Infinium® Genotyping Data Analysis

A guide for analyzing Infinium genotyping data using the GenomeStudio® Genotyping Module

Introduction

Illumina DNA Analysis BeadChips using the Infinium Assay provide researchers unparalleled genomic access and accuracy for analyzing genetic variation. Illumina's portfolio of powerful products offers many benefits to researchers. In particular, BeadChips provide highly multiplexed sample analysis of 300,000 to more than 5,000,000 markers simultaneously. The density of arrayed features used for this high level of parallel analysis also allows for a high level of intrinsic assay redundancy. Each locus is assayed with a minimum 12- to 18-fold redundancy for each sample, yielding a highly robust intensity estimate that enables highly accurate genotype calls. Another strong contributor to data quality is outlier removal when averaging raw intensity data.

Each SNP is analyzed independently to identify genotypes. Genotypes are called by comparing customer-generated data with those in the supplied cluster file. Genotype calls are highly accurate and unambiguous for high-quality samples. Generally, high-quality data with 99.5% call rates can be expected. However, accuracy is highly sample dependent. When samples do not perform as expected, experimenters can choose to reprocess (requeue) these samples to confirm or potentially improve results. Poorly performing samples can be systematically excluded from the project before recalling genotypes. Additionally, some samples (e.g., samples that have been whole-genome amplified) may not fit standard cluster positions well. A custom-generated cluster file may improve the call rate and accuracy for these aberrant cases.

This technical note describes the method used by the Illumina FastTrack Genotyping Services Group for analyzing Infinium genotyping data with the Illumina GenomeStudio Genotyping Module to optimize call rates. Analysis begins with preliminary sample quality evaluation to determine which samples may require reprocessing or removal. If a custom cluster file is required, clustering should be done after removal of failed or suboptimal samples. After selected samples have been reprocessed and included in the analysis, sample quality can be evaluated in more detail. If necessary, individual locus analysis, where loci are identified by sorting on per-locus metrics such as call rate or cluster separation, can optimize data quality before generating a final report.

Note: For polyploid samples, the Polyploid Clustering Module is used rather than the standard Genotyping Module. The Polyploid Clustering Module identifies clusters in polyploid organisms like wheat and potato, for which the standard diploid clustering algorithm is not particularly efficient. This module performs cluster assignment but does not assign genotypes because the assignment of genotypes in polyploid species is highly dependent on the population and biology of the organism. Further information on using this module can be found in the Performing Clustering in the GenomeStudio Polyploid Clustering Module user guide[1].

GenCall Score Cutoff

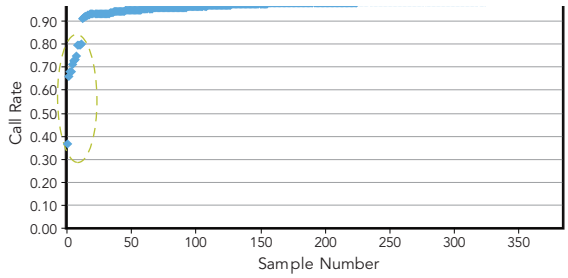
The first GenomeStudio parameter that should be optimized to obtain the highest genotyping accuracy is the GenCall score cutoff, or no-call threshold. The GenCall score is a quality metric calculated for each genotype (data point), and ranges from 0 to 1. GenCall scores generally decrease in value the further a sample is from the center of the cluster to which the data point is associated. The no-call threshold is the lower bound for calling genotypes relative to its associated cluster. Illumina FastTrack Genotyping Project Managers typically use a no-call threshold of 0.15 with Infinium data. This means that genotypes with a GenCall score less than 0.15 are not assigned genotypes because they are considered too far from the cluster centroid to make reliable genotype calls. They are instead assigned a "no call" for that locus. No calls on successful DNAs at successful loci contribute to lowering the call rate for the overall project. The standard 0.15 no-call threshold for Infinium data was determined using projects with trio and replicate information. A range of no-call thresholds were evaluated to optimize the call rate without compromising reproducibility or Mendelian consistency. The no-call threshold in GenomeStudio can be changed by selecting Tools | Options | Project | No-call Threshold. Sample and SNP statistics must be recalculated after an adjustment to the no-call threshold.

Identifying and Reprocessing Poorly Performing Samples

There are two options to ensure optimal analysis with samples that have not performed well. A sample may be reprocessed in an attempt to improve the results, or the sample can be excluded before reclustering. An initial assessment of DNA sample quality, as described here, can indicate which samples, if any, should be reprocessed. This procedure should be used after the Infinium Genotyping assay is complete and all BeadChips are imaged in a project.

