Custom cluster file creation for improved copy number analysis

Improve the performance of cytogenomic analysis by creating a custom cluster file that captures variation from site-specific factors

illumına®

# Copy number analysis from genotyping arrays

High-density Infinium™ BeadChips, such as the Infinium CytoSNP-850K BeadChip, Infinium Global Screening Array with Cytogenetics-24, and Infinium Global Diversity Array with Cytogenetics-8, enable high-resolution copy number analysis and the potential for discovery of meaningful cytogenetic aberrations. Copy number analysis with the Infinium assay is based on the normalized intensity values LogR Ratio (LRR) and normalized B allele frequency (BAF) of associated genotypes. A standard canonical cluster position is used to compute both LRR and BAF information from each locus.

The standard cluster file (*.egt file) that is supplied by Illumina for each Infinium BeadChip type is generated using a diverse set of more than 200 HapMap[1] DNA samples in an Illumina laboratory. This cluster file is expected to yield the specified performance metrics of the BeadChip. However, because all calculations for LRR and BAF are made by comparing the experimental data to a canonical genotype cluster, individual laboratories can obtain optimal copy number performance by implementing a custom-generated cluster file that captures local experimental conditions (Figure 1).

By following Infinium assay best practices and implementing a custom-generated cluster file, it is possible to reduce the LogR Deviation metric and the number of spurious region calls and increase the accuracy of the results of experimental copy number analysis. Therefore, customers using Infinium BeadChips should consider generating and implementing use of a custom cluster file. Custom cluster file creation can be performed at the beginning of a study or retroactively applied to previously processed BeadChips. This technical note provides guidelines and a detailed workflow for generating a custom cluster file that captures the unique experimental conditions of a laboratory site.

🔗 For a detailed description of LRR and BAF calculations, read the Interpreting Infinium Assay Data for Whole-Genome Structural Variation technical note.

🔗 For more information about Infinium Best Practices, including essential equipment and operating procedures, read the Infinium Lab Setup and Best Practices guide.
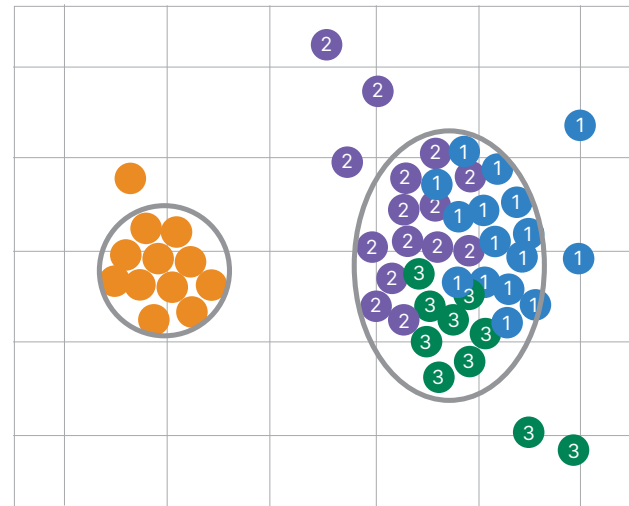
A. No experimental variation captured   B. Normal experimental variation captured

Figure 1: Representation of how custom cluster files should capture variation associated with experimental or site-specific conditions—Laboratories can generate custom cluster files with their own reference samples to capture common, site-specific variation. The custom cluster file can help labs generate enhanced copy number profiles as shown. (A) No variation captured. The data points and cluster on the left (orange) are from a single processing run that does not capture normal experimental variation. (B) Normal variation captured. A larger number of data points, derived from multiple processing runs, indicated here with different colored data points numbered 1–3, more accurately reflects the normal variation specific to a lab. Capturing normal variation will lead to a more robust cluster file that will be applicable for a longer period.

Custom cluster files can improve copy number analysis

Site-specific factors (ie, sample processing, lots, operators, etc.) can vary and a custom cluster file can facilitate more accurate genotyping results or provide a more representative reference for LRR and BAF calculations in copy number analysis. The recommendations in this technical note are specific to cluster files used primarily for copy number analysis and differ slightly from standard genotyping arrays, as copy number analysis software loads LRR and BAF values regardless of whether single nucleotide polymorphisms (SNPs) have been zeroed.

