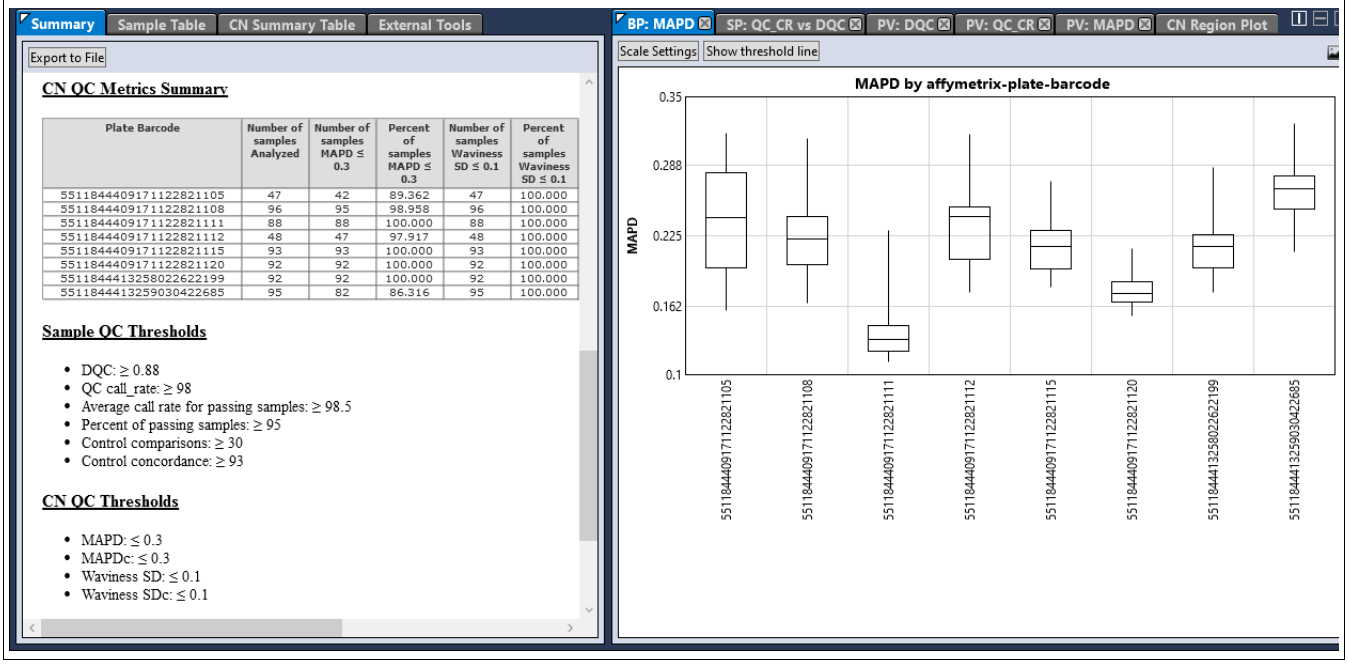


Plate Barcode	Number of samples Analyzed	Number of samples MAPD ≤ 0.3	Percent of samples MAPD ≤ 0.3	Number of samples Waviness SD ≤ 0.1	Percent of samples Waviness SD ≤ 0.1
551184440917112282110 5	47	42	89.362	47	100.000
551184440917112282110 8	96	95	98.958	96	100.000
551184440917112282111 1	88	88	100.000	88	100.000
551184440917112282111 2	48	47	97.917	48	100.000
551184440917112282111 5	93	93	100.000	93	100.000
551184440917112282112 0	92	92	100.000	92	100.000
551184441325802262219 9	92	92	100.000	92	100.000
551184441325903042268 5	95	82	86.316	95	100.000

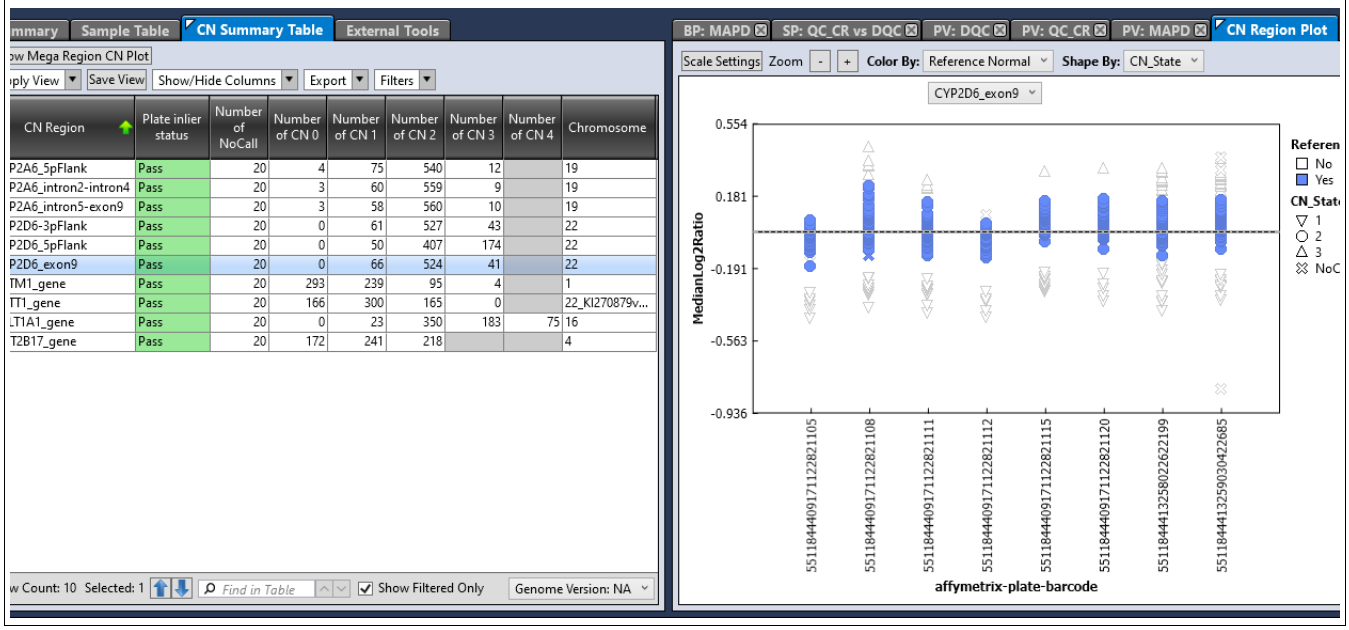
The MAPD Box Plot shows how appropriate it is for each plate to be used for the tested CN reference, when doing CN analysis. Samples with higher MAPD values behave more differently from the samples used to make the tested CN reference. One reason to make a new CN reference using this workflow is to lower the MAPD values for most samples. Lowering the MAPD values should reduce the

number of false segment calls from the Copy Number Discovery workflow and increase the call rate and accuracy of fixed regions from the Copy Number Fixed Regions workflow.

In the MAPD Box Plot, you can click **Show threshold line**, and then drag the threshold line to the CN QC Threshold for MAPD (or type the exact value in the box by the threshold line). These MAPD values are computed using the initial CN reference. In this example, the plate used to build Initial CN Reference has the lowest overall MAPD values as expected.

12. To review the results for each fixed region, navigate to the CN Summary Table and in the right section, click on the CN Region Plot. Click on each region in the CN Summary Table and view the CN Region Plot.

Figure 17 CN Summary Table tab window



The purpose of step 1 of Best Practices Copy Number Reference Creation is to select the normal CN samples for each CN region, in order to make a good multiplate CN reference. The selected normal CN samples are labeled 'Reference Normal' in the CN Region Plot. The CN Summary Table, the CN Region Plot, and Mega Region CN Plot are the primary visualizations to help you decide if the software made good Reference Normal selections.

In the CN Summary Table, **Plate inlier status** column has values of Pass or Review for each region. 'Pass' means that, for a given region, AxAS does not detect more than 10% of the plates in the batch to be outliers (threshold set in library files). An outlier plate is one that has unusually few measured normal CN samples, or has BOTH:

a. An unusual distribution of CN calls compared to multiple known populations AND

b. An unusual plate-averaged CN_State compared to the other plates in the batch

When a plate is labeled as an outlier, NO samples from that plate are typically selected

as Reference Normal. If a sample is not labeled as Reference Normal, it will not be used when making a CN reference for this CN Region.

IMPORTANT! If a CN Region has 'Plate inlier status' of 'Review', it should be reviewed to confirm or edit the handling of outlier plates. It is possible that the CN calling algorithm assigned the wrong CN State to the samples.

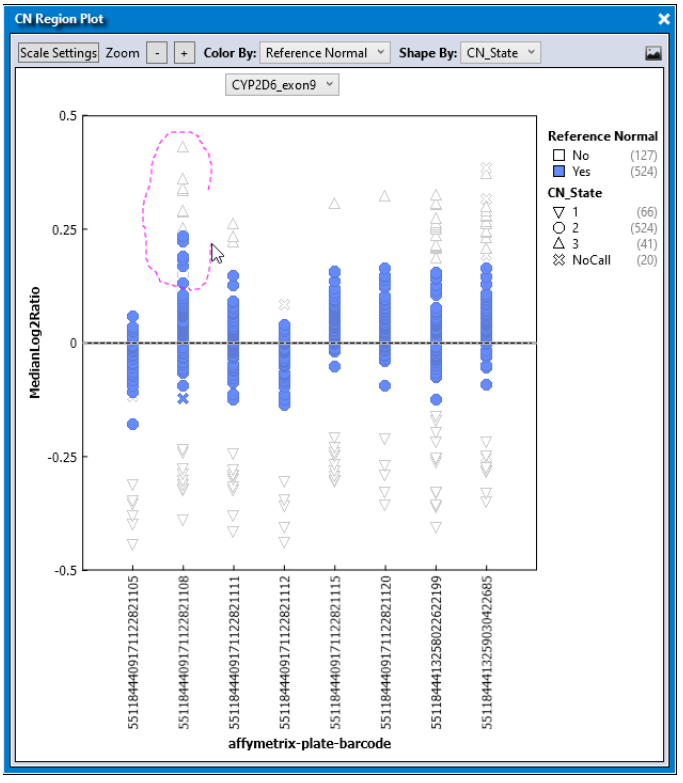
For details on outlier tests, see Chapter 5.

How to evaluate the CN Region Plot for a step1 Best Practices CN Reference Creation batch:

- The blue data points are the samples AxAS selected as appropriate to use as normal samples when building the CN reference for that region. Different regions will have different 'Reference Normal' samples selected.
- The normal CN samples for each plate should be near the horizontal dashed line, which represents MedianLog2ratio=0. The default shape of the samples is by 'CN_State'. Since there are between-plate effects, it is not unusual for the normal CN sample cluster (usually CN2) to shift a little above and below the line. If the labeled normal CN cluster has shifted too far away from the dashed line, miscalls become more likely.
- Since the normal CN state for most regions is 2, most or all of the CN2-shaped data points should be blue. It is possible for a sample to be labeled as Reference Normal even if the labeled CN_State is not 2, and vice versa. This can happen if the sample has matched the CEL file to a recognized sample and has knowledge of the expected CN_State.
- If any plate for a region has NO blue data points (Reference Normal samples), that plate is considered an outlier. A plate with no Reference Normal samples will not be used to create a reference for the region even if it has some samples labeled by shape as CN2.

