Axiom copy number analysis

Introduction

Applied Biosystems™ Axiom™ custom and catalog genotyping arrays offer innovative capabilities in genotyping applications for biobanking studies, population-optimized analyses, and pharmacogenomics. Copy number variations (CNVs), or alterations in the number of copies of a stretch of a genome, are important in several areas of human disease research. CNVs can manifest themselves as copy number (CN) gains or losses, and have been associated with increased risk for human diseases, physical abnormalities, psychiatric conditions, and childhood developmental disorders [1]. CNVs have also been implicated in diseases and traits in nonhuman species such as canine, bovine, porcine, poultry, and several diploid and polyploid plants. Therefore, in addition to genotyping single nucleotide polymorphisms (SNPs) and insertions or deletions (indels), Axiom arrays are designed to detect CN changes and allelic imbalances such as loss of heterozygosity (LOH). Axiom arrays can be used for targeted CNV applications or whole-genome CN discovery applications in diploid species including humans, animals, and plants.

In previous years, data analyzed with Axiom arrays have been used with Applied Biosystems™ Array Power Tools (APT)* or Axiom™ CNV Summary Tool Software and third-party applications such as PennCNV or Nexus software by BioDiscovery to calculate the $\log_2$ ratio and B allele frequency (BAF) for identifying CNVs. For example, the Applied Biosystems™ UK Biobank Axiom™ Array has been used to identify CNVs that were associated with adverse health outcomes [1]. CNVs have been identified with Axiom arrays in various other studies, including the Canadian longitudinal study, the Korea biobank study, and the Taiwan biobank study for understanding relationships between CNVs and specific human traits. CNV analysis has also been performed in chickens using the Applied Biosystems™ Axiom™ Chicken Genotyping Array [2].

The launch of Applied Biosystems™ Axiom™ Analysis Suite 4.0 now offers the capability to perform CNV analysis in regions with predefined boundaries and whole-genome CN discovery applications. All visualization is performed using the Integrative Genomics Viewer (**https://software.broadinstitute.org/software/igv/**) [3,4]. This technical note provides the background information on CNV calling within Axiom Analysis Suite and the corresponding quality control (QC) metrics for fixed-region CN analysis, whole-genome *de novo* CN analysis, and LOH analysis.

Basics of CNV analysis

The $\log_2$ ratio and BAF, which form the basis for CNV calling, are now calculated directly within the Axiom Analysis Suite application and used with the workflows established for CN analysis. The $\log_2$ ratio is the $\log_2$ of the ratio of signal intensity of a probeset** to reference total intensity for the same probeset. The reference total intensity is an estimate of the total A and B allele intensities for the probeset, representing the normal diploid state at that location. Probesets used for CN calculations include those at polymorphic and nonpolymorphic markers. Probeset selection for CN analysis may be based on sequence homology and signal response to CN changes.

* Array Power Tools was previously called Affymetrix Power Tools.

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

Calculation of the $\log_2$ ratio and subsequent CNV analysis is a function of the reference total intensity for the probeset, so it is important to get an accurate reference value. This reference value is calculated by taking the median total intensity for that probeset in a set of selected reference samples. The reference set, therefore, should represent the normal CN state for each probeset. The reference set may be created using one of the following approaches:

- **Plate or self reference:** One approach is to create the reference based on all samples genotyped on the plate, provided that for each probeset, the vast majority of individual samples on the plate are expected to have the normal CN state. This reference may then be applied to all samples on the plate.
- **Universal reference:** Another approach is to create a separate reference based on a selected set of normal CN samples that have been genotyped on one or more sets of plates. The universal reference is then applied to all subsequent plates. The number of samples used for such a reference should be as large as possible to account for all variations in experimental conditions. The analysis may be carried out with any number of samples but will be less accurate for smaller reference sets.
- **Regions with common CN variations:** Generating a reference in genomic regions where variations in CN are common within or across populations poses a special challenge. An example is the *GSTM1* gene: the majority of individuals in a population may not be diploid. In such a region, the reference for probesets should be generated from a set of samples that are carefully selected and known to be diploid in the region. An array may have multiple such regions. So, the final reference for all probesets on the array should be assembled from several such individual reference builds. This is possible using a special workflow in Axiom Analysis Suite or with appropriate inputs to the APT reference generation tool.