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

Guidelines for custom cluster files

The following guidelines are recommended for creating a custom cluster file. Using these guidelines and the subsequent workflow (Figure 2) captures common variation, mitigating the need for creation of additional custom cluster files. The performance of a custom cluster file in copy number analysis should be monitored for significant changes over time. This data can inform when a new custom cluster file needs to be generated. Follow these general guidelines when generating the file:

- Use samples of comparable quality and quantity—samples should represent the sample type/source and population that will be studied and should be normalized to the same concentration

- Use at least 100 samples, roughly balanced between males and females—increasing the number of samples can increase the robustness of the custom cluster file

- Use unaffected samples (eg, nontumor)—samples should not have major/large chromosomal abnormalities (eg, deletions, duplications)

- Exclude samples with failing or poor-performing call rates—do not use runs with obvious outliers, processing errors, or known deviations

- To capture common variation, use data from:

  - A minimum of three runs

  - A minimum of three reagent lots, including user-supplied reagents

  - Runs from a representative number of operators

  - Runs from all automation instruments, if using automation

  - Runs with roughly even sampling of various conditions and sample types

Steps for creating a cluster file

Create a GenomeStudio™ project

1. Create a GenomeStudio project using the current manifest and cluster file available on the product support web page (see the GenomeStudio user guide for detailed instructions on project creation).
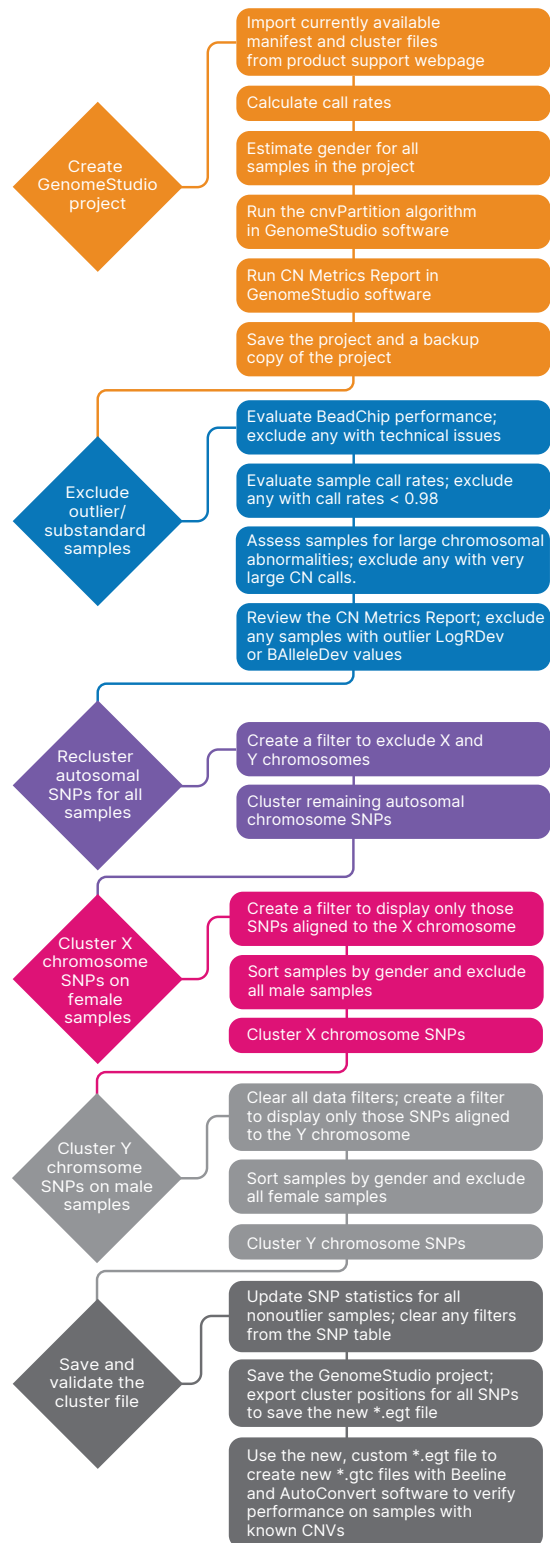

Figure 2: Custom cluster file generation workflow—Six main processes involved in creating a custom cluster file are shown with a summary of the steps in the workflow.

2. Calculate call rates for the imported samples from the samples table.

3. Select all samples, right click, and select "Estimate Gender for All Samples." If gender was not assigned to samples during project creation, select the option to populate the Gender Column.