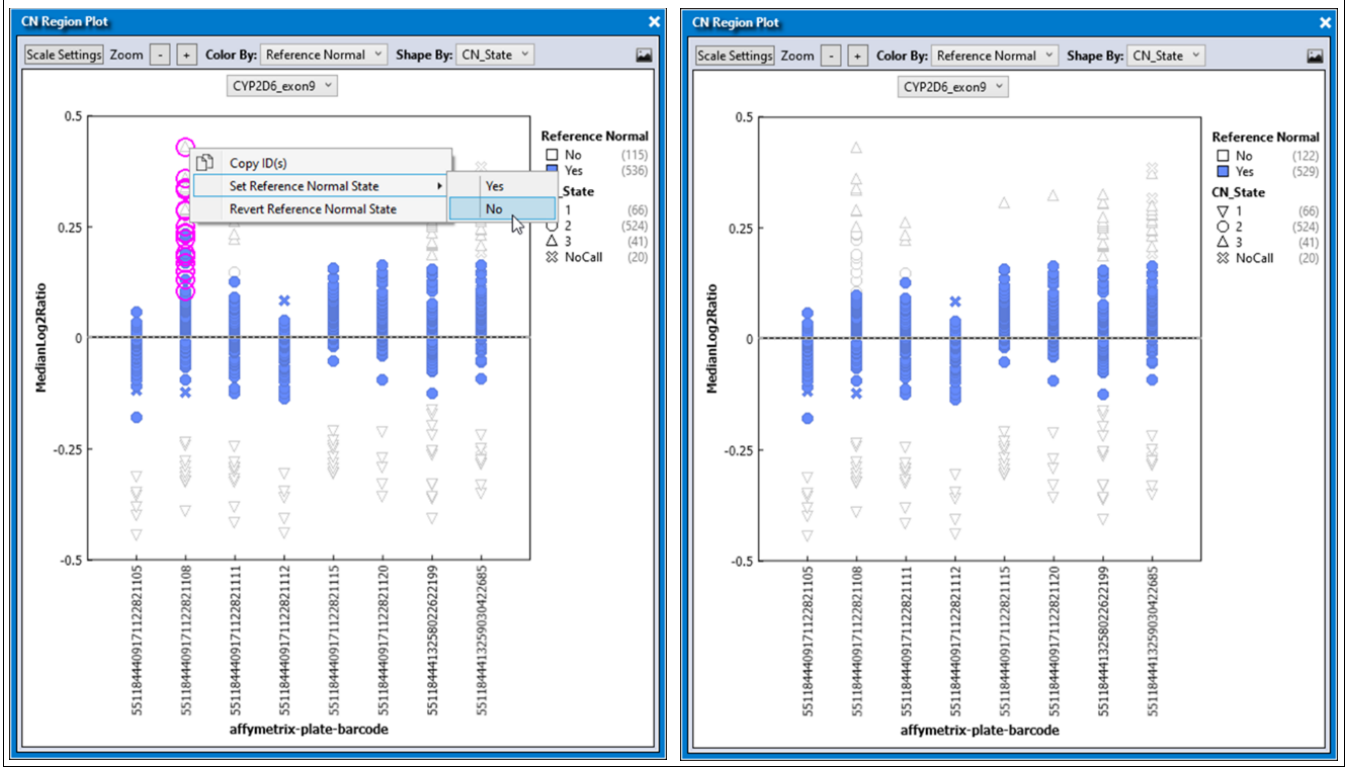
13. To change which samples are selected as Reference Normal for a region, drag-select (lasso) the data points to edit. (Figure 18)

Figure 18 CN Region Plot window Lasso



14. 16.Use the secondary mouse button to click on the selected samples. Then select Set Reference Normal State, and in the sub-menu, select the desired option. (Figure 19)
    a. a.To remove samples from the Reference Normal set, select **Set Reference Normal State > No**.
    b. b.To add samples to the Reference Normal set, select **Set Reference Normal State > Yes**.

Figure 19 CN Region Plot window Change Reference

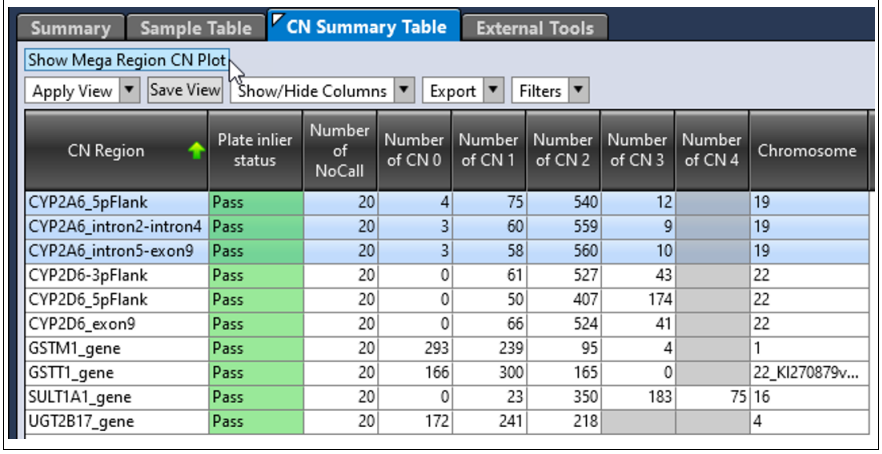


15. To undo all changes to the Reference Normal set, select some or all samples, then secondary mouse click on them, and select **Revert Reference Normal**. Reverting changes must be done separately for each region.

    Note: To make a good CN Reference, it helps if the Reference Normal set includes ONLY normal samples for each region, not samples that appear to have CN gains or losses. It is OK if some of the normal sample are not included, as long as most plates include a high percentage of their normal samples labeled as Reference Normal.

16. 18.When there are multiple regions for the same gene, it helps to use neighboring region CN States to help confirm your choices for Reference Normal. The steps below show how to evaluate the selection of Reference Normal samples for CYP2A6_intron2-intron4 in the example dataset. From the CN Summary Table, select all CYP2A6 rows, and click **Show Mega Region CN Plot**. (Figure 20)

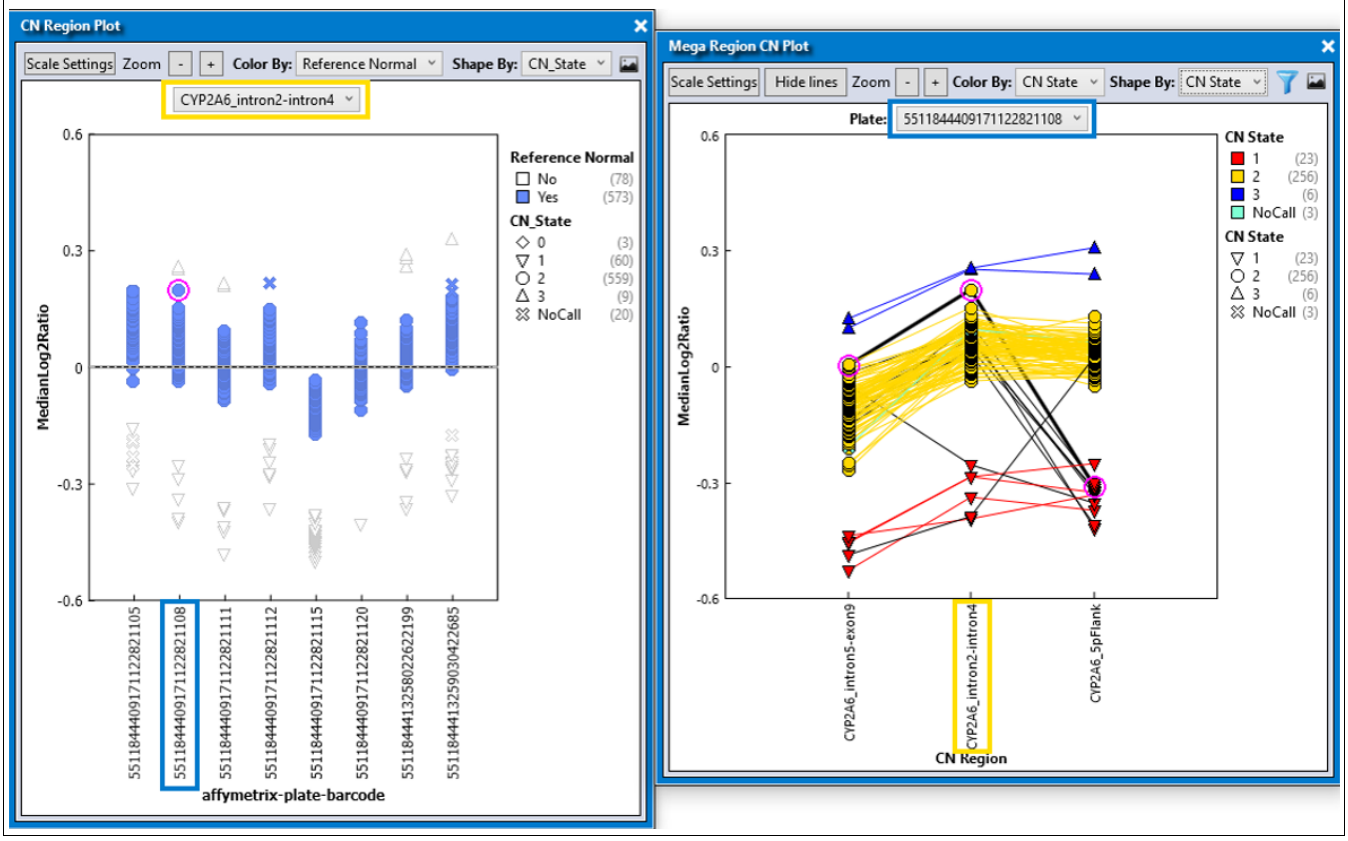
Figure 20 Show Mega Region CN Plot



CN Region	Plate inlier status	Number of NoCall	Number of CN 0	Number of CN 1	Number of CN 2	Number of CN 3	Number of CN 4	Chromosome
CYP2A6_5pFlank	Pass	20	4	75	540	12		19
CYP2A6_intron2-intron4	Pass	20	3	60	559	9		19
CYP2A6_intron5-exon9	Pass	20	3	58	560	10		19
CYP2D6-3pFlank	Pass	20	0	61	527	43		22
CYP2D6_5pFlank	Pass	20	0	50	407	174		22
CYP2D6_exon9	Pass	20	0	66	524	41		22
GSTM1_gene	Pass	20	293	239	95	4		1
GSTT1_gene	Pass	20	166	300	165	0		22_KI270879v...
SULT1A1_gene	Pass	20	0	23	350	183	75	16
UGT2B17_gene	Pass	20	172	241	218			4

a. In order to see both the CN Region Plot and the Mega Region CN Plot at same time, you can pop out one or both tabs. This can be done by clicking the white triangle (upper left corner of the tab). While the CN Region plot shows one region across all plates, the Mega Region CN Plot shows one plate across all selected regions of same chromosome. A line connects one sample across its regions. (Figure 21)

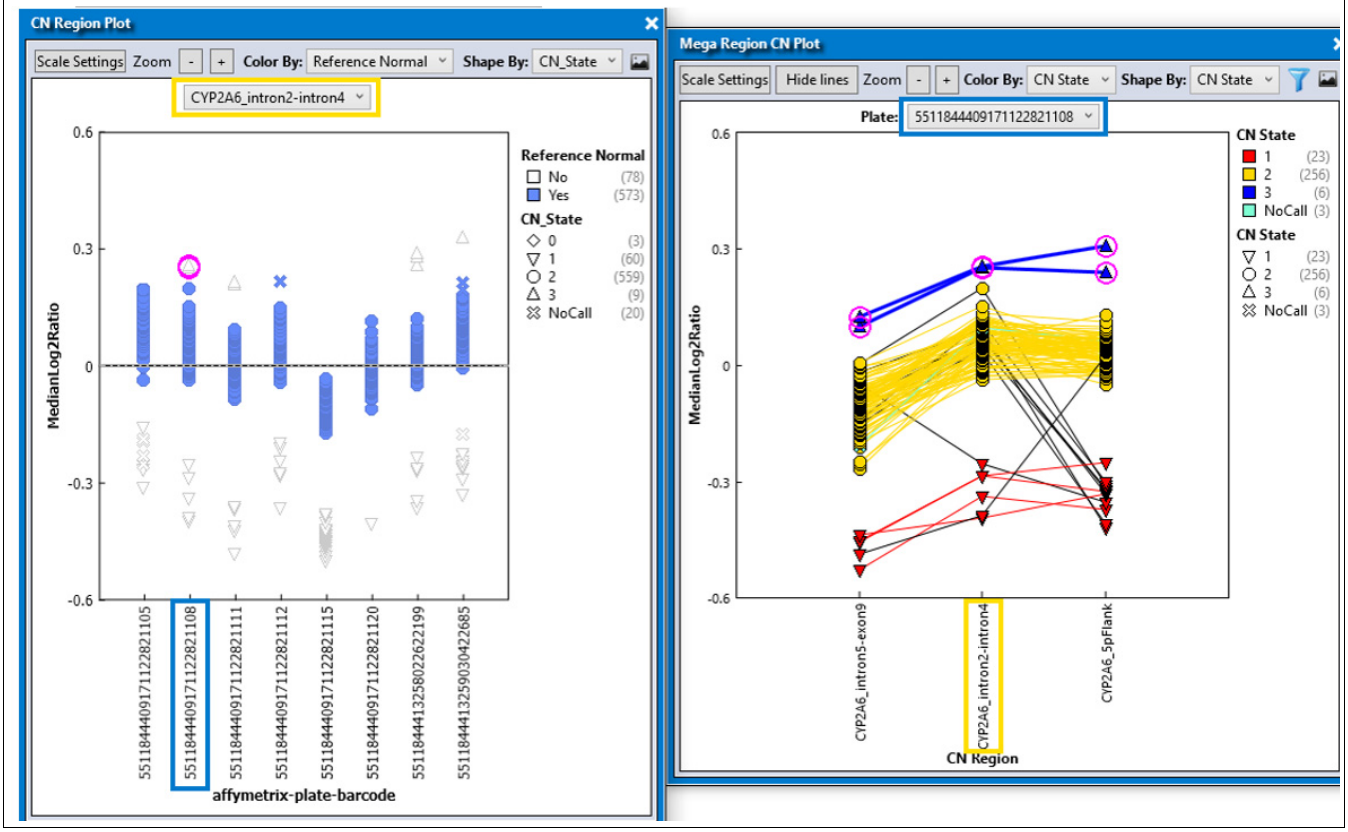
Figure 21 CN Region Plot vs. Mega Region CN Plot 1



b. If you want to determine the CN States of the flanking regions of the selected sample in the CN Region Plot, then we need to select the same plate from the Mega Region Plot. In the shown example, the highlight sample is reported as CN2 for CYP2A6_intron2-intron4, and is reported as CN2 for intron5-exon9, and CN1 for 5pFlank. It is reasonable to think that this sample is NOT CN3 for intron2-intron4, and therefore is properly assigned as Reference Normal for this region.