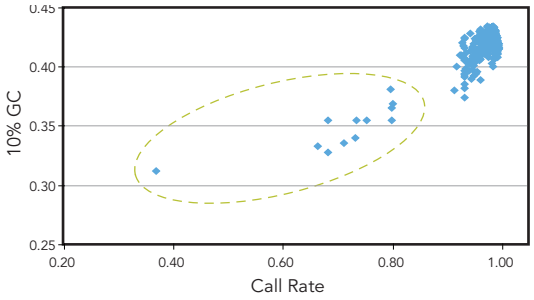
1. Open a new project in the GenomeStudio GT Module.
2. Load your data from a sample sheet or from LIMS using the SNP manifest file (*.bpm) and standard cluster file (*.egt) that correspond to the Infinium product used. For instructions on loading data from a sample sheet, refer to the GenomeStudio GT Module User Guide (part # 11319113).
3. After the GenomeStudio tables are populated with data, click the Calculate button () above the Samples Table. This updates the average call rate in the Call Rate column for each sample across all loci. Average call rate expectations for standard human products are typically greater than 99%. Refer to the product literature for specific call rate specifications.
4. To identify problematic samples, create scatter plots of call rate as a function of sample number (Figure 1), and 10% GenCall (10% GC) score as a function of sample call rate (Figure 2). To

Figure 1: Scatter Plot of Call Rates Across a Set of Samples



Viewing the call rate of all samples quickly reveals poorly performing samples with abnormally low call rates (green circle).

Figure 2: Scatter Plot of 10% GC Score Compared to Call Rates of a Sample Set



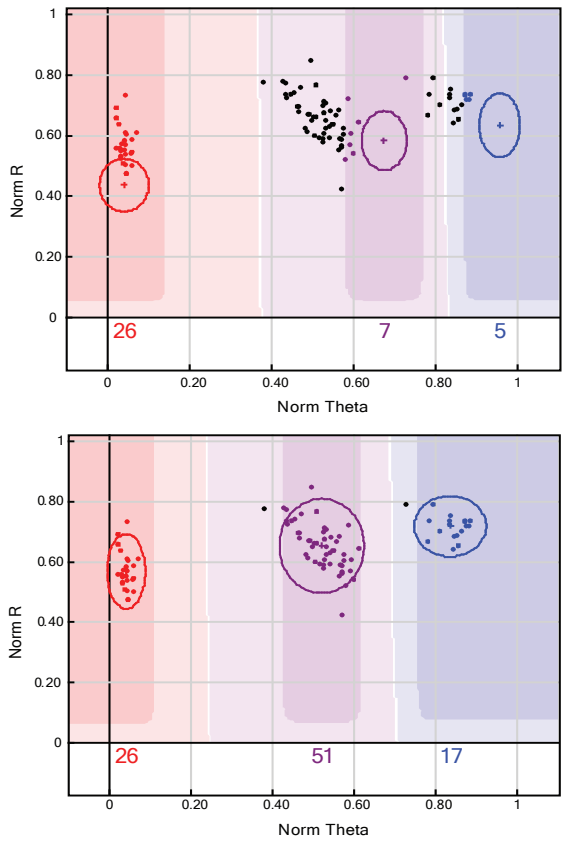
Poorly performing samples are obvious outliers from most of the samples when 10% GC Score is plotted against sample call rate (green circle).

create graphs, use the graphing functions in GenomeStudio, or export the Sample Table data and reopen it in another graphing program. Poorly performing samples can be identified as having low sample call rates and low 10% GC scores. Samples with low call rates and 10% GC scores that are outliers from the main population should be considered for reprocessing.

However, if there are samples expected to have large amounts of loss of heterozygosity (LOH) or copy number variation (CNV), such as tumor or WGA samples, there may be outliers in call rate and 10% GC. This may be due to biological reasons and not because of poor DNA quality. Therefore, these samples are not good candidates for reprocessing.

In human projects, if call rates below 99% are observed across most samples, the measured sample intensities may not match the intensities from the standard cluster file provided by Illumina. To test for this possibility, recluster SNPs on good samples only and recalculate the average call rate. Save this new preliminary cluster file with a meaningful name (e.g., Preliminary_Hap550_Project Name.egt). If the sample call rate is still low after reclustering samples with the highest call rates and 10% GC scores, there could be a systematic non-biological issue affecting data quality.

Figure 3: Example of a Successfully Reclustered SNP



Three distinct clusters exist but are poorly represented by the standard cluster file. This leads to an initial SNP call frequency of 39.6% for this locus (before). After reclustering, the cluster positions better represent the data and yield a call frequency of 99.7% (after).

Determining Final Sample Set

After poorly performing samples have been reprocessed, the new requeued samples should be screened in GenomeStudio for the best performing instance of each sample. The higher-quality instance of each sample (across all SNPs) should be retained for analysis. If both instances of a sample perform poorly, the entire sample should be excluded from the project before generating final reports. A procedure for including only samples that perform best follows:

1) Load reprocessed DNA samples into the existing GenomeStudio project by either selecting File | Update Project from LIMS or File | Load Additional Samples.

2) Select the appropriate sample sheet and data repository.

3) After all additional samples have been included, click the Calculate button.

4) Select Analysis | Exclude Samples by Best Run. This automatically excludes the lower quality requeued samples throughout the entire project.