4. Run the cnvPartition plug-in from the analysis tab.

5. Run the CN metrics report.

6. Save the project (saving a copy of the project is also recommended).

Sample inclusion/exclusion

1. Evaluate BeadChip performance on the controls dashboard and exclude any BeadChips that demonstrate a profile suggestive of a technical issue.

2. Evaluate sample call rates in the samples table and exclude samples with call rates < 0.98.

3. Assess samples for large chromosome abnormalities by selecting "Show CNV Region Display" from the analysis menu. From the display, review the imported samples and identify samples with large copy number calls and exclude these samples.

4. Review the CN Metrics report for any samples that have outlier LogRDev or BAlleleDev values and exclude those samples.

How to exclude samples

1. From the samples table, highlight samples to be excluded, right click, and select "Exclude Selected Samples."

2. Select "No" for "Do you wish to update SNP statistics for all SNPs?" in the popup.

3. Assign the excluded outlier samples an Aux value (for example, "0") to filter them out of subsequent steps

Recluster autosomal chromosome SNPs (excluding X and Y chromosomes) using all samples

1. From the SNP table, display only the autosomal SNPs by creating a filter (funnel icon) to exclude SNPs aligned to the X and Y chromosomes.

2. Return to the full data table and sort by "Chr" to ensure chromosome X and Y SNPs are not displayed.

    a. Select all displayed SNPs.

    b. Right click and select "Cluster Selected SNPs."

    c. Do not select the option to update SNP statistics.

Cluster X chromosome SNPs with female samples

1. Clear filters from the SNP and sample tables.

2. Filter to display only the SNPs aligned to Chromosome X.

3. In the Samples Table, sort by Gender and select all male samples.

    a. Right click and select "Exclude Selected Samples."

    b. Do not select the option to update SNP Statistics.

4. Return to the full data table and sort by "Chr" to ensure only chromosome X SNPs are displayed.

    a. Select all displayed SNPs.

    b. Right click and select "Cluster Selected SNPs."

    c. Do not select the option to update SNP statistics.

Cluster Y chromosome SNPs with male samples

1. Clear filters from the SNP Table.

2. Filter to display only the SNPs aligned to Chromosome Y.

3. In the Samples Table, include all male samples; be sure to keep the outlier samples excluded (Aux value of "0").

4. In the Samples Table, sort by Gender to select all female samples.

    a. Right click and select "Exclude Selected Samples."

    b. Do not select the option to Update SNP Statistics.

5. Return to the full data table and sort by "Chr" to ensure only chromosome Y SNPs are displayed.

    a. Select all displayed SNPs.

    b. Right click and select "Cluster Selected SNPs."

    c. Do not select the option to update SNP statistics.

6. Save the GenomeStudio project.

Additional editing or zeroing of SNPs (optional)

- BlueFuse™ Multi software and GenomeStudio software load LRR and BAF values regardless of whether SNPs have been zeroed.

- For copy number analysis applications in which genotypes are not being reported, additional zeroing of poorly performing SNPs after reclustering may not be warranted.

    When genotype analysis is also important, refer to the Infinium Genotyping Data Analysis technical note for detailed instructions on SNP evaluation and generating a custom cluster file.

Make sure that all nonoutlier samples are included in the samples table

1. Select all nonoutlier samples from the samples table.

2. Right click and select "Include Samples." Update SNP statistics when prompted.

3. Clear any filters from the SNP table.

4. Save the GenomeStudio project.

5. Export the cluster positions for all SNPs from the File menu to save the new *.egt file.

Validate the new cluster file

Use the new custom *.egt file to create *gtc files with Beeline™ and AutoConvert softwares to verify performance on samples with known copy number variations. For the best evaluation, the samples should not have been included within the GenomeStudio project to create the custom cluster file.

To read more about how Beeline software enables flexible filtering of array results, visit the Introduction to Beeline Software page

Summary

Illumina supplies a standard cluster file for each Infinium BeadChip type that is expected to yield the specified performance metrics of the BeadChip. Variation or drift in data over time may require updates to custom cluster files to maintain optimal performance. By following the guidelines in this technical note, users can create a custom cluster file to increase the accuracy of the results, providing optimal data for copy number analysis.

Learn more

Illumina training resources

GenomeStudio Genotyping: Creating Custom Cluster Files for Infinium Arrays webinar

References

1. International HapMap Consortium. The International HapMap Project. *Nature.* 2003;426(6968):789-796. doi:10.1038/nature02168

illumına®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com