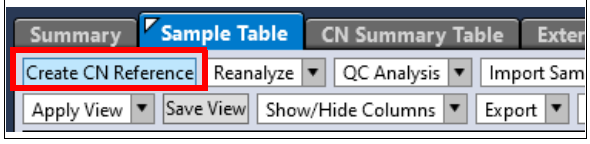
c. If instead you select in the CN Region Plot the two samples from the same plate that are reported as CN3, we can see in the Mega Region Plot that those same samples are also reported as CN3 in both flanking regions. Therefore, it is reasonable that these samples are excluded from the Reference Normal set for intron2-intron4.

Figure 22 CN Region Plot vs. Mega Region CN Plot 2



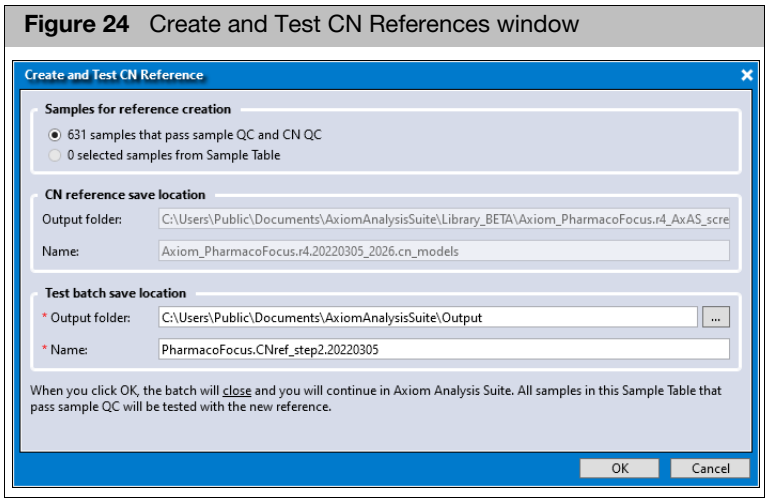
17. After reviewing and optionally editing the samples selected as Reference Normal, it is time to start the second half of this workflow. Click **Create CN Reference** button in the Sample Table tab. (Figure 23)

Figure 23 Create CN Reference button



The **Create and Test CN Reference** window appears. (Figure 24)

Figure 24 Create and Test CN References window



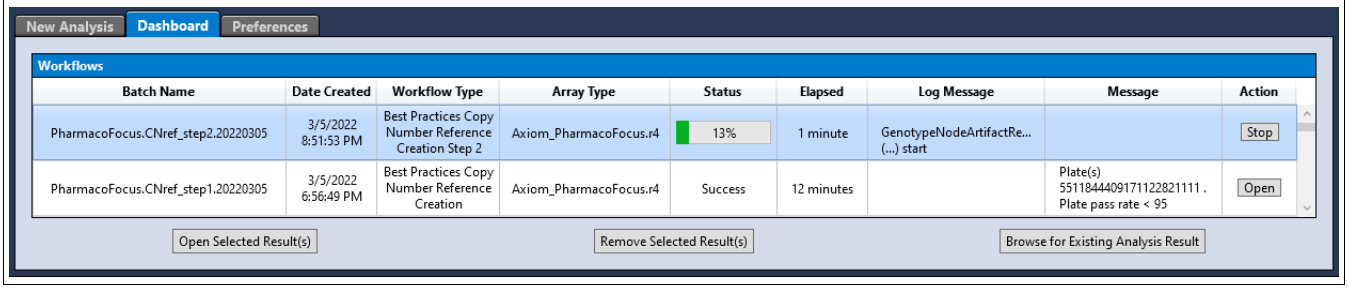
18. In the **Samples for reference creation** section, select the set of samples used to make the new CN Reference. The default option is to use all samples that pass both sample QC and CN QC (Sample Table "Pass/Fail" column = Pass, and "CN passes QC" = yes). If you want to select a different set from the Sample table, then 1) select the desired sample rows from the Sample Table, 2) click **Create and Test CN Reference**, and 3) select the **selected samples from Sample Table** option.

    Note: Regardless of which samples are selected to make the new CN reference, all samples in this batch that pass sample QC will be tested using the new CN reference.

    - The CN reference save location is displayed and cannot be changed.
    - The test batch save location is pre-filled, but can be changed.

19. Click **OK** to start the second half of this workflow. The Viewer closes and the Dashboard displays the progress of the new workflow 'Best Practices Copy Number Reference Creation Step 2'. (Figure 25)

Figure 25 Dashboard Best Practices Copy Number Reference Creation Step 2



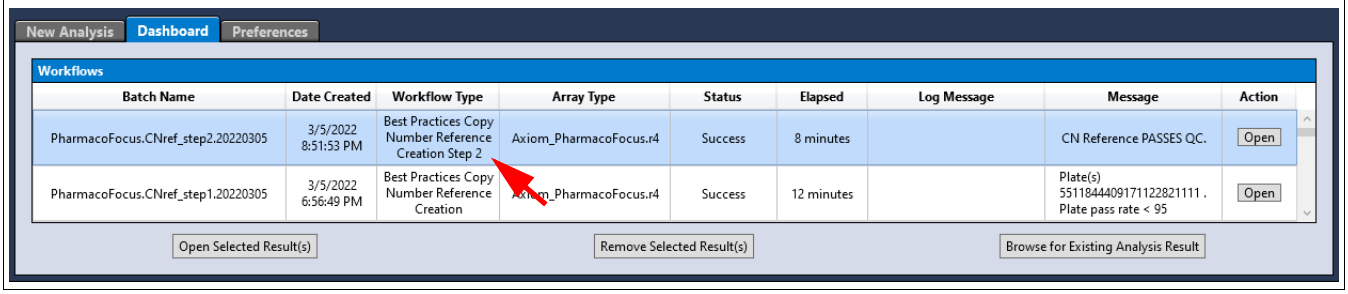
Once the run completes, confirm that the message in the Message column of the Dashboard for this analysis is CN Reference PASSES QC.

IMPORTANT! At this point the new CN reference exists, and you can use either the previous CN reference or the new CN reference for another workflow. However, it is essential to first review the test results of this step 2 batch. Since the CN results are from samples processed using a CN reference trained on the same data, future CN results on new plates may not look better than these test results. If the current results are not satisfactory, now is the best time to try to improve the CN reference, before using it routinely.

Reviewing the best practices CN reference

1. From the Dashboard, select the batch that tests the best practices CN reference you just created. Its Workflow Type is **Best Practices Copy Number Reference Creation Step 2**. (Figure 26)

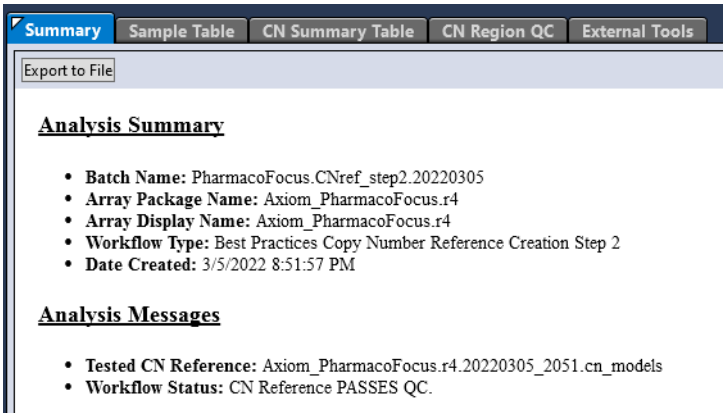
Figure 26 Dashboard Reviewing Best Practices Copy Number Reference Creation Step 2



After the run completes, confirm that the message in the Message column of the Dashboard for this analysis is **CN Reference PASSES QC**.

2. To open this batch, click **Open**.
3. From the Summary tab window, review the Analysis Messages section. (Figure 27)

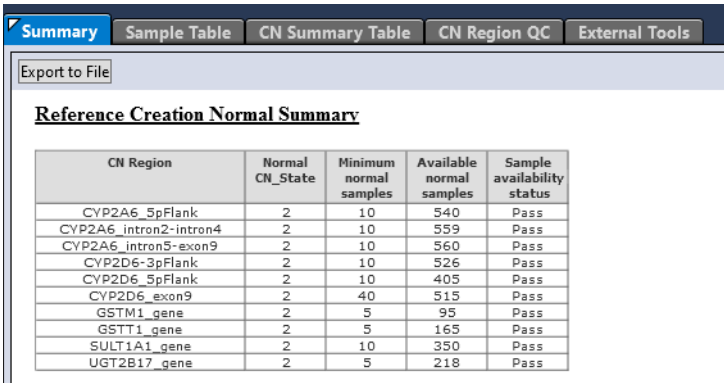
Figure 27 Summary tab window Analysis Messages



Make sure the Sample Summary and Plate QC Summary are the same as the Step1 batch from which this Step2 batch was created.

4. Review the number of normal samples used to create the reference for each CN Region in the **Reference Creation Normal Summary** table. (Figure 28) A good CN Reference requires Sample availability status to be a Pass for all regions.

Figure 28 Reference Creation Normal Summary table



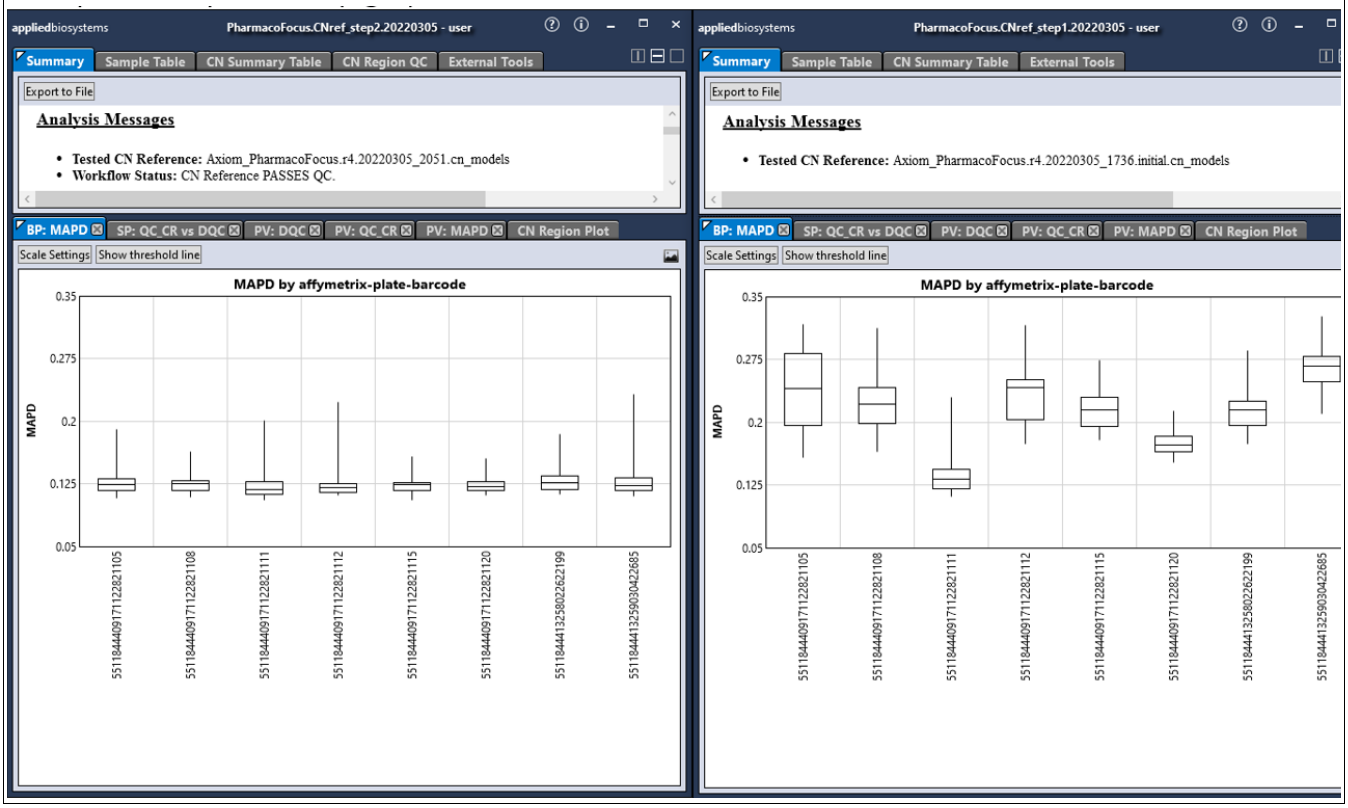
CN Region	Normal CN_State	Minimum normal samples	Available normal samples	Sample availability status
CYP2A6_5pFlank	2	10	540	Pass
CYP2A6_intron2-intron4	2	10	559	Pass
CYP2A6_intron5-exon9	2	10	560	Pass
CYP2D6-3pFlank	2	10	526	Pass
CYP2D6_5pFlank	2	10	405	Pass
CYP2D6_exon9	2	40	515	Pass
GSTM1_gene	2	5	95	Pass
GSTT1_gene	2	5	165	Pass
SULT1A1_gene	2	10	350	Pass
UGT2B17_gene	2	5	218	Pass

5. Scroll down to the CN QC Metrics Summary table. Confirm that a high percentage of samples pass both MAPD and WavinessSD thresholds. A good CN Reference requires that the percentage of samples passing CN QC also meets the CNQC Pass Rate threshold.
6. Scroll down to the CN QC Metrics Summary table. Confirm that a high percentage of samples pass both MAPD and WavinessSD thresholds. A good CN Reference requires that the percentage of samples passing CN QC also meets the CNQC Pass Rate threshold.
7. Switch to the Sample Table tab. One of the columns is "Used in Reference" (usually last column), which identifies the samples in this table that were used to create the multiplate reference being tested. If the proportion of rows that have

"Used in Reference = No" is unexpectedly high, you may decide to create another CN reference from a re-opened Step 1 batch, this time manually selecting samples from the Step 1 batch Sample Table.