$\log_2$ ratios are negative for losses and positive for gains (e.g., a CN loss is identified as $\log_2 (1/2) = –1.0$ while a CN gain is identified as $\log_2 (3/2) = 0.59$). Differences in $\log_2$ ratios between consecutive CN states become smaller with higher CN states, such as CN4 and CN5. Empirical $\log_2$ ratios based on measured intensities show smaller differences between consecutive CN states (i.e., become more compressed). For this reason, it can be difficult to distinguish between high consecutive CN gain states. An example of CN2 and CN3 states is shown in Figure 1.

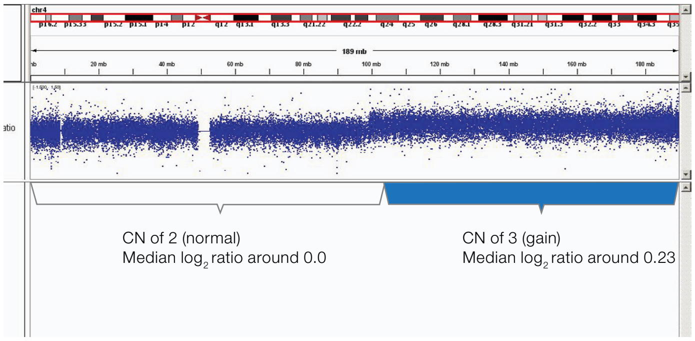


Figure 1. The $\log_2$ ratio track in IGV for a sample shows a large gain on chromosome 4.

Measures of heterozygosity and LOH can be used to identify mosaic sex chromosome aneuploidies, situations of consanguinity, or specific traits such as Turner syndrome [9]. The BAF is one such measure of the heterozygosity at any location and is calculated for each probeset: Raw BAF = (signal intensity of B allele)/(signal intensity of A allele + signal intensity of B allele). A complete deletion of one copy is indicated by a BAF of 0 or 1 (no middle-band BAF at 0.5) and with a $\log_2$ ratio of less than 0.

In Axiom Analysis Suite or APT, final BAF values for a probeset are scaled and normalized based on nominal values for different genotypes. An example for a CN2 and a CN3 region is shown in Figure 2.

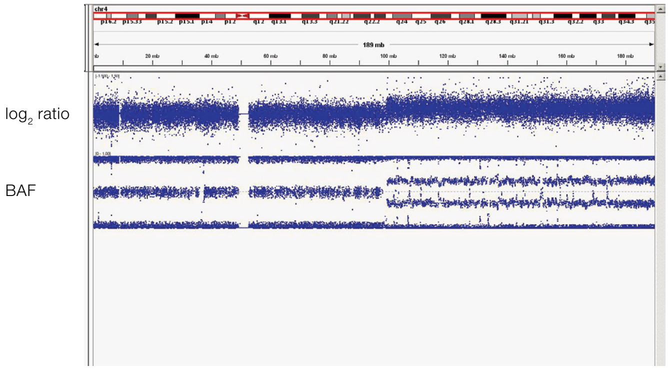


Figure 2. BAF and $\log_2$ ratio tracks in IGV for the same sample as in Figure 1. The BAF track in the normal CN2 region shows the expected three bands (AA, AB, BB genotypes) while the CN3 region shows four bands (AAA, AAB, ABB, BBB genotypes).