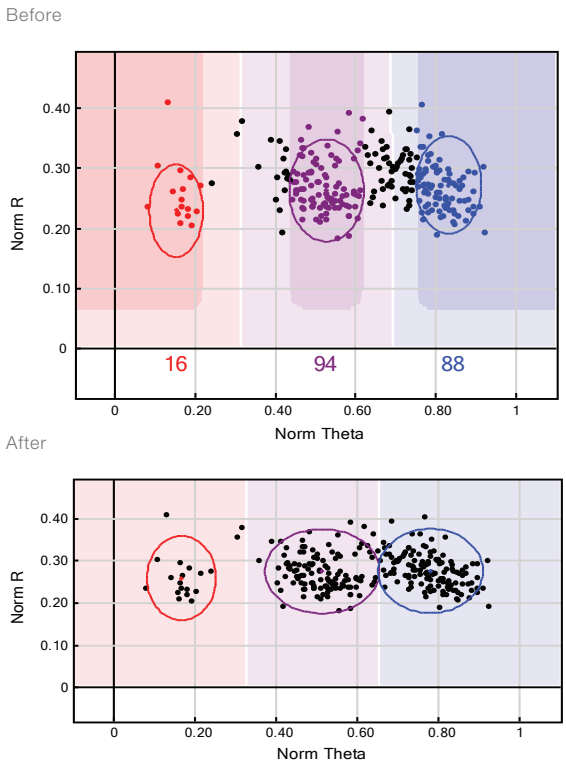
5) To determine whether any additional samples should be excluded from the project, reevaluate samples as described in the previous section.

6) When the GenomeStudio project contains only final and unique samples, update the SNP statistics.

Locus Analysis and Reclustering

Standard cluster files provided with Infinium products identify expected intensity levels of genotype classes for each SNP. For human projects, comparing sample intensities to this cluster file is sufficient for generating high-quality data. Standard cluster files for standard human Infinium products are created using a diverse set of over 100 samples from the Caucasian (CEU), Asian (CHB+JPT), and Yoruban (YRI) HapMap populations. Therefore, much of the genetic diversity in these populations is incorporated. Use this standard file when call rates are high and the genetic diversity of a sample set is well represented by the standard cluster file.

Figure 4: Example of an Unsuccessfully Reclustered SNP



At this locus, project samples form diffuse clusters and do not correlate well with the standard cluster positions (before). The genotype calls are unreliable because of poor cluster separation. Reclustering on this locus does not improve data quality. If not zeroed automatically by the algorithm (as shown), this locus should be zeroed based on the low score for cluster separation.

However, in some situations, sample intensities might not overlay perfectly onto the standard cluster positions. Reclustering some or all SNPs can optimize the GenomeStudio software's ability to call genotypes and may result in higher overall call rates (Figure 3). In some cases, reclustering is not helpful and the locus must be zeroed (Figure 4). Reclustering using samples from a particular project will redefine cluster positions, making them more representative of the samples in the project.

All or a subset of loci can be reclustered to generate a custom cluster file. To recluster all SNPs in the GenomeStudio project select Analysis | Cluster All SNPs. To recluster a subset of SNPs, highlight relevant rows of the SNP Table, right-click the SNP Table, and select Cluster Selected SNPs. An alternative method is to use filters to identify and select SNPs that should be excluded from reclustering. Consult the GenomeStudio Framework User Guide (part # 11318815) for information about how to create filters for various metrics in the Full Data Table, SNP Table, or Samples Table. Generating a custom cluster file will require a thorough evaluation of newly defined SNP cluster positions to ensure the accuracy of the genotype data. See Evaluating and Editing SNP Cluster Positions later in this document.

An important consideration in the decision to recluster is that the GenomeStudio clustering algorithms require approximately 100 samples to produce cluster positions that are representative of the overlying population. Therefore, in projects with fewer than 100 different samples, it is best to use the standard cluster file for calling genotypes.

Special consideration must be given to mitochondrial (mtDNA) and Y chromosome (Y-chr) SNPs. The GenomeStudio clustering algorithms do not automatically accommodate loci that lack heterozygous clusters. The mtDNA (Figures 5 and 6) and Y-chr SNPs (Figures 7 and 8) must be manually edited following reclustering. Before reclustering all SNPs, evaluate the mtDNA and Y-chr SNPs to decide whether the predefined Illumina standard cluster positions are appropriate for the current project samples. If appropriate, exclude these loci from reclustering.

Evaluating and Editing SNP Cluster Positions

To identify loci that should be manually edited or zeroed due to poor performance (< 1%), evaluate newly reclustered SNPs using the metrics listed in the SNP Table. These metrics are based on all samples for each locus, and thus provide overall performance information for each locus. To find loci that may require editing or removal before generating final reports, sequentially sorting the SNP Table by columns and screening loci by each quality metric identified in Table 1 is recommended. To determine hard cutoffs and grey zones, sort data by one column at a time and explore values starting at the extremes of the ranges. A hard cutoff, defined as the level above (or below) which most of the loci are unsuccessful, can be applied and the loci zeroed when necessary. To improve analysis further, the grey zone, defined to contain loci that are 80–90% successful, can be manually adjusted. The upper (or lower) limit of the grey zone is the point at which all loci are successful. SNPs falling in the grey zone should be visually evaluated and either accepted, zeroed, or manually edited by moving cluster positions. Hard cutoffs and grey zones may not transfer between projects because they are highly dependent on