At this point it helps to see how the best practices CN reference tested in step 2 compares to the previous CN reference used in step 1, when tested on the same data. The following screen captures compares the MAPD box plot of the current step 2 batch (left), to the MAPD box plot from a re-opened step 1 batch (right), as shown in Figure 29.

Figure 29 Step 2 vs. Step 1 Compare 1



It is better for MAPD values to be low and more consistent across plates. The Best Practices CN Reference built on these plates (step 2 batch) is a better reference for CN analysis for this data than the previous reference used in the step 1 batch.

8. For the step 2 batch that evaluates the multiplate reference, switch to the CN Region QC tab, and check whether all regions pass all tests. (Figure 30)

Figure 30 CN Region QC tab window

CN Region QC

Show Mega Region CN Plot | Apply View | Save View | Show/Hide Columns | Export | Filters

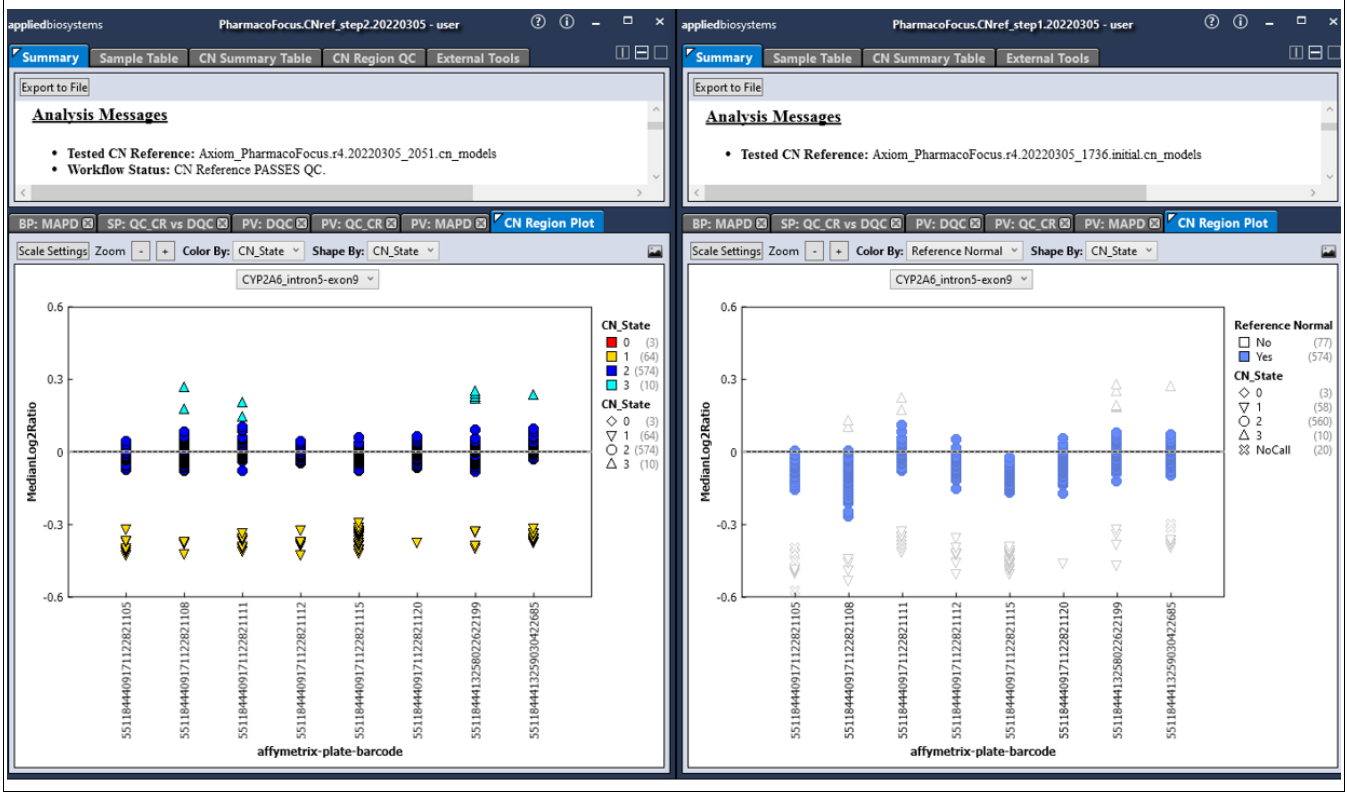
CN Region	Chromosome	Cluster solution status	Integer rate status	Plate inlier status	Concordance status	Cluster solution rate	Minimum solution rate	Integer CN_Raw rate	Minimum integer CN_Raw...	Plate inlier rate	Concordance	Minimum concordance
CYP2A6_5pFlank	19	Pass	Pass	Pass	Pass	100	90	100	90	100	100	90
CYP2A6_intron2-intron4	19	Pass	Pass	Pass	Pass	100	90	100	90	100	100	90
CYP2A6_intron5-exon9	19	Pass	Pass	Pass	Pass	100	90	100	90	100	100	90
CYP2D6-3pFlank	22	Pass	Pass	Pass	Pass	100	90	100	90	100	99.2	90
CYP2D6_5pFlank	22	Pass	Pass	Pass	Pass	100	90	100	90	100	99.2	90
CYP2D6_exon9	22	Pass	Pass	Pass	Pass	100	90	100	90	100	99.2	90
GSTM1_gene	1	Pass	Pass	Pass	Pass	100	90	100	90	100	100	90
GSTT1_gene	22_KI270879v...	Pass	Pass	Pass	Pass	100	90	100	90	100	100	90
SULT1A1_gene	16	Pass	Pass	Pass	Pass	100	90	100	90	100	99.2	80
UGT2B17_gene	4	Pass	Pass	Pass	Pass	100	90	100	90	100	100	90

Row Count: 10 Selected: 1 | Find in Table | Show Filtered Only | Genome Version: NA

Note: Concordance columns appear if at least one sample is matched to samples for which the library files have expected CN calls. Known samples include Ref103 and many 1000Genomes project samples.

9. Review the CN Region Plot. See if all plates for all regions appear to have the CN2 samples near the horizontal dashed line. To see the benefit of the new CN reference vs previous one, it helps to have the CN Region Plot from BOTH the step 2 and previous step 1 batches open side by side, as shown in Figure 31.

Figure 31 Step 2 vs. Step 1 Compare 1



In the step2 vs step1 comparison, you can see the impact that the new reference has on the results. The right plot shows the blue samples selected for building a new multiplate CN reference. The left plot shows how the multiplate reference is better able to correct for between-plate variability. Having the CN2 samples more centered at MedianLog2Ratio makes correct CN calling easier. Also, the clusters are sometimes smaller, which makes it easier to resolve the different copy number states within each plate.

If the CN Reference PASSES QC and a visual inspection of the CN Region Plot across all regions reveals all plates made CN2 calls near the MedianLog2Ratio=0 line, then this CN Reference is ready for routine use. If the CN Reference File passes QC, it is OK to use it for other CN analyses.

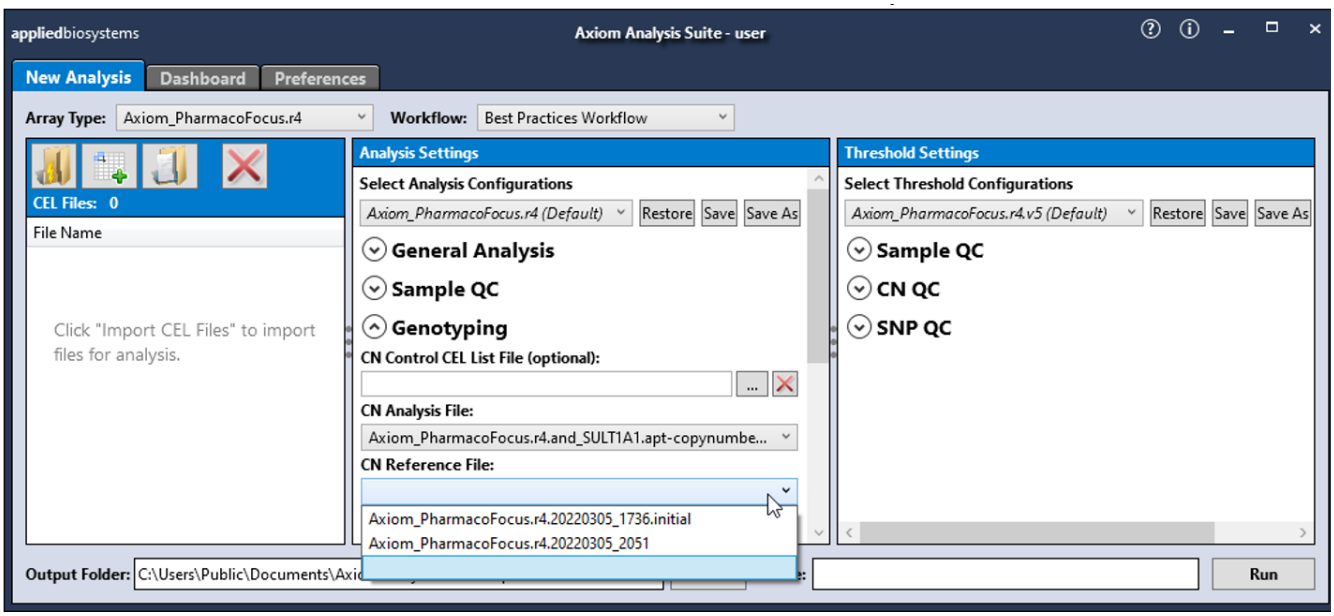
10. Close the batches.

Saving a preferred CN reference for later use

At this point you may have created multiple CN references. However, you still need to select the preferred reference every time you use it. This section describes how to save the preferred CN reference to an analysis configuration.

1. From New Analysis tab, select your Array Type, and an analysis configuration.
2. From Workflow menu, select each workflow that can use the CN Reference File, like Copy Number Discovery, Copy Number Fixed Regions, and (for CN-aware genotyping arrays) the Genotyping workflow.
3. Notice that the CN Reference File default is blank. (Figure 32) Select it to see available options. Confirm you can see both the Initial and Best Practices Reference files you created earlier.

Figure 32 Dashboard Blank CN Reference File entry



IMPORTANT! For routine use, you should use the most recent reference that has passed QC tests when being created. Multiplate references are preferred over "initial" single plate references.

4. From the CN Reference File drop-down menu, select the best practices multiplate reference you created. Confirm you did NOT choose any "initial" reference.
5. Click **Save As** and save the custom analysis configuration, maybe as "custom".
6. The next time you use the custom analysis configuration in a workflow, that configuration is auto-selected by default for later workflows.

Copy number reference creation

Library files that do not include the Best Practices Copy Number Reference Creation workflow may support the Copy Number Reference Creation workflow. This workflow generates a reference file using all selected CEL files. Because there is no sample QC filtering, you should only select samples that pass Sample QC as part of the Sample QC workflow. Each sample should be mostly CN neutral. It is recommended that at least 80 samples be supplied, with at least 40 female and 40 male samples. AxAS will warn you if these minimum counts are not met.

If you intend to use a new CN reference for Copy Number Fixed Regions analysis, you should instead use the Best Practices Copy Number Reference Creation workflow if it is available. This is because the Copy Number Reference Creation workflow does not assist you in making one of the optional inputs, the CN Region Calls file. This file tells AxAS which samples are CN normal for each fixed region.

Steps to create a CN reference are as follows:

1. Open Axiom Analysis Suite.
2. Select your Array Type, then an Analysis Configuration, then Workflow Copy Number Reference Creation.
3. Add the CEL files.

   If you intend to use the CN reference only for **Copy Number Discovery** workflow, you do not need to select a **CN Region Calls File**. However, be aware that CN hypervariable fixed regions may give unreliable CN segment calls.

   If you intend to also perform a **Copy Number Fixed Regions** analysis either as a stand-alone workflow or as part of the **Genotyping** or **Best Practices** workflows for supported arrays, then:

   a. Find an existing Copy Number Fixed Regions or Genotyping or Best Practices batch that was analyzed on the **same** set of CEL files.

   b. Using Windows Explorer, confirm that the existing batch has the folder 'CNData' with the file 'AxiomCNVmix.cnregioncalls.txt' or 'AxiomCN.cnregioncalls.txt'.

c. Open the existing batch in AxAS Viewer, then review the CN Region Plot for each fixed region. Depending on the AxAS version you have, you may need to select the region from the **CN Summary Table** to update the **CN Region Plot**. Check if the likely CN normal samples are usually assigned a CN_State of 2.

d. Close the existing batch.

e. Back in the AxAS **New Analysis** window, click the **Copy Number Reference Creation** down arrow to expand the section, then click the **Browse** button for the **CN Region Calls File**, and browse to the ***.cnregioncalls.txt** file that you found earlier.

