Custom Cluster File Creation for Improved Copy Number Analysis

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

Introduction

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

The standard cluster file (*egt file) 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, optimal copy number performance may be achieved by implementing a custom-generated cluster file that captures the experimental conditions of the processing laboratory (Figure 1).

By following Infinium best practices and implementing a custom-generated cluster file, it is possible to reduce the LogR Deviation metric, reduce the number of spurious region calls, and increase the accuracy of the results to provide optimal data for copy number analysis. Therefore, customers using Infinium BeadChips, such as the Infinium CytoSNP-850K BeadChip for copy number analyses, should consider implementation of a custom cluster file. Custom cluster file creation can be performed at the beginning of a study or on previously processed BeadChips. This technical note provides guidelines and a detailed workflow for generating a custom cluster file to capture the unique experimental conditions of each 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 for an Infinium lab, read the Infinium Assay Lab Setup and Procedures Guide.

Custom Cluster Files Improve Copy Number Analysis

The goal of creating a custom cluster file is to capture the common variation (ie, processing, lots, operators, etc) within a given lab and/or sample type that is reflective of performance at that site (Figure 1). Site-specific factors can vary, and often a custom cluster file can result in 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 load LRR and BAF values regardless of whether SNPs have been zeroed.

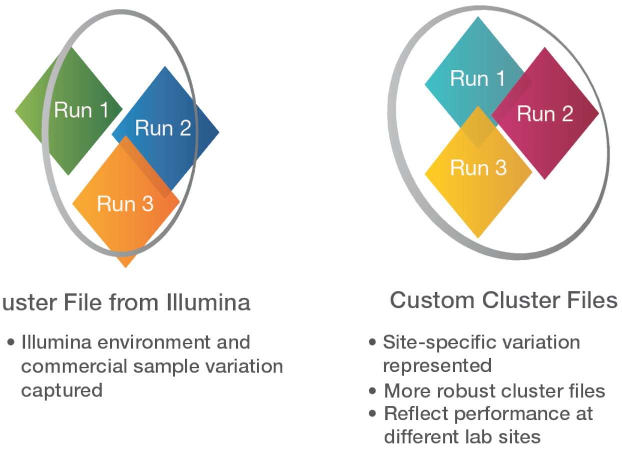

Run 1
Run 2
Run 3
Run 1
Run 2
Run 3
Cluster File from Illumina
• Illumina environment and commercial sample variation captured
Custom Cluster Files
• Site-specific variation represented
• More robust cluster files
• Reflect performance at different lab sites

Figure 1: Illustration of the Common Variation Captured with Custom Cluster Files—Illumina provides a standard cluster file for each Infinium BeadChip type that is used for calculating LRR and BAF values, which are then used in copy number analysis. Laboratory sites that generate custom cluster files can capture common, site-specific variation that can result in enhanced copy number profiles.

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

Guidelines for Creating a Custom Cluster File

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 frequent creation of additional custom cluster files. The performance of a custom cluster file in copy number analysis should be monitored for significant variation changes over time. This data can inform when a new custom cluster file needs to be generated.

- **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 lead to increased 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 reagents lots (include user-supplied reagents)