Figure 5: A Successfully Edited Mitochondrial SNP

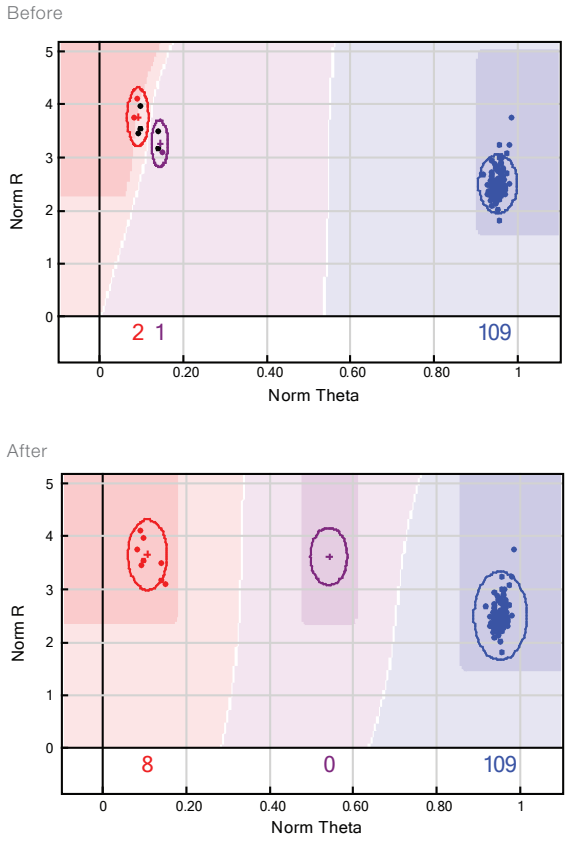
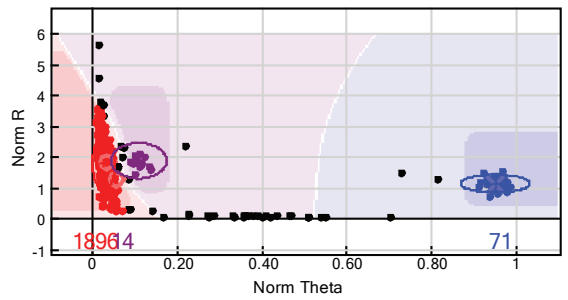


Figure 6: An Unsuccessful Mitochondrial SNP



The clustering is ambiguous and genotypes are unreliable. This SNP should be zeroed.

sample quality. Additional metrics may be required with each new project and should be determined on an individual basis.

The choice to edit loci manually should be carefully considered. Any changes made should be consistent with principles of population genetics and prevent introduction of subjective bias into the data set. Additionally, visually inspecting cluster plots and manually editing loci can be very time consuming. Increasing the number of SNPs used for data generation is certainly beneficial, but the process should be evaluated for diminishing returns. For example, with projects requiring large numbers of SNPs such as genome-wide association studies (GWAS), applying hard cutoffs and zeroed SNPs may be fully adequate, excluding only 1–3% of the markers in the panel while significantly increasing data quality.

The following procedure describes a method for evaluating the quality of newly created cluster positions by sequentially sorting the SNP Table by various column metrics. For each step, it may be helpful to determine and record hard cutoff and grey zone thresholds.

1. For projects with 100 samples or more, sort the SNP Table by **Cluster Sep**. Cluster Sep measures the separation between the three genotype clusters in the theta dimension and varies from 0–1. Evaluate individual SNPs for overlapping clusters, starting with those having low Cluster Sep. If clusters are well separated, the SNP can be manually edited (Figure 9). SNPs with overlapping clusters should be zeroed (Figure 10).

2. Sort the SNP Table by call frequency (**Call_Freq**). Call_Freq is the proportion of all samples at each locus with call scores above the no-call threshold. The value varies from 0–1. Evaluate SNPs starting with those having low Call_Freq values. Zero the SNP if the low call frequency cannot be attributed to a potential biological effect (Figures 11 and 12).

3. Sort the SNP Table by **AB R Mean**, the mean normalized intensity (R) of the heterozygote cluster. This metric helps identify SNPs with low intensity data and has values increasing from 0. Evaluate SNPs from low to high AB R Mean and zero any SNPs where the intensities are too low for genotypes to be called reliably (Figure 13).

4. Sort the SNP Table by **AB T Mean**, the mean of the normalized theta values of the heterozygote cluster. This value ranges from 0–1. Evaluate SNPs with AB T Mean ranging from 0–0.2 and 1–0.8 (or more, if necessary) to identify SNPs where the heterozygote cluster has shifted toward the homozygotes. Edit the SNP if clusters can be reliably separated (Figure 14). Otherwise, zero the locus.

5. To identify errors in Mendelian inheritance (**MI**), sort the SNP Table by columns **P-P-C Errors** and **P-C Errors**. Both columns measure deviations from expected allelic inheritance patterns in matched parent and child samples. Values range from zero to two times the maximum number of trios included in the project. In the GenomeStudio SNP Graph, MI errors are displayed as circles (o) for parents and a cross (×) for a child. Sort by this column and evaluate all SNPs with one or more errors. Zero SNPs with both MI errors and ambiguous clusters (Figure 15). SNPs associated with copy number polymorphisms may exhibit P-P-C and P-C errors that are biologically meaningful. Therefore,

loci with errors should be zeroed only if the clustering is also ambiguous (Figure 16).