4. (Optional) To change the default annotation file, click the **Copy Number Reference Creation** down arrow to expand the section, then click to highlight the **Annotation File** you want to use to create your reference file.

5. Enter a name for the CN reference in the **Name** field, then click **Run**.

   The Dashboard appears and displays the progress of this workflow. When it completes, the new CN reference will be available for future workflows. This CN The analysis batch cannot be opened for viewing. The new CN reference is saved to the 'custom' sub-folder of the current library folder.

Reference creation using APT

Overview

The copy number reference creation workflows described in Chapter 3 automate the many steps required to generate a robust reference. It is therefore recommended that these workflows in AxAS be used for reference creation.

A reference file (.cn_models) may also be created using command line options in APT, but this may require manual intermediate steps.

For details on individual APT programs, refer to:
https://www.affymetrix.com/support/developer/powertools/changelog/index.html

The library file package for the array is assumed to be downloaded into an accessible directory in the system.

The example scripts below are directly usable by Linux users. For readability, the Linux commands are broken into several lines with the backslash character: "`\`". When executing these commands in Windows, the backslash character should be removed, the input folder path should be specified using Windows conventions, and the entire command should be put in one line.

Where command line options point to directories such as the input, output, or analysis directories, the full path to the appropriate directory should be specified. For example, `<analysis_directory>` is the full path to the library file package directory for the array. Input and output directory names are specific to a particular analysis run. Input directory names in the scripts below may point to directories created in earlier steps.

Where command line options point to library files, their typical extensions are shown in the scripts. Library file names typically include the name of the array. They may also include additional tags such `na<#>` or `r<#>` where `#` is a revision number.

Where command line options point to input or output files, an example filename is provided.

Command line usage

The copy number reference file is created by the `apt-copynumber-axiom-ref` engine. This file may be generated using either CEL files or the results of a genotyping run as input. If the array is enabled for fixed regions with variable copy number, users may specify normal samples for reference in the fixed regions.

For a description of all options type: `apt-copynumber-axiom-ref --help`

Generating a reference file starting with CEL files

An example command with input cel_list.txt is shown below. This file has one column with the header 'cel_files' and a list of all input CEL files with full paths specified.

```
apt-copynumber-axiom-ref \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<#>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml \
--apt-axiom-cnv-node:annotation-file <analysis_directory>/arrayname.na<#>.r<#>.annot.db \
--cel-files <input_directory>/cel_list.txt \
--out-dir <output_directory> \
--log-file <output_directory>/apt-copynumber-axiom-ref.log \
--new-reference-file <output_directory>/arrayname.r<#>.cn_models
```

If an annotation file is specified in the .xml file and exists in `--analysis-files-path`, then the `--apt-axiom-cnv-node:annotation-file` option is not required in the script.

Generating a reference file starting with the results of a previous genotyping run

An example command with input summary.a5, calls.txt, and report.txt is shown below. Genotyping outputs are assumed to be in the `<genotyping_directory>`.

```
apt-copynumber-axiom-ref \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<#>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml \
--apt-axiom-cnv-node:annotation-file <analysis_directory>/arrayname.na<#>.r<#>.annot.db \
--skip-genotyping TRUE \
--summary-file <genotyping_directory>/AxiomGT1.summary.a5 \
--adjusted-intensities-node:calls-file <genotyping_directory>/AxiomGT1.calls.txt \
--cn-library-file-node:report-file <genotyping_directory>/AxiomGT1.report.txt \
--out-dir <output_directory> \
--log-file <output_directory>/apt-copynumber-axiom-ref.log \
--new-reference-file <output_directory>/arrayname.r<#>.cn_models
```

The command shown above assumes that the summary file is in .a5 format. If using `AxiomGT1.summary.txt file`, include: `--use-text-format-summary-file TRUE`

The a5 format is set by default and recommended for the input summary file, because it is faster than the txt format especially for larger datasets. The AxiomGT1.summary.txt file can be converted to a5 format by using apt2-a5-table-converter.

If an annotation file is specified in the .xml file and exists in `--analysis-files-path`, then the `--apt-axiom-cnv-node:annotation-file` option is not required in the script.

Generating a reference file specifying normal samples in fixed regions

Normal reference samples in fixed regions on the array may be specified in an input tab delimited text file using the `--fixed-cn-region-calls-file` option. The required columns are 'cel_files', 'CN_Region' and 'CN_State'. For each fixed region, CEL files with CN_State=2 are used to create the reference. Other rows in the input are ignored.

```
apt-copynumber-axiom-ref \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<#>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml \
--apt-axiom-cnv-node:annotation-file <analysis_directory>/arrayname.na<#>.r<#>.annot.db \
--cel-files <input_directory>/cel_list.txt \
--out-dir <output_directory> \
--log-file <output_directory>/apt-copynumber-axiom-ref.log \
--fixed-cn-region-calls-file <input_directory>/arrayname.cnregioncalls.txt \
--new-reference-file <output_directory>/arrayname.r<#>.cn_models
```

An advanced feature allows generation of reference when the normal reference state for a fixed region is not 2. Use:

```
--normal-cn-fixed-regions-file <input_directory>/arrayname.normalcnstate.txt
```

The input is a tab delimited text file. The required columns are 'CN_Region' and 'CN_State_Normal' where CN_State_Normal specifies the normal CN_State for the corresponding CN_Region. CN reference for CN_Regions will be computed using CEL files that have CN_State equal to CN_State_Normal in the cnregioncalls.txt file.

If annotation file is specified in the .xml file and exists in `--analysis-files-path`, then the `--apt-axiom-cnv-node:annotation-file` option is not required in the script.

Generating a wave-correction enabled reference

Wave correction is a method to correct batch effects in log2 ratios and improve accuracy of CNV analysis, especially for the Discovery method. Probeset specific wave-correction values are computed from the data used for reference generation and are stored in the reference file. If wave-correction parameters are specified in the array analysis file, arrayname.r<#>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml, then a wave-correction enabled reference file is created. If wave-correction parameters are not specified in the analysis file, a wave-correction enabled reference can be generated using the additional options to apt-copynumber-axiom-ref

```
--use-wave-correction true
--wave-count -1
```

This generates a reference file with the number of wave correction values set equal to the number of unique plate barcodes in the input dataset.

Initial Copy Number Reference Creation

This reference creation workflow is primarily intended for the Axiom Precision Medicine Diversity Research Array and derivative custom arrays that are enabled for Fixed Region CNV analysis in genes of interest in pharmacogenomics and blood typing. The copy number reference for such arrays must be created from known normal samples and the Axiom Training Plate should be used to generate the Initial Reference.

It is recommended that users run the Initial Copy Number Reference Creation workflow in Axiom Analysis Suite, which automates all the steps in this process. Command line options for running key steps in APT are shown below. The Initial Reference Creation workflow in Axiom Analysis Suite performs additional steps that must be done manually when running the process in APT. Workflow parameters and thresholds are specified in analysis XML files and in the ax_threshold file in the analysis directory.

The input is a list of CEL files (cel_list.txt). This text file has one column with the header 'cel_files' and a list of all input CEL files with full paths specified. All XML files that are passed using the `--arg-file` option or the `--xml-file` option must exist in the analysis directory.

Sample and plate genotyping QC

All samples used for reference generation must pass genotyping QC checks.

Calculate DishQC by running:

```
apt-geno-qc \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-geno-qc.AxiomQC1.xml \
--cel-files <input_directory>/cel_list.txt \
--out-dir <genoqc_directory> \
--log-file <genoqc_directory>/ apt-geno-qc.log
```

The column 'axiom_dishQC_DQC' in the output apt-geno-qc.txt has DishQC values. Remove samples that fail Dish QC and create cel_list2.txt for input in the next step.

Calculate QC Call Rate by running:

```
apt-genotype-axiom \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.96orMore_Step1.r<>.apt-genotype-axiom.SnpSpecificPriors.AxiomGT1.apt2.xml \
--cel-files <input_directory>/cel_list2.txt \
--out-dir <qccr_directory> \
--log-file <qccr_directory>/apt-genotype-axiom.log
```

The column 'call_rate' in the output AxiomGT1.report.txt has the QC Call Rate values. Remove samples that fail QC Call Rate criteria and create cel_list3.txt for input in the next step.

For more details on genotyping QC, see *Chapter 8 in Axiom Genotyping Solutions Data Analysis Guide*.

Plate average QCCR for passing samples must be above the minimum rate. Workflow requires a minimum number of passing samples. Thresholds are specified in the analysis XML and ax_thresholds files in the analysis directory.

Sample and plate copy number QC

All samples used for reference generation must also pass copy number QC checks. QC metrics - MAPD and WavinessSD - are calculated from log2 ratios which require a reference file, which may not exist at this stage of the analysis. The command shown below generates a temporary reference file and then calculates copy number QC metrics for all samples. QC metrics are written out in the `qc-results-file`. The temporary reference file is not used in subsequent steps of the analysis.

```
apt-copynumber-axiom-ref \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml \
--apt-axiom-cnv-node:annotation-file <analysis_directory>/arrayname.na<>.r<>.annot.db \
--use-wave-correction False \
--cel-files <input_directory>/cel_list3.txt \
--out-dir <cnqc_directory> \
--log-file <cnqc_directory>/apt-copynumber-axiom-ref.log \
--new-reference-file <cnqc_directory>/TempReference.cn_models \
--qc-results-file <cnqc_directory>/TempReference.cnref_qc_report.txt
```

The output cnref_qc_report.txt file has MAPD and WavinessSD values. Remove samples that fail MAPD and WavinessSD criteria and create cel_list4.txt for input in the next step. Workflow requires a minimum number of passing samples. Thresholds are specified in the analysis XML and ax_thresholds files in the analysis directory.

Create input file of normal reference samples for fixed regions

Copy number reference must be created from known normal samples. Normal reference samples are specified in a required tab delimited text file called cnregionnormalcalls.txt with columns 'cel_files', 'CN_Region', and 'CN_State'. Such a file must be created manually by appropriately joining information from the following sources:

- CEL files that pass all QC checks in the steps above should be used (cel_list4.txt).
- The mapping from CEL file to and plate well may be known. If unknown, the CEL file ('cel_files') to plate well mapping ('affymetrix-plate-peg-wellposition') can be obtained from the AxiomGT1.report.txt output of `apt-genotype-axiom`.
- The Axiom Training Plate well mapping file (plate_map.txt) is a tab-delimited text file in the analysis directory. It provides the mapping between plate well ('well') and sample ('sample_name').

- The cn_region_refcalls file in the analysis file directory is a tab-delimited text file with columns 'Sample', 'CN_Region', 'CN_State' and 'Population'. The file contains known CN states in fixed regions for many samples from the International HapMap Project including all samples on the Axiom Training Plate.

Generate initial reference

The initial reference can now be created using cel_list4.txt with all QC passing samples and cnregionnormalcalls.txt file with normal reference samples as inputs.

```
apt-copynumber-axiom-ref \

--analysis-files-path <analysis_directory> \

--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-
axiom-ref.AxiomCNref.apt2.xml \

--apt-axiom-cnv-node:annotation-file <analysis_directory>/
arrayname.na<>.r<>.annot.db \

--cel-files <input_directory>/cel_list4.txt \

--fixed-cn-region-calls-file <input_directory>/
arrayname.cnregionnormalcalls.txt \

--out-dir <initialref_directory> \

--log-file <initialref_directory>/apt-copynumber-axiom-ref.log
\

--new-reference-file <initialref_directory>/
arrayname.r<>.initial.cn_models
```

This initial reference (initial.cn_models) should be tested. Testing may be done on the same data that was used to generate the reference.

Run analysis with the initial reference

The initial reference is used to perform Fixed Region CNV analysis (Chapter 6) on the same plate of data. Use the initial.cn_models, summary.a5 and report.txt files generated in the step above as inputs.

```
apt-copynumber-axiom-cnvmix \

--analysis-files-path <analysis_directory> \

--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-
axiom-cnvmix.AxiomCNVmix.apt2.xml \

--summary-file <initialref_directory>/AxiomGT1.summary.a5 \

--report-file <initialref_directory>/AxiomGT1.report.txt \

--reference-file <initial_directory>/
arrayname.r<>.initial.cn_models \

--out-dir <cnvmitest_directory> \

--log-file <cnvmixtest_directory>/apt-copynumber-axiom-
cnvmix.log
```

The results of the CNV analysis using the initial reference file are in the cnvmixtest_directory.

Create a sample name lookup file for the test data

Testing the data includes a concordance check against known CN states of HapMap samples. A sample name lookup file (samplenamelookup.txt) in tab delimited text format is required. The required columns are 'cel_files' and 'Sample'. This file must be created manually by appropriately joining information from the following sources

- CEL files in the reference that should be used (cel_list4.txt).
- The mapping from CEL file to and plate well may be known. If unknown, the CEL file ('cel_files') to plate well mapping ('affymetrix-plate-peg-wellposition') can be obtained from the AxiomGT1.report.txt output of apt-genotype-axiom.
- The Axiom Training Plate well mapping file (plate_map.txt) is a tab-delimited text file in the analysis directory. It provides the mapping between plate well ('well') and sample ('sample_name').