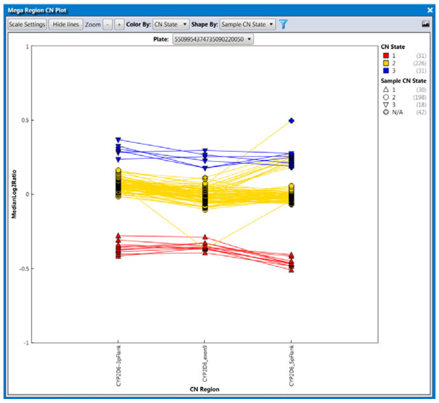
CN analysis

Axiom Analysis Suite and APT use the $\log_2$ ratio and BAF to make a determination of CN gain or loss. This can include fixed-region CN analysis or whole-genome *de novo* analysis for CN discovery.

Fixed-region analysis

Fixed-region analysis is one of two methods for performing CN analysis in Axiom Analysis Suite or APT. This method is used when the breakpoints of CN regions of interest are known from publications or prior work. Fixed-region CN analysis requires designing probesets for a specific gene or a specific region within the genome (e.g., a CN analysis of specific exons of the human *DMD* gene is achieved by designing multiple probesets across the *DMD* gene). By saturating the targeted region with multiple probesets, one is able to get higher resolution of the CN changes within the region. Probesets used for CN analysis can interrogate nonpolymorphic markers or SNPs in the region. Probesets at nonpolymorphic markers are typically ignored during the genotyping process by not including them in the Step2 probeset list [10] by default.

In fixed-region CN analysis, probesets are selected and assigned to fixed CN regions of interest at the time of array design. This probeset mapping information is stored in library files, and median $\log_2$ ratios are calculated during analysis for all selected probesets for each region for each sample. A novel clustering algorithm then performs a multisample analysis of the data for each region and assigns CN states to each sample. Model parameters and state priors such as mean and standard deviation of the median $\log_2$ ratio for desired CN states are required as inputs for model fitting and analysis. Example results are shown in Figure 3.



Mega Region CN Plot

Color By: CN State

Shape By: Sample CN State

Plate: 5509954374735090220050

Figure 3. CN changes for 3 regions of *CYP2D6* across 96 samples analyzed with the Applied Biosystems™ Axiom™ Precision Medicine Diversity Array (Axiom PMDA). Results are visualized with the Mega Region CN Plot, a built-in Axiom Analysis Suite capability.

Whole-genome *de novo* analysis

Discovery or *de novo* analysis is another method for performing CN analysis in Axiom Analysis Suite or APT. This method is used when CN breakpoints are not known and must be determined. CN states are determined by implementation of a hidden Markov model (HMM). Breakpoints are discovered and CN segments are labeled by states. Regions of interest are specified as inputs based on intent at the time of array design. State transition probabilities, model parameters, and state priors must also be specified as inputs. Visualization of results from a discovery analysis in IGV is shown in Figure 4.

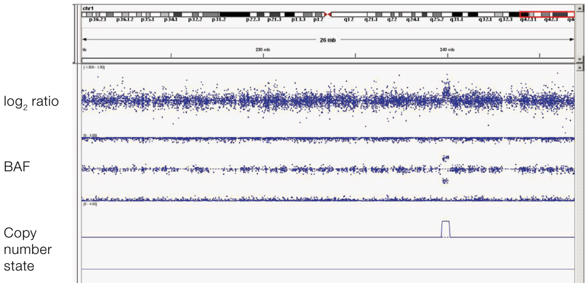


Figure 4. Example of a CN gain on chromosome 1 of a sample as detected with an HMM. Also shown are the corresponding $\log_2$ ratio and BAF tracks in IGV.

Loss or absence of heterozygosity

LOH regions are those where SNPs do not display heterozygosity (Figure 5). Probesets at markers with high minor allele frequencies are used to detect regions of LOH. Such regions may also indicate absence rather than LOH in agrigenomics applications (e.g., inbreeding).