6. The **Rep Errors** column of the SNP Table measures the reproducibility of genotype calls for replicate samples at each SNP. A range from 0 to the maximum number of replicates is included in the project. Sort by this column and evaluate all SNPs with one or more errors (Figures 17 and 18). Reproducibility errors are displayed as squares in the GenomeStudio SNP Graph.

7. The SNP Table metric, **Het Excess**, is an indicator of the quantity of excess heterozygote calls relative to expectations based on Hardy-Weinberg Equilibrium. This metric varies from -1 (complete deficiency of heterozygotes) to 1 (100% heterozygotes). Evaluate SNPs with Het Excess values less than -0.3 or greater than 0.2 (or more, if necessary). If clusters are unambiguous (Figure 19), edit the SNP and retain it for the final report. Otherwise, the SNP should be zeroed (Figure 20). Because males are not expected to be heterozygous for X chromosome loci, the Het Excess metric might not be relevant for X-linked SNPs. To evaluate X chromosome SNPs, see step 9.

8. The SNP Table column, **Minor_Freq**, measures the SNP minor allele frequency and can help identify loci where homozygotes have been incorrectly identified. Minor_Freq values vary from 0–1 and all SNPs with Minor_Freq less than 0.1 should be evaluated to detect false homozygotes. A SNP can be edited unless clusters overlap or cannot be separated unambiguously (Figure 21).

9. Because males are not expected to be heterozygous for any X-linked markers, these loci should be evaluated while taking gender into account. In GenomeStudio, the gender of each sample is estimated using X chromosome SNPs (Gender Est column in the Samples Table). Males can be selected in the Samples Table and marked with a different color on the SNP Graph to evaluate SNPs on the X chromosome. If males are called heterozygote at X chromosome SNPs, clusters may need to be manually adjusted so that they are called homozygote (Figure 22). However, if clustering is ambiguous, the SNP should be zeroed (Figure 23).

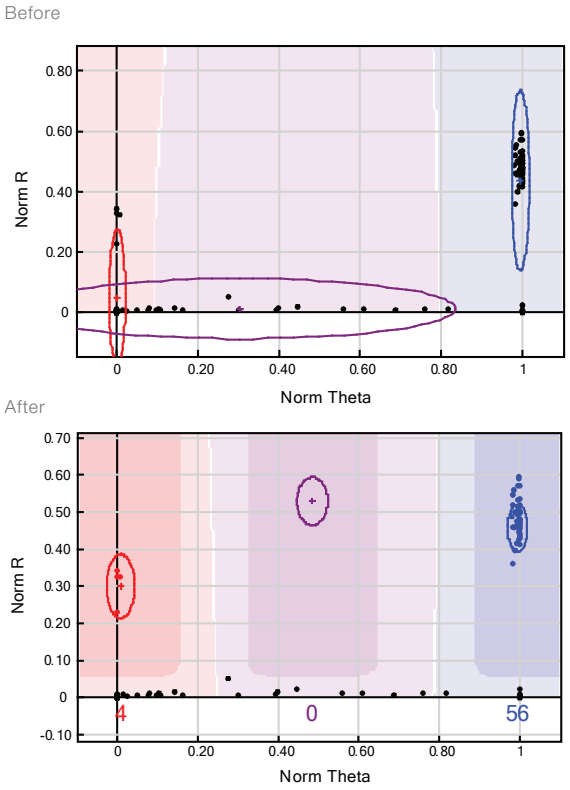
To watch a pre-recorded webinar featuring cluster evaluation, reclustering, and cluster editing, visit the Support Webinars list at icom.illumina.com/Webinar/Index. Select View Recorded Webinars, and choose the Infinium Exome Webinar.

Final Report

After completing the suggested data analysis and editing SNP clusters, calculate sample statistics by clicking the Calculator button in the Samples Table toolbar. If necessary, also update sample reproducibility and heritability statistics by selecting Analysis | Update Heritability/Reproducibility Errors. The GenomeStudio project is now final. Save changes using File | Save Project Copy As. The newly created cluster file (*.egt) can be exported from the GenomeStudio project by selecting File | Export Cluster Positions.

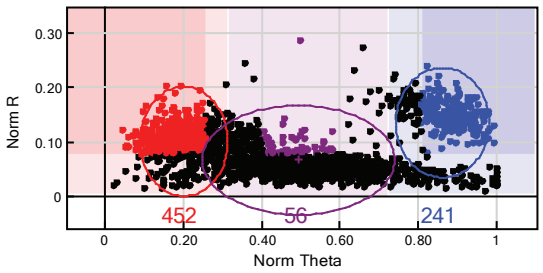
The Final Report Wizard lets you export the genotyping data from GenomeStudio for use in downstream analysis applications. To run the Final Report Wizard, select Analysis | Reports | Report Wizard and

Figure 7: A Successfully Edited Y-Chromosome SNP



Female samples have very low intensities, a broad range of theta values, and are incorrectly called AB after reclustering (before). After manually adjusting the cluster positions, female samples are not called, and the two homozygous clusters correctly define male genotypes (after).