Perform QC check of the analysis results

The results of the CNV analysis using the Initial CN Reference can now be tested using `apt-copynumber-cnvmix-util` with the `--write-qc true` option. The input files are either in the analysis directory or in the cnvmixtest directory

```
apt-copynumber-cnvmix-util \

--write-qc true \

--cn-region-refsettings-file <analysis_directory>/
arrayname.r<>_cn_region_refsettings.txt \

--refcalls-file <analysis_directory>/
arrayname.r<>_cn_region_refcalls.txt

--fixed-cn-region-details-file <cnvmixtest_directory>/
batchname.cnregions.details.txt \

--fixed-cn-region-calls-file <cnvmixtest_directory>/
batchname.cnregioncalls.txt \

--fixed-cn-region-priors-file <analysis_directory>/
arrayname.r<>.cn_priors \

--samplenamelookup-file <input_directory>/samplenamelookup.txt
\

--out-dir <writeqc_directory> \

--log-file <writeqc_directory>/apt-copynumber-cnvmix-util.log
```

Review the output files. The cnregions.qc.txt file contains a table of QC metrics and corresponding Pass/Fail/Review status for all fixed regions on the array. The reference passes if it does not fail any of the status checks. The header of this file includes status check messages. For more details see the Definitions table in Chapter 3.

An Initial Reference file (cn_models) that passes QC may be used to do preliminary copy number analysis of other plates of data and to select samples to build a more robust reference from more diverse samples using the Best Practices Reference Creation workflow.

Best Practices Reference Creation

The Best Practices Reference Creation workflow is used a create a robust reference using a diverse set of samples. It assumes that an Initial Reference file (cn_models) exists.

It is recommended that users run the Best Practices Reference Creation workflow in Axiom Analysis Suite, which automates all the steps in this process. Command line options for running key steps in APT are shown below. The Best Practices Reference Creation workflow in Axiom Analysis Suite performs additional steps that must be done manually when running the process in APT. The workflow also allows user interaction in selecting samples for reference generation which cannot be replicated in APT. Workflow parameters and thresholds are specified in analysis XML files and in the ax_thresholds file in the analysis directory.

The input is a list of CEL files (cel_list.txt). This text file has one column with the header 'cel_files' and a list of all input CEL files with full paths specified. All XML files that are passed using the `--arg-file` option must exist in the analysis directory.

Sample and plate genotyping QC

All samples used for reference generation must pass genotyping QC checks.

Calculate DishQC by running:

```
apt-geno-qc \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-geno-qc.AxiomQC1.xml \
--cel-files <input_directory>/cel_list.txt \
--out-dir <genoqc_directory> \
--log-file <genoqc_directory>/ apt-geno-qc.log
```

The column 'axiom_dishQC_DQC' in the output apt-geno-qc.txt has DishQC values. Remove samples that fail Dish QC and create cel_list2.txt for input in the next step.

Calculate QC Call Rate by running:

```
apt-genotype-axiom \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/
arrayname.96orMore_Step1.r<>.apt-genotype-
axiom.SnpSpecificPriors.AxiomGT1.apt2.xml \
--cel-files <input_directory>/cel_list2.txt \
--out-dir <qccr_directory> \
--log-file <qccr_directory>/apt-genotype-axiom.log
```

The column 'call_rate' in the output AxiomGT1.report.txt has the QC Call Rate values. Remove samples that fail QC Call Rate criteria and create cel_list3.txt for the next step.

For more details, see *Chapter 8 in Axiom Genotyping Solutions Data Analysis Guide*.

Plate average QCCR for passing samples must be above the minimum rate. Workflow requires a minimum number of passing samples. Thresholds are specified in the analysis XML and .ax_thresholds files in the analysis directory.

Select normal reference samples for fixed regions

Copy number reference must be created from known normal samples. However, normal reference samples for fixed regions are often not known in the diverse set of samples used for Best Practices Reference Creation. To deal with this, the Initial Reference is used to generate a set of putative copy number normal samples from the diverse set of samples. After a filtering step that omits samples showing suspicious calls and a relabeling step that adjusts copy number calls for samples with known copy number states, these putative normal samples are used to generate the Best Practices CN Reference.

1. Probeset summarization

   The first step in the process is to do probeset summarization of the input samples. Only samples that pass genotyping QC must be used (cel_list3.txt).

```
apt-genotype-axiom \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-genotype-axiom.AxiomCN_PS1.apt2.xml \
--cel-files <input_directory>/cel_list3.txt \
--out-dir <summarization_directory> \
--log-file <summarization_directory>/apt-genotype-axiom.log
```

2. Fixed Region analysis

   The next step is to do Fixed Region copy number analysis of these samples using summarization results from above and the Initial Reference.

```
apt-copynumber-axiom-cnvmix \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-axiom-cnvmix.AxiomCNVmix.apt2.xml \
--summary-file <summarization_directory>/AxiomGT1.summary.a5 \
--report-file <summarization_directory>/AxiomGT1.report.txt \
--reference-file <initialref_directory>/
arrayname.r<>.initial.cn_models
--out-dir <putativenormals_directory> \
--log-file <putativenormals_directory>/apt-copynumber-axiom-cnvmix.log \
```

See Chapter 6 for detailed description of outputs. The `cnregioncalls.txt` file contains a putative list of normal samples for each fixed region. Only samples that pass copy number QC may report a copy number.

The output report.txt file shows which samples passed CNQC (values in column 'CN passes QC' is 'yes'). Starting with cel_list3.txt which has all genotyping QC passing samples, create an updated CEL list cel_list4.txt restricting to CEL files that also pass CNQC.

3. Write reference normals.

   Reference normal must be determined from the putative list of normal samples. The output of the Fixed Region analysis is tested for plate outliers and sample concordance. Copy number calls are set to NoCall or adjusted to known states. This step includes concordance check against known CN states of HapMap samples. A sample name lookup file (samplenamelookup.txt) in tab delimited text format is required. The required columns are 'cel_files', 'Sample' with column names in the header. This file must be created manually by appropriately joining information from the following sources:

   - CEL files used for Fixed Region analysis must be used (cel_list3.txt).
   - The mapping from CEL file to and plate well may be known. If unknown, the CEL file ('cel_files') to plate well mapping ('affymetrix-plate-peg-wellposition') can be obtained from the AxiomGT1.report.txt output of apt-genotype-axiom
   - The Axiom Training Plate well mapping file (plate_map.txt) is a tab-delimited text file in the analysis directory. It provides the mapping between plate well ('well') and sample ('sample_name').

Reference normal samples are written out using `apt-copynumber-cnvmix-util` with the `--write-reference-normals true` option. Chapter 5 describes the process in detail. The input files are either in the analysis directory or in the cnvmixtest directory.

```
apt-copynumber-cnvmix-util \

--write-reference-normals true \

--cn-region-refsettings-file <analysis_directory>/
arrayname.r<>_cn_region_refsettings.txt \

--refcalls-file <analysis_directory>/
arrayname.r<>_cn_region_refcalls.txt

--fixed-cn-region-details-file <putativenormals_directory>/
batchname.cnregions.details.txt \

--fixed-cn-region-calls-file <putativenormals_directory>/
batchname.cnregioncalls.txt \

--fixed-cn-region-priors-file <analysis_directory>/
arrayname.r<>.cn_priors \

--samplenamelookup-file <input_directory>/samplenamelookup.txt
\

--out-dir <writerefnormals_directory> \

--log-file <writerefnormals_directory>/apt-copynumber-cnvmix-
util.log
```

Review the outputs. The cnregionnormalcalls.txt file contains the best estimate of normal samples in each region after testing and adjustment. This file is tab delimited with three columns - 'cel_files', 'CN_Region' and 'CN_State'. Samples that are normal (usually CN_State=2) for each region will be used to generate a reference. This file should be edited only if there is clear evidence that specific CN_States must be adjusted further.

Generate Best Practices Reference

All samples used for reference generation must pass both genotyping and copy number QC checks (cel_list4.txt). The normal reference sample text file cnregionnormalcalls.txt from the above step is also required.

```
apt-copynumber-axiom-ref \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml \
--apt-axiom-cnv-node:annotation-file <analysis_directory>/arrayname.na<>.r<>.annot.db \
--cel-files <input_directory>/cel_list4.txt \
--fixed-cn-region-calls-file <writerefnormals_directory>/arrayname.cnregionnormalcalls.txt \
--out-dir <bestpracticesref_directory> \
--log-file <bestpracticesref_directory>/apt-copynumber-axiom-ref.log \
--new-reference-file <bestpracticesref_directory>/arrayname.r<>.cn_models
```

The Best Practices Reference (cn_models) should be tested. Testing may be done on the same data that was used to generate the reference.

Run analysis with the reference

The Best Practices Reference is used to perform Fixed Region CNV analysis (Chapter 6) on the same plate of data. Use the cn_models, summary.a5 and report.txt files generated in the step above as inputs.

```
apt-copynumber-axiom-cnvmix \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-axiom-cnvmix.AxiomCNVmix.apt2.xml \
--summary-file <bestpracticesref_directory>/AxiomGT1.summary.a5 \
--report-file <bestpracticesref_directory>/AxiomGT1.report.txt \
--reference-file <bestpracticesref_directory>/arrayname.r<>.cn_models \
--out-dir <cnvmixtest_directory> \
--log-file <cnvmixtest_directory>/apt-copynumber-axiom-cnvmix.log
```

The results of the CNV analysis using the Best Practices Reference file are in the cnvmixtest_directory.

Create a sample name lookup file for the test data

Testing the data includes a concordance check against known CN states of HapMap samples. A sample name lookup file (samplenamelookup.txt) in tab delimited text

format is required. The required columns are 'cel_files', 'Sample' with column names in the header. This file must be created manually by appropriately joining information from the following sources

- CEL files in the reference should be used (cel_list4.txt).
- The mapping from CEL file to and plate well may be known. If unknown, the CEL file ('cel_files') to plate well mapping ('affymetrix-plate-peg-wellposition') can be obtained from the AxiomGT1.report.txt output of `apt-genotype-axiom`.
- The Axiom Training Plate well mapping file (plate_map.txt) is a tab-delimited text file in the analysis directory. It provides the mapping between plate well ('well') and sample ('sample_name').

Perform QC check of the analysis results

The results of the CNV analysis using the Best Practices Reference can now be tested using `apt-copynumber-cnvmix-util` with the `--write-qc true` option. Chapter 5 describes the QC checking criteria in detail. The input files are either in the analysis directory or in the cnvmixtest directory.

```
apt-copynumber-cnvmix-util \

--write-qc true \

--cn-region-refsettings-file <analysis_directory>/
arrayname.r<>_cn_region_refsettings.txt \

--refcalls-file <analysis_directory>/
arrayname.r<>_cn_region_refcalls.txt

--fixed-cn-region-details-file <cnvmixtest_directory>/
batchname.cnregions.details.txt \

--fixed-cn-region-calls-file <cnvmixtest_results_directory>/
batchname.cnregioncalls.txt \

--fixed-cn-region-priors-file <analysis_directory>/
arrayname.r<>.cn_priors \

--samplenamelookup-file <input_directory>/samplenamelookup.txt
\

--out-dir <writeqc_directory> \

--log-file <writeqc_directory>/apt-copynumber-cnvmix-util.log
```

Review the output files. The cnregions.qc.txt file contains a table of QC metrics and corresponding Pass/Fail/Review status for all fixed regions on the array. The reference passes if it does not fail any of the status checks. The header of this file includes status check messages. For more details see the Definitions table in Chapter 3.

A Best Practices Reference that passes QC can be used for CNV analysis of other plates of data.

Copy Number Reference Creation

If a Best Practices Reference can be generated for an array, it may be used for Discovery analysis as well. Best Practice Reference generation may not be enabled on some arrays such as those that are designed for Discovery analysis only. In such situations, a reference may be created from all QC passing samples.

The input is a list of CEL files (cel_list.txt). This text file has one column with the header 'cel_files' and a list of all input CEL files with full paths specified. All XML files that are passed using the `--arg-file` option must exist in the analysis directory.

Sample and plate genotyping QC

All samples used for reference generation must pass genotyping QC checks.

Calculate DishQC by running:

apt-geno-qc \

--analysis-files-path <analysis_directory> \

--arg-file <analysis_directory>/arrayname.r<>.apt-geno-qc.AxiomQC1.xml \

--cel-files <input_directory>/cel_list.txt \

--out-dir <genoqc_directory> \

--log-file <genoqc_directory>/ apt-geno-qc.log

The column 'axiom_dishQC_DQC' in the output apt-geno-qc.txt has DishQC values. Remove samples that fail Dish QC and create cel_list2.txt for input in the next step.