  Runs from a representative number of operators

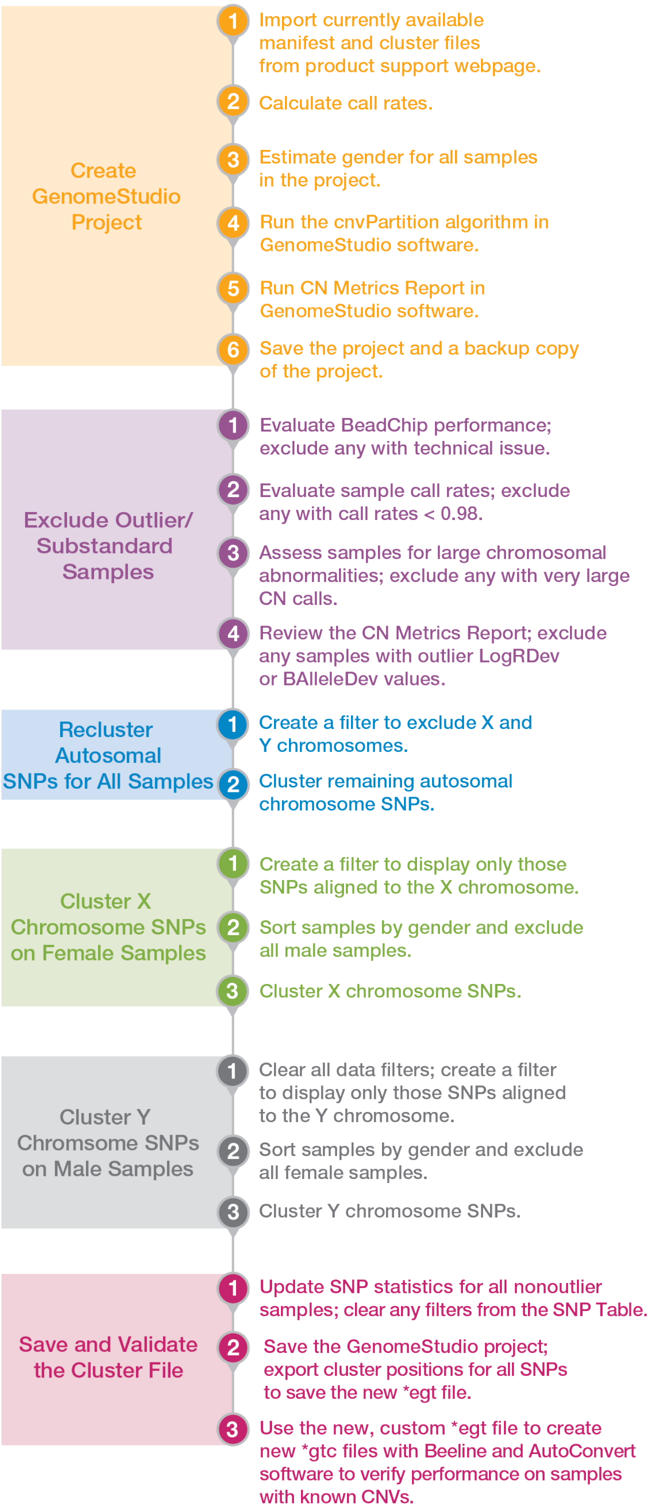
  Runs from all Tecans (if using automation)

  Runs with roughly even sampling of various conditions and sample types

Outline of Steps to Create a Cluster File

Create a GenomeStudio® project

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



Create GenomeStudio Project

1 Import currently available manifest and cluster files from product support webpage.

2 Calculate call rates.

3 Estimate gender for all samples in the project.

4 Run the cnvPartition algorithm in GenomeStudio software.

5 Run CN Metrics Report in GenomeStudio software.

6 Save the project and a backup copy of the project.

Exclude Outlier/ Substandard Samples

1 Evaluate BeadChip performance; exclude any with technical issue.

2 Evaluate sample call rates; exclude any with call rates < 0.98.

3 Assess samples for large chromosomal abnormalities; exclude any with very large CN calls.

4 Review the CN Metrics Report; exclude any samples with outlier LogRDev or BAlleleDev values.

Recluster Autosomal SNPs for All Samples

1 Create a filter to exclude X and Y chromosomes.

2 Cluster remaining autosomal chromosome SNPs.

Cluster X Chromosome SNPs on Female Samples

1 Create a filter to display only those SNPs aligned to the X chromosome.

2 Sort samples by gender and exclude all male samples.

3 Cluster X chromosome SNPs.

Cluster Y Chromsome SNPs on Male Samples

1 Clear all data filters; create a filter to display only those SNPs aligned to the Y chromosome.

2 Sort samples by gender and exclude all female samples.

3 Cluster Y chromosome SNPs.

Save and Validate the Cluster File

1 Update SNP statistics for all nonoutlier samples; clear any filters from the SNP Table.

2 Save the GenomeStudio project; export cluster positions for all SNPs to save the new *egt file.

3 Use the new, custom *egt file to create new *gtc files with Beeline and AutoConvert software to verify performance on samples with known CNVs.

Figure 2: Custom Cluster File Generation Workflow—The process for creating a custom cluster file includes six main steps with detailed instructions.

Sample inclusion/exclusion criteria

1. Evaluate BeadChip performance on the Controls Dashboard. 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 very 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' in the popup screen "Do you wish to update SNP statistics for all SNPs?".
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 Samples 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.

Save and validate the cluster file

Final steps for exporting cluster positions

Ensure 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 new *gtc files with Beeline™ software and AutoConvert software to verify its 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 Beeline software enables flexible filtering of array results, visit www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/beeline.html.

Summary

Illumina supplies a standard cluster file for each Infinium BeadChip type that 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, optimal copy number performance may be achieved by implementing a custom-generated cluster file that captures the experimental conditions of the processing laboratory. 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

To access a catalog of Illumina online training courses for arrays, visit support.illumina.com/training/online-courses/array.html.

To view an Illumina technical support webinar on the topic of "GenomeStudio Genotyping: Creating Custom Cluster Files for Infinum Assays," visit support.illumina.com/training/webinars/array/array-archive.html.

References

1. International HapMap Consortium. The International HapMap Project. *Nat.* 2003;426:789–796.

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

Pub. No. 770-2017-017-A