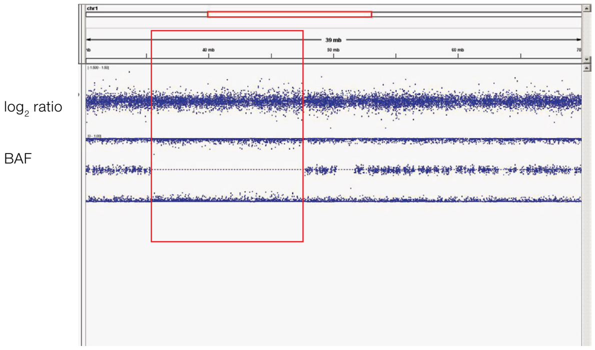


Figure 5. Example of LOH in a canine sample. LOH is detected in a ~20 Mb genomic region.

QC metrics in CNV analysis

CNV analysis in Axiom Analysis Suite offers QC of the results. The QC metrics include median of the absolute values of all pairwise differences (MAPD) and waviness-SD metrics.

MAPD

The MAPD metric is the median of the absolute values of all pairwise differences between $\log_2$ ratios for a given sample. It is a global measure of the variation observed in probeset signal intensities distributed across the genome. Any two markers that are adjacent in the genomic coordinates are a pair. MAPD is robust against high biological variability in CN states, and hence, $\log_2$ ratios.

Samples with high MAPD, indicative of high variability in $\log_2$ ratios between adjacent probes, are excluded from the analysis because increased variability decreases the quality of CN calls. Variability in $\log_2$ ratios in a chip arises from two distinct sources: (1) intrinsic variability in the DNA material, assay preparation, and instrumentation, and (2) apparent variability induced by the fact that the reference total intensity may have systematic differences from this array.

Variability in general will be reduced by using a reference set generated from arrays run in the same laboratory. As in genotyping, there can be lab-to-lab systematic effects. If a reference is generated from arrays run in another lab, such systematic differences may inflate apparent variability. If MAPD metrics are consistently high when using an external reference, it is recommended to recalculate MAPD with an intralab reference. If the MAPD metrics drop significantly, then the high MAPD is an artifact introduced by a systematic difference between current samples and the samples that made up the reference, rather than it being a quality issue. Examples of $\log_2$ ratio tracks for different MAPD values are shown in Figure 6.

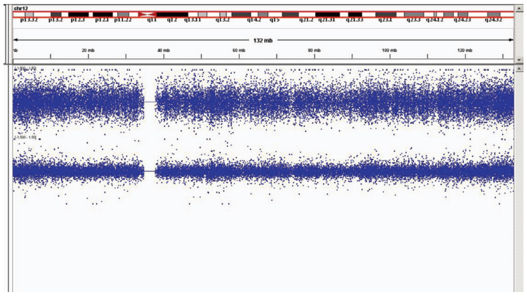


Figure 6. The $\log_2$ ratio tracks for 2 samples with different MAPD values. The MAPD value is 0.36 for the sample shown in the upper track and 0.15 for the sample shown in the lower track.

Waviness-SD

The waviness-SD metric is a global measure of variation of probesets that is insensitive to short-range variation and focuses on long-range variation. High waviness-SD usually indicates too much noise in the data and implies either sample or processing effects that will reduce the quality of the CN calls. However, elevated waviness-SD with good MAPD can also occur in samples with many CN changes (e.g., cancer samples) or very large regions of change. In such situations, it is important to inspect the data more carefully.

Verification results from CN analysis algorithms

Algorithms were experimentally verified using Axiom arrays for *de novo* and fixed-region CN analysis.

De novo analysis

Algorithm verification was performed using data from the UK Biobank Axiom Array (Cat. No. 902502) [5,6]. Whole-genome CN analysis was performed using 3 plates of samples from the International HapMap Project, a plate reference, and the default HMM parameters and priors. CN calls were compared to calls on the same samples from Applied Biosystems™ CytoScan™ HD arrays. All samples used for performance evaluation passed QC metrics. For CN detection using Axiom arrays, CN0 events must have a minimum of 25 probesets, and CN1 and CN3+ events must be at least 50 kb with a minimum of 50 probesets. In individual samples with fewer than 3 events of the same kind, sample performance metrics were not calculated.