Calculate QC Call Rate by running:

```
apt-genotype-axiom \

--analysis-files-path <analysis_directory> \

--arg-file <analysis_directory>/
arrayname.96orMore_Step1.r<>.apt-genotype-
axiom.SnpSpecificPriors.AxiomGT1.apt2.xml \

--cel-files <input_directory>/cel_list2.txt \

--out-dir <qccr_directory> \

--log-file <qccr_directory>/apt-genotype-axiom.log
```

The **call_rate** column in the output AxiomGT1.report.txt has the QC Call Rate values. Remove samples that fail QC Call Rate criteria and create `cel_list3.txt` for input in the next step.

For more details, see *Chapter 8 in Axiom Genotyping Solutions Data Analysis Guide*.

Plate average QCCR for passing samples must be above the minimum rate. Workflow requires a minimum number of passing samples. Thresholds are specified in the analysis **XML** and **ax_thresholds** files in the analysis directory.

Sample and plate copy number QC

All samples used for reference generation must also pass copy number QC checks. QC metrics - MAPD and WavinessSD - are calculated from log2 ratios which require a reference file, which may not exist at this stage of the analysis. The command shown

below generates a temporary reference file and then calculates copy number QC metrics for all samples. QC metrics are written out in the `qc-results-file`. The temporary reference file is not used in subsequent steps of the analysis.

```
apt-copynumber-axiom-ref \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml \
--apt-axiom-cnv-node:annotation-file <analysis_directory>/arrayname.na<>.r<>.annot.db \
--use-wave-correction False \
--cel-files <input_directory>/cel_list3.txt \
--out-dir <cnqc_directory> \
--log-file <cnqc_directory>/apt-copynumber-axiom-ref.log \
--new-reference-file <cnqc_directory>/TempReference.cn_models \
--qc-results-file <cnqc_directory>/TempReference.cnref_qc_report.txt
```

The `output cnref_qc_report.txt` file has MAPD and WavinessSD values. Remove samples that fail MAPD and WavinessSD criteria and create `cel_list4.txt` for input in the next step. Workflow requires a minimum number of passing samples. Thresholds are specified in the analysis **XML** and **ax_thresholds** files in the analysis directory.

Generate a reference

To generate a reference use:

```
apt-copynumber-axiom-ref \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-axiom-ref.AxiomCNref.apt2.xml \
--apt-axiom-cnv-node:annotation-file <analysis_directory>/arrayname.na<>.r<>.annot.db \
--cel-files <input_directory>/cel_list4.txt \
--out-dir <output_directory> \
--log-file <output_directory>/apt-copynumber-axiom-ref.log \
--new-reference-file <output_directory>/arrayname.r<>.cn_models
```

Note: This reference file may be used for CNV analysis of other plates of data.

5 Identifying normal reference samples in fixed regions

Overview

Copy number calling in Axiom requires a copy number reference, which is a set of probeset signal intensity values generated from copy number normal samples run on the array. It is important that samples used to generate the reference capture the diversity of signal intensities expected from samples that will be analyzed with the reference.

The Best Practices Copy Number Reference Generation Workflow creates a robust fixed-region wave-correction enabled copy number reference using a set of known copy number normal samples and a subset of the customer's own samples. This ensures that the reference includes samples that reflect the diversity of populations analyzed by the customer. The workflow relies on several methods to identify which customer samples are copy number normal (usually copy number 2 for autosomal fixed regions) and should be included in the reference.

A Best Practices Copy Number Reference is typically composed of a training plate and six to ten customer plates. The Axiom Training Plate (provided by Thermo Fisher Scientific) consists of well-characterized HapMap samples with known copy number. The training plate is used to generate an Initial Copy Number Reference and perform a copy number analysis with the customer's plates. A putative set of copy number normal samples are determined using CNVmix, the algorithm that assigns copy number states to each sample (for details see Chapter 6).

Because the population of samples used to generate the Initial Copy Number Reference may not capture the full diversity of the customer's samples, CNVmix may report incorrect copy number calls for some samples. To prevent samples that are not copy number normal from being used to build the Best Practices Copy Number Reference, a filtering step attempts to omit samples showing suspicious calls and a relabeling step adjusts copy number calls for samples with known copy number states. The remaining putative normal samples are used to generate the Best Practices Copy Number Reference.

Outlier tests

For each fixed region, outlier tests attempt to identify plates showing a distribution of normal copy number calls that is 1) extremely low, 2) inconsistent with the distribution of copy number calls of the other plates in the batch, and 3) inconsistent with the expected distribution in known HapMap populations. Samples from plates that are determined to be outliers for a fixed region are relabeled as 'NoCall' for that region.

The outlier tests identify possible errors in the putative normal samples by analyzing copy number calls for samples grouped by plate for each fixed region. It is recommended that customers randomize their samples across plates (see "Recommendations" on page 15). This is essential for the assumptions made by the three outlier tests.

Minimum plate normal copy number rate

If the number of copy number normal samples on a plate is extremely low, it is likely due to a clustering error made by the CNVmix algorithm. For each fixed region, the percentage of samples on a plate that are labeled copy number normal is calculated. If the percentage is below a minimum rate, samples are relabeled as 'NoCall' for that region so that they are not included in the Best Practices Copy Number Reference.

The minimum rate threshold was determined for each fixed region based on the expected frequency of normal copy number states in HapMap populations. These expected frequencies were derived from a large Thermo Fisher Scientific data set of samples with known copy number spanning 5 HapMap populations (AFR, AMR, EAS, EUR, SAS). For each fixed region, the population showing the lowest frequency of normal copy number samples was determined. The minimum threshold is approximately the lowest frequency divided by two. The minimum thresholds are defined for each region in the `cn_region_refsettings.txt` library file.

Mean copy number outlier test

If the distribution of copy number calls on a plate is very different from the distribution of copy number calls across most plates in the batch, the results may need to be adjusted.

Only samples that pass the minimum plate normal copy number threshold (described above) are tested. For each fixed region on each plate, the mean copy number across samples is calculated. The grand median and quartiles of the plate means are calculated. A plate is flagged if the plate mean is an outlier based on both quartiles and the grand median. The thresholds *iqr-coefficient* and *distance-from-grand-median* are defined for each region in the `cn_region_refsettings.txt` library file.

Fisher's exact test

If the distribution of copy number calls on a plate is very different from the distribution of copy number calls in known HapMap populations, the results may need to be adjusted.

Only samples that pass the minimum plate normal copy number threshold (described above) are tested. For each fixed region and each HapMap population included in the *cn_region_refcalls.txt* library file, Fisher's Exact test is performed, testing the null hypothesis that the distribution of copy number states observed on a plate is similar to the distribution of copy number states expected in the HapMap population. Only plates showing extreme deviations from expected copy number state distributions in all populations are flagged (according to p-value threshold alpha defined for each region in the cn_region_refsettings.txt library file.). The populations tested can be specified as a comma-delimited list with flag `--fisher-populations` when using the APT reference generation workflow.

The outlier tests described above are used for both selecting copy number normal samples for reference creation and for testing the reference (See Plate Inlier Rate in Chapter 3).

Relabeling known samples

Any known samples included in the customer data are identified. Many HapMap and 1000 genomes project samples can be identified according to their genotype (See *Chapter 5 in Genotyping Solutions Data Analysis Guide*). If any samples are found, for each fixed region, any discordant copy number calls are relabeled with the correct copy number state.

Selecting copy number normal samples

The final set of copy number normal samples used to generate the Best Practices Copy Number Reference are chosen by combining all the methods described above.

For each fixed region

1. Flag plates with too few normal copy number samples according to the minimum plate normal copy number rate.
2. Restrict to plates that are not flagged:
    a. Perform Mean Outlier Test.
    b. Perform Fisher's Exact Test.
    c. Plate is flagged if outlier by BOTH tests in (a) and (b).
3. Output cnregionnormalcalls.txt with columns cel_files CN_Region and CN_State.
    - cel_files associated with plates flagged by (1) or (2) will have CN_State = NoCall
    - cel_files corresponding to known samples identified by signitureSNP will have CN_State = Known CN_State
    - All remaining cel_files will have CN_State = original CNVmix call
    - Finally, if CN_State is not the normal copy number state, then CN_State = NoCall

Note: The rules are implemented in the apt-copynumber-cnvmix-util engine and used by the Initial Copy Number Reference Creation and Best Practices Copy Number Reference Creation workflows.

6 Fixed region copy number analysis

Overview

Fixed Region analysis is used when breakpoints or endpoints of copy number regions of interest are known from publications or prior work. To perform such analysis, fixed regions of interest must be included in the array design and array library files.

Fixed Region analysis uses a novel algorithm called CNVmix which clusters the calculated median log2ratios for each region and assigns copy number states to each sample. The clustering is done separately for each plate.

For details on individual APT programs, refer to: https://www.affymetrix.com/support/developer/powertools/changelog/index.html

Fixed region copy number analysis in AxAS

The Analysis Files Folder must contain all required library files including a reference **.cn_models** file. Follow the steps below:

1. Click the **Array Type** drop-down to select the appropriate array type.
2. Import your CEL files.
3. From the Workflow drop-down, select **Copy Number Fixed Regions**.
4. Click the **Copy Number Discovery** drop-down button.
5. Click the **CN Reference File** drop-down to select the appropriate CN Reference file.
6. In the **Name** field (lower right) enter a batch name, then click **Run**.

For more information, see *Axiom Analysis Suite User Guide* (*Chapter 5*). Details on viewing results and exporting Copy Number Fixed Regions batches can also be found in Chapter 5 of the *Axiom Analysis Suite User Guide*.

Fixed region copy number analysis using APT

The library file package for the array is assumed to be downloaded into an accessible directory in the system. This directory must contain all required library files including a reference cn_models file. If the reference cn_models file does not exist, refer to Chapters 3 and 4 of this Guide on how to generate one.

The example scripts below are directly usable by Linux users. For readability, the Linux commands are broken into several lines with the backslash character: "\". When executing these commands in Windows, the backslash character should be removed, the input folder path should be specified using Windows conventions, and the entire command should be put in one line.

Where command line options point to directories such as the input, output, or analysis directories, the full path to the appropriate directory should be specified. For example, `<analysis_directory>` is the full path to the library file package directory for the array. Input and output directory names are specific to a particular analysis run. Input directory names in the scripts below may point to directories created in earlier steps.

Where command line options point to library files, their typical extensions are shown in the scripts. Library file names typically include the name of the array. They may also include additional tags such na<#> or r<#> where # is a revision number.

Where command line options point to input or output files, an example filename is provided.

The input is a list of CEL files (cel_list.txt). This text file has one column with the header 'cel_files' and a list of all input CEL files with full paths specified. All XML files that are passed using the `--arg-file` option must exist in the analysis directory.

Sample and plate genotyping QC

All samples used for CNV analysis must pass genotyping QC checks.

Calculate DishQC by running:

```
apt-geno-qc \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-geno-qc.AxiomQC1.xml \
--cel-files <input_directory>/cel_list.txt \
--out-dir <genoqc_directory> \
--log-file <genoqc_directory>/ apt-geno-qc.log
```

The column 'axiom_dishQC_DQC' in the output apt-geno-qc.txt has DishQC values. Remove samples that fail Dish QC and create cel_list2.txt for input in the next step.

Calculate QC Call Rate by running:

```
apt-genotype-axiom \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/
arrayname.96orMore_Step1.r<>.apt-genotype-
axiom.SnpSpecificPriors.AxiomGT1.apt2.xml \
--cel-files <input_directory>/cel_list2.txt \
--out-dir <qccr_directory> \
--log-file <qccr_directory>/apt-genotype-axiom.log
```

The column 'call_rate' in the output AxiomGT1.report.txt has the QC Call Rate values. Remove samples that fail QC Call Rate criteria and create cel_list3.txt for input in the next step.

For more details, see *Chapter 8 in Axiom Genotyping Solutions Data Analysis Guide*.

Probeset Summarization

Only samples that pass genotyping QC must be used (cel_list3.txt).

```
apt-genotype-axiom \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-genotype-axiom.AxiomCN_PS1.apt2.xml \
--out-dir <summarization_directory> \
--log-file <summarization_directory>/apt-genotype-axiom.log \
--cel-files <input_directory>/cel_list3.txt
```

Run CNVmix

The reference file must be specified; it is typically found in the analysis files directory. Probeset summarization files - summary.a5 and report.txt - are also required.

```
apt-copynumber-axiom-cnvmix \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-axiom-cnvmix.AxiomCNVmix.apt2.xml \
--summary-file <summarization_directory>/AxiomGT1.summary.a5 \
--report-file <summarization_directory>/AxiomGT1.report.txt \
--reference-file <analysis_directory>/arrayname.r<>.cn_models \
--out-dir <cnvmixresults_directory> \
--log-file  <cnvmixresults_directory>/apt-copynumber-axiom-cnvmix.log
```

The command shown above assumes that the summary file is in a5 format. If using AxiomGT1.summary.txt file, include:

```
--use-text-format-summary-file TRUE
```

The a5 format is set by default and recommended for the input summary file, because it is faster than the txt format especially for larger datasets. The AxiomGT1.summary.txt file can be converted to a5 format by using apt2-a5-table-converter.