Figure 8: An Unsuccessful Y Chromosome SNP



The clustering is ambiguous and the genotypes are unreliable. This SNP should be zeroed.

Table 1: GenomeStudio Clustering Algorithm Metrics

Single-Variable Metrics

Variable	Hard Cut Off	Grey Zone	Notes
Call_Freq*	< 0.97	≤ 0.99	User-defined
Rep Errors	> 2	> 0	
P-P-C Errors	> 2	> 0	
Cluster Sep	≤ 0.3	< 0.45	Some good quality SNPs identified in this range
AA R Mean	≤ 0.2	< 0.4	
AB R Mean	≤ 0.2	< 0.4	
BB R Mean	≤ 0.2	< 0.4	
10%_GC_Score	–	≤ 0.3	
Hex Excess	–	> 0.2	Hardy Weinberg equilibrium
A/B_Freq	–	≥ 0.4	Hardy Weinberg equilibrium
AB T Mean	–	(< 0.2) or (> 0.8)	Identifies nearby polymorphisms

Multi-Variable Metrics

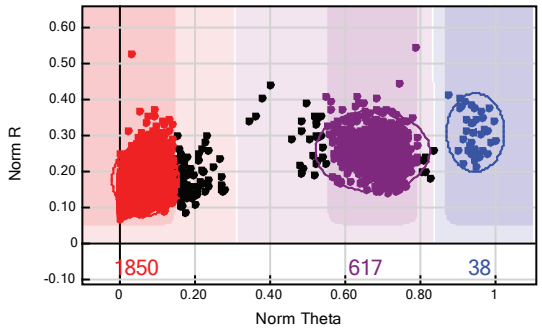
Filter Condition	Variable	Hard Cut Off	Grey Zone	Notes
A/A_Freq = 1	AA T Mean	> 0.3	> 0.2	AA homozygote clustering far from axis
A/A_Freq = 1	AA T Dev	> 0.06	> 0.04	Wide AA homozygote clustering
B/B_Freq = 1	BB T Mean	< 0.7	< 0.8	BB homozygote clustering far from axis
B/B_Freq = 1	BB T Dev	> 0.06	> 0.04	Wide BB homozygote clustering
A/A_ or B/B_Freq = 0	AB T Dev	–	> 0.05	Identifies possible nearby polymorphisms
A/B_Freq = 0	Minor_Freq (MAF)†	–	> 0	Identifies heterozygote clusters incorrectly classified as homozygote
0.998 > Call_Freq > 0.990	Minor_Freq (MAF)	–	< 0.05	Low MAF clusters identified

* Y chromosome markers should be excluded from call rate metric before analysis.

† X chromosome, Y chromosome, and mtDNA SNPs should be excluded from the Minor_Freq metric before analysis.

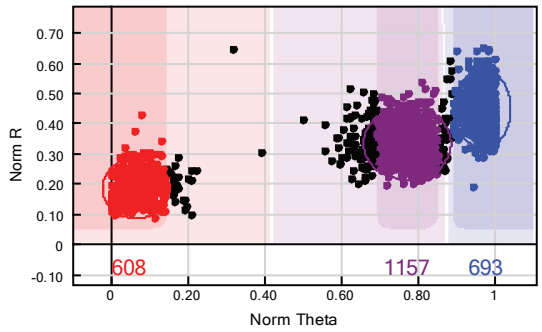
To increase data quality, various quality metrics defined in the GenomeStudio clustering algorithm are considered. Loci are either zeroed automatically (Hard Cutoff) or visually inspected for accuracy (Grey Zone). These metrics allow evaluation of cluster width, deviation, and intensity, helping to identify errors in cluster assignment. Metric definitions can be found in the appendix at the end of this document. **Note:** The metric values in this table are for human studies. Non-human studies may require different cut-off values.

Figure 9: Example of a Successful SNP Locus



Clusters are sufficiently separated so that clustering is not ambiguous. Manually editing this locus can increase the call frequency.

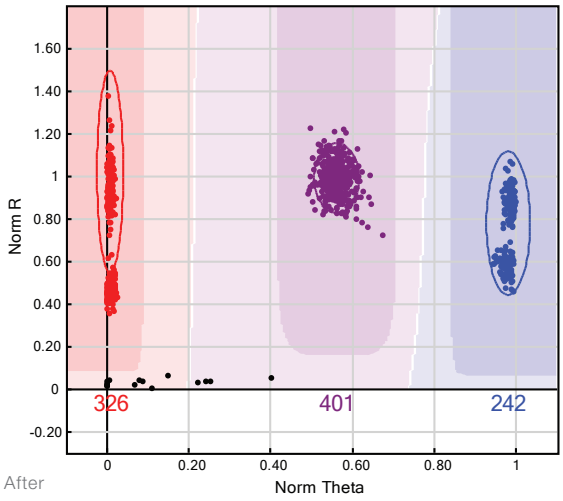
Figure 10: Example of an Unsuccessful Locus