CN calls in hypervariable regions were removed (Table 1). Hypervariable regions are genomic regions that show CN changes across many samples. They were determined empirically.

Table 1. List of hypervariable regions on the UK Biobank Axiom Array (in the hg19 reference genome) that were excluded from analysis.

Chromosome	Start	End
1	248,681,754	248,835,053
5	180,376,952	180,432,918
6	29,850,000	29,940,000
7	38,273,345	38,419,181
8	39,226,075	39,390,890
11	55,347,529	55,481,854
14	22,329,745	23,005,312
14	106,000,000	108,000,000
17	44,107,114	44,854,730
19	55,240,000	55,370,000
22	22,680,000	23,260,000

There was only one known CN0 event in the data set, and this event was detected. There were 79 known CN1 events in the verification data set, of which 72 events were detected. One additional previously unknown CN1 event was also detected. There were 173 known CN3 events in the verification data set, of which 154 events were detected. Twenty-eight additional previously unknown CN3 events were also detected.

Overall, CN1 events were detected with sensitivity >90% and positive predictive value >80%. CN3 events were detected with sensitivity >80% and positive predictive value >70%.

Fixed-region CN analysis

Algorithm verification was performed using theApplied Biosystems™ Axiom™ Transplant Genotyping Array (Cat. No. 902865) [7,8]. Fixed-region analysis of ~2,000 regions of interest was performed using 3 plates of samples, a plate reference, and the default fixed-region algorithm parameters and priors. CN calls were compared to call results from a prior analysis of the same data set using a different algorithm.

All samples used for performance evaluation passed QC metrics. To be detected on an Axiom array, CN0 events must have at least 25 probesets, CN1 events must be at least 50 kb with at least 50 probesets, and CN3+ events must be at least 50 kb with at least 50 probesets. In individual samples with fewer than 3 events of the same kind, sample performance metrics were not calculated.

There were 7 known CN0 events in this verification data set, all of which were detected. One additional CN0 event was also detected. There were 57 known CN1 events in this verification data set, of which 55 events were detected. Four additional CN1 events were detected. There were 68 known CN3 events in this verification data set, of which 67 events were detected. Twenty-one additional CN3 events were also detected.

Overall, CN1 events were detected with sensitivity >90% and positive predictive value >80%. CN3 events were detected with sensitivity >80% and positive predictive value >70%.

References

1. Kendall KM, Rees E, Escott-Price V et al. (2017) Cognitive performance among carriers of pathogenic copy number variants: analysis of 152,000 UK Biobank subjects. *Biol Psychiatry* 82:103–110.
2. Strillacci MG, Cozzi MC, Gorla E et al. (2017) Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. *Animal* 11:737–745.
3. Robinson JT, Thorvaldsdóttir H, Winckler W et al. (2011) Integrative Genomics Viewer. *Nat Biotechnol* 29:24–26.
4. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief in Bioinform* 14:178-192.
5. UK Biobank Axiom Array. Data sheet: **https://assets.thermofisher.com/TFS-Assets/LSG/brochures/uk_axiom_biobank_genotyping_arrays_datasheet.pdf**
6. Bycroft C, Freeman C, Petkova D et al. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562:203–209.
7. Li YR, van Setten J, Verma SS et al. (2015) Concept and design of a genome-wide association genotyping array tailored for transplantation-specific studies. *Genome Med* 7:90.
8. Axiom Transplant Genotyping Array. Data sheet: **https://assets.thermofisher.com/TFS-Assets/LSG/brochures/axiom_tx_genotyping_array_datasheet.pdf**
9. Shankar RK, Backeljauw PF (2018) Current best practice in the management of Turner syndrome. *Ther Adv Endocrinol Metab* 9:33–40.
10. Axiom Genotyping Solution. Data analysis guide: **https://assets.thermofisher.com/TFS-Assets/LSG/manuals/axiom_genotyping_solution_analysis_guide.pdf**

Find out more at **thermofisher.com/microarrays**