Outputs

The output files of Fixed Region CNV analysis are:

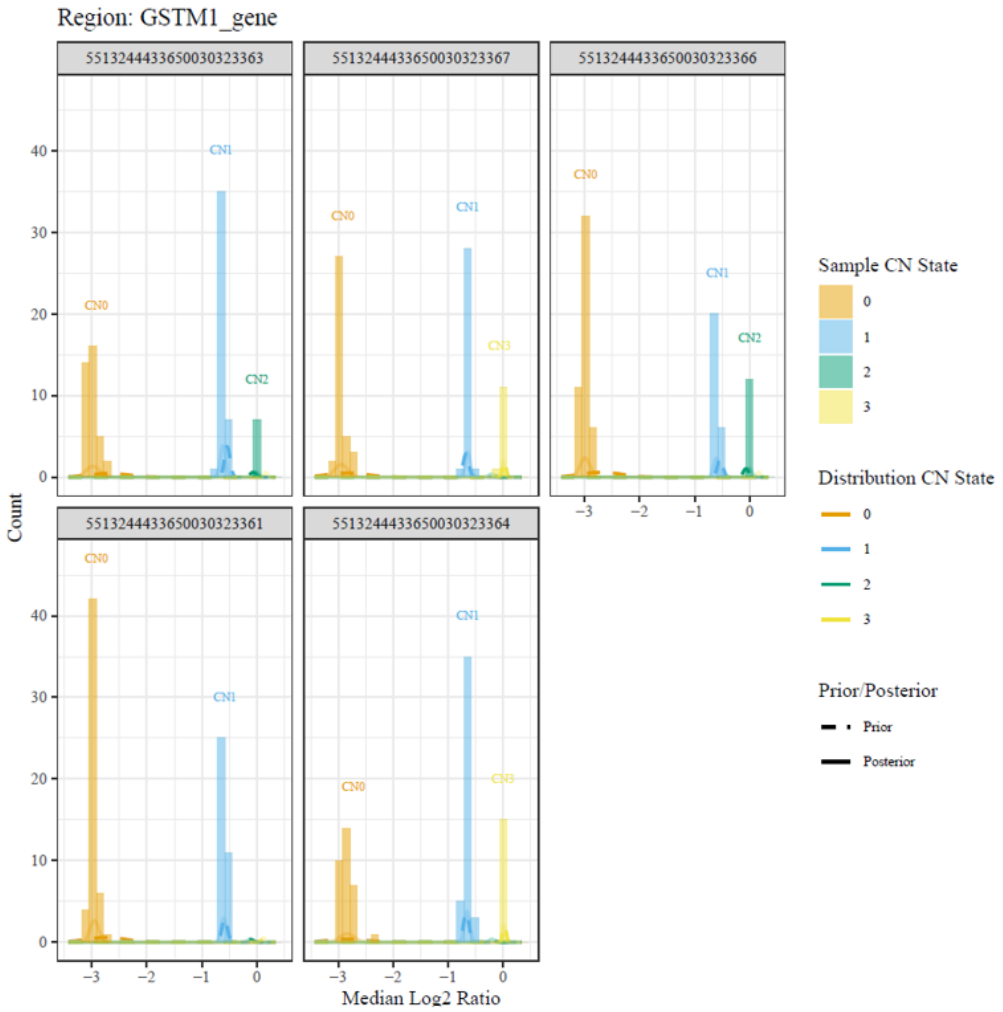
- report.txt file with copy number QC metrics and QC passing status for all samples
- cnregions.summary.txt file with a summary of the copy number calls for all the fixed regions of interest
- cnregioncalls.txt file with copy number calls for each sample for each region. This also includes a confidence score, median log2ratio, and plate bar code

- cnregions.details.txt file with details on CNVmix mixture model fitting and cluster labeling. This includes observed means, standard deviations, plate shifts, bin counts, probabilities, and other statistics.
- cn_posteriors.txt file in same format as the cn_priors file
- cnpscalls.txt that is used only when copy number aware genotyping is enabled

Visualization using SNPolisher

CN Visualization is a function in SNPolisher (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0017790_SNPolisher_UG.pdf) that generates plots from output files of fixed copy number analysis. Plots display the genotype calls, prior and posterior information, and the region details. (Figure 33) If the true copy number state is known, then a text file with true states may also be included as input.

Figure 33 Histograms of median log2 ratios for a Fixed Region called GSTM1_gene on five 96-format Axiom plates.



7 Discovery copy number analysis

Overview

Discovery copy number analysis is used when breakpoints or endpoints of copy number regions are not known a priori. It uses a Hidden Markov Model (HMM) implementation to perform segmentation. Analysis is done on each sample independently.

For details on individual APT programs, refer to: https://www.affymetrix.com/support/developer/powertools/changelog/index.html

Discovery copy number analysis in AxAS

The Analysis Files Folder must contain all required library files including a reference **.cn_models** file. Follow the steps show below.

1. Click the **Array Type** drop-down to select the appropriate array type.
2. Import your CEL files.
3. From the Workflow drop-down, select **Copy Number Discovery**.
4. Click the Copy Number Discovery drop-down button.
5. Click the **CN Reference File** drop-down to select the appropriate CN Reference file.
6. In the **Name** field (lower right) enter a batch name, then click **Run**.

For more details, see the *Axiom Analysis Suite User Guide* (*Chapter 5*). Instructions on viewing and exporting results can also be found in Chapter 5 of the *Axiom Analysis Suite User Guide*.

Discovery copy number analysis in APT

The library file package for the array is assumed to be downloaded into an accessible directory in the system. This directory must contain all required library files including a reference cn_models file. If the reference cn_models file does not exist, refer to Chapters 3 and 4 of this Guide on how to generate one.

The example scripts below are directly usable by Linux users. For readability, the Linux commands are broken into several lines with the backslash character: "\". When executing these commands in Windows, the backslash character should be removed, the input folder path should be specified using Windows conventions, and the entire command should be put in one line.

Where command line options point to directories such as the input, output, or analysis directories, the full path to the appropriate directory should be specified. For example,

`<analysis_directory>` is the full path to the library file package directory for the array. Input and output directory names are specific to a particular analysis run. Input directory names in the scripts below may point to directories created in earlier steps.

Where command line options point to library files, their typical extensions are shown in the scripts. Library file names typically include the name of the array. They may also include additional tags such na<#> or r<#> where # is a revision number.

Where command line options point to input or output files, an example filename is provided.

The input is a list of CEL files (cel_list.txt). This text file has one column with the header 'cel_files' and a list of all input CEL files with full paths specified. All XML files that are passed using the `--arg-file` option must exist in the analysis directory.

Sample genotyping QC

All samples used for CNV analysis must pass genotyping QC checks.

Calculate DishQC by running:

```
apt-geno-qc \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-geno-qc.AxiomQC1.xml \
--cel-files <input_directory>/cel_list.txt \
--out-dir <genoqc_directory> \
--log-file <genoqc_directory>/ apt-geno-qc.log
```

The column 'axiom_dishQC_DQC' in the output apt-geno-qc.txt has DishQC values. Remove samples that fail Dish QC and create cel_list2.txt for input in the next step.

Calculate QC Call Rate by running:

```
apt-genotype-axiom \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/
arrayname.96orMore_Step1.r<>.apt-genotype-
axiom.SnpSpecificPriors.AxiomGT1.apt2.xml \
--cel-files <input_directory>/cel_list2.txt \
--out-dir <qccr_directory> \
--log-file <qccr_directory>/apt-genotype-axiom.log
```

The column 'call_rate' in the output AxiomGT1.report.txt has the QC Call Rate values. Remove samples that fail QC Call Rate criteria and create cel_list3.txt for input in the next step.

For more details, see *Chapter 8 in Axiom Genotyping Solutions Data Analysis Guide*.

Probeset Summarization and Genotyping

Only samples that pass genotyping QC must be used as input (cel_list3.txt).

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

```
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-genotype-axiom.AxiomCN_GT1.apt2.xml \
--out-dir <genotyping_directory> \
--log-file <genotyping_directory>/apt-genotype-axiom.log \
--cel-files <input_directory>/cel_list3.txt
```

Run HMM

The reference file must be specified; it is typically found in the analysis files directory. Probeset summarization files, summary.a5, and report.txt are also required.

```
apt-copynumber-axiom-hmm \
--analysis-files-path <analysis_directory> \
--arg-file <analysis_directory>/arrayname.r<>.apt-copynumber-axiom-cnvmix.AxiomHMM.apt2.xml \
--summary-file <genotyping_directory>/AxiomGT1.summary.a5 \
--report-file <genotyping_directory>/AxiomGT1.report.txt \
--reference-file <analysis_directory>/arrayname.r<>.cn_models \
--out-dir <hmm_directory> \
--log-file  <hmm_directory>/apt-copynumber-axiom-cnvmix.log
```

The command shown above assumes that the summary file is in a5 format. If using AxiomGT1.summary.txt file, include:

```
--use-text-format-summary-file TRUE
```

The a5 format is set by default and recommended for the input summary file, because it is faster than the txt format especially for larger datasets. The AxiomGT1.summary.txt file can be converted to a5 format by using apt2-a5-table-converter.

Filtering output copy number calls

Copy number calls may be filtered based on minimum segment length and minimum number of probesets in a segment to improve accuracy of calls. Segments that are longer and are based on more probesets are more likely to be true events. The argument XML file has a default set of filter values based on copy number states. These values may be overridden by command line options.

For APT 2.11.6 and later, filter options and recommended values are:

```
--seg-min-bases-CN-zero 25000 \
--seg-min-bases-CN-one 50000 \
--seg-min-bases-CN-two 100000 \
--seg-min-bases-CN-three 100000 \
--seg-min-bases-CN-fourormore 100000 \
--seg-min-probesets-CN-zero 25 \
--seg-min-probesets-CN-one 50 \
--seg-min-probesets-CN-two 100 \
```

```
--seg-min-probesets-CN-three 100 \
--seg-min-probesets-CN-fourormore 100
```

For earlier APT versions, filter options are:

```
--seg-min-bases-CN-zero
--seg-min-bases-CN-oneormore
--seg-min-probesets-CN-zero
--seg-min-probesets-CN-oneormore
```

Run HMM with custom regions file

By default, the HMM algorithm detects CNV segments in regions defined in the hmm_regions.txt file in the analysis directory. This tab delimited file has columns 'region', 'chromosome', 'start', 'stop' followed by several HMM specific parameters. The regions are usually chromosome arms 1p, 1q, 2p, etc. If required, users may override the default file by providing a custom regions file using the option:
`--regions-file <input_directory>/custom.hmm_regions.txt`

Run HMM with LOH analysis

Loss of Heterozygosity (LOH) analysis may be enabled by the array library file package. Probesets at markers with high minor allele frequencies are used to detect regions of LOH, and such probesets must be flagged in the reference template and reference files.

To perform LOH analysis, the calls.txt and confidences.txt files from a genotyping analysis are required as additional inputs as shown below:

```
--loh-calls-file <genotyping_directory>/AxiomGT1.calls.txt \
--loh-confidences-file <genotyping_directory>/
AxiomGT1.confidences.txt
```

Outputs

Outputs of `apt-copynumber-axiom-hmm` are:

- `report.txt` file with copy number QC metrics, QC passing status, and segment counts per copy number state for all samples
- `cnv.a5` file in HDF5 format containing multiple datasets with results from the Discovery data analysis. This file contains the following datasets:
  - `/cn_segment` includes copy number segmentation information for each CEL file.
  - `/cnv` includes summary tables for log2ratios, BAFs, copynumber and smooth signal for all samples.
  - `/loh_segment` includes LOH segmentation information for each CEL file, if LOH analysis was performed.

To export segments in VCF format from the cnv.a5 file, use:

```
apt-format-result \
```

```
--cn-region-calls-file <hmm_directory>/AxiomHMM.cnv.a5 \
--annotation-file <analysis_directory>/
arrayname.na<>.r<>.annot.db \
--export-chr-shortname true \
--export-vcf-file <output_directory>/vcfsegments.vcf \
--log-file <output_directory>/apt-format-result.log
```

To restrict the VCF export to segments of specific CEL files, include:

```
--sample-filter-file <input_directory>/cel_files.txt
```

The `cnv.a5` file can be viewed using HDFView (https://www.hdfgroup.org).

The `cnv.a5` file may also be unpacked into multiple text files corresponding to its datasets using apt2-dset-util

```
apt2-dset-util --input-file AxiomHMM.cnv.a5 --output-type txt
```

Visualization in Axiom Analysis Suite

Folders for use with Axiom Analysis Suite can be created from the output folders of copy number and genotyping analysis performed in APT using:

```
apt-package-util \
--copynumber-data-dir <hmm_directory> \
--batch-folder <AxAS_batch_folder>
```

Whole Genome View in Axiom Analysis Suite may be used to visualize log2 ratio, BAF, copy number and smooth signal tracks.

Visualization using Integrative Genomics Viewer (IGV)

Discovery copy number analysis results may be imported to the Integrative Genomics Viewer (https://software.broadinstitute.org/software/igv/) for visualization. To format BAF, log2 ratio and smooth signal results for import to IGV, use apt-format-result.

```
apt-format-result
--run-format-igv True
--igv-cndata-file <cnv.a5 file>
--igv-export-cnv-baf True
--igv-export-cnv-log2ratio True
--igv-export-cnv-smooth-signal True
--igv-out-dir <output dir location name>
```