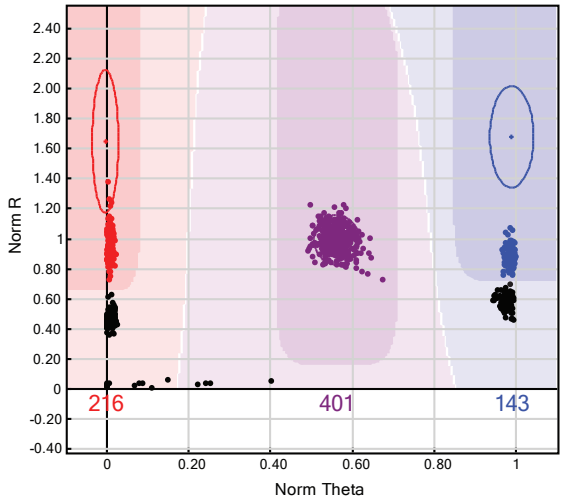
AB and BB clusters overlap and genotyping cannot be considered reliable. This SNP should be zeroed.

Figure 11: A Successful Locus with Low Call Frequency

Before

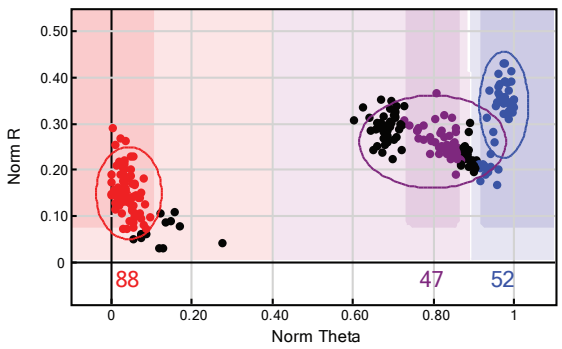


After



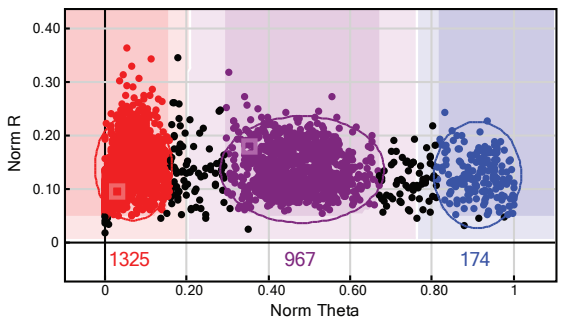
Although this SNP has a call frequency that falls below expectations, it has a phenotype that may indicate the presence of a deletion or a third polymorphic allele. There are two AA clusters (possibly AA and A/-), an AB cluster, and two BB clusters (possibly BB and B/-). In this case, samples that are not called (the black samples along the X-axis) may represent a biologically meaningful failure of those samples (e.g., -/-). Illumina recommends no-calling the A/- and B/- clusters.

Figure 12: An Unsuccessful Locus with Low Call Frequency



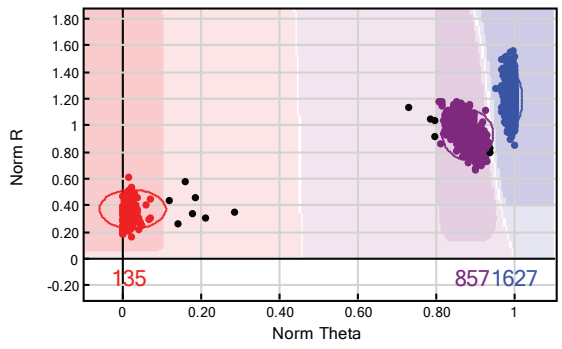
AB and BB clusters are not distinct and genotyping cannot be considered reliable. This SNP should be zeroed.

Figure 13: SNP with Low Intensities



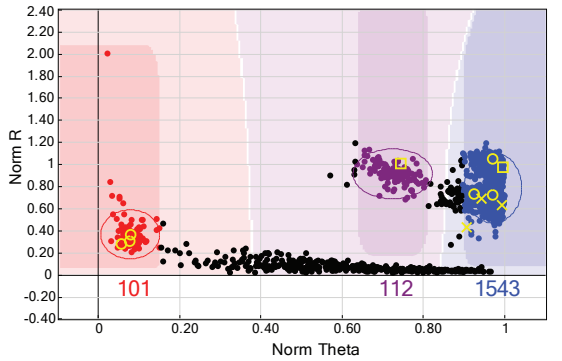
Intensities (Norm R) are too low for genotypes to be reliably called. This SNP should be zeroed.

Figure 14: A Successful SNP Locus With High AB T Mean Value



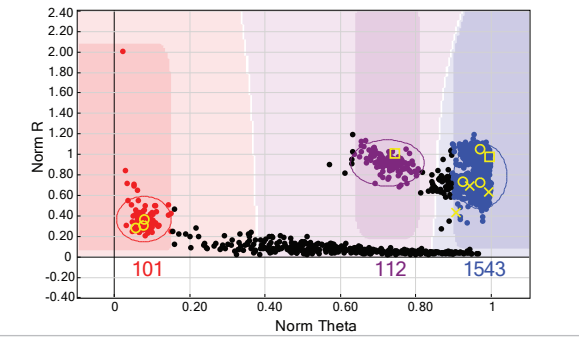
This locus can be manually edited to include more AA samples to improve call frequency. There is no ambiguity in clusters at this locus.

Figure 15: An Unsuccessful Locus with MI Errors



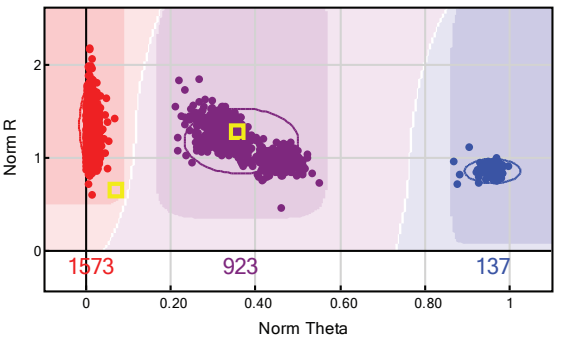
This locus exhibits six MI errors and ambiguous clusters (left BB cluster could be AB). This locus should be zeroed.

Figure 16: Example of a SNP With Erroneous Trios



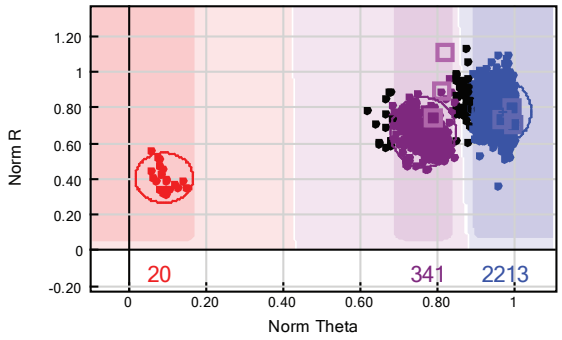
Clustering is not ambiguous and the locus has a phenotype that may indicate the presence of a chromosomal deletion or third polymorphic allele. This locus might be of special interest.

Figure 17: A Successful Locus with a Reproducibility Error



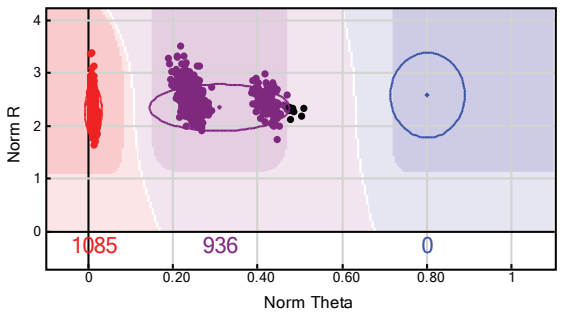
There is one erroneous genotype (the AA replicate should be AB, and could be manually no-called) and the clustering is not ambiguous at this locus.

Figure 18: A Failed Locus with Reproducibility Errors



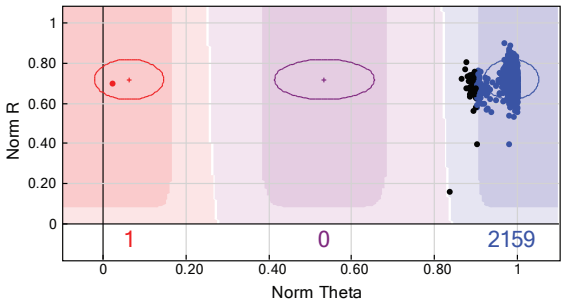
AB and BB clusters overlap. Genotypes cannot be called reliably so this locus should be zeroed.

Figure 19: Example of a SNP with High Heterozygote Excess



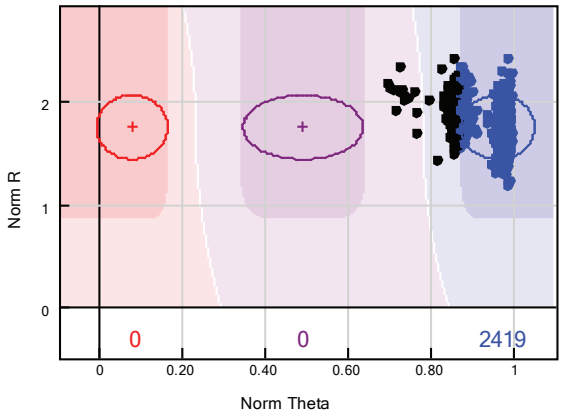
Clusters can be separated and are not ambiguous. If the algorithm minimum cluster separation parameters allow it, this locus can be manually edited to call genotypes correctly.

Figure 20: Example of a SNP with Heterozygote Deficiency



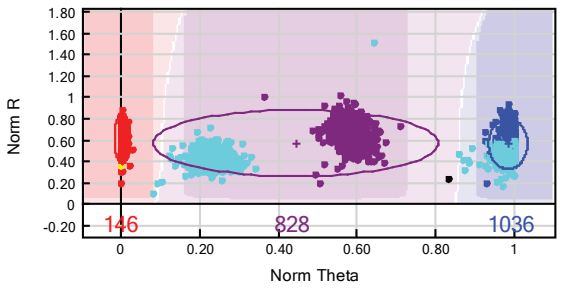
The clustering is ambiguous (left BB cluster could be AB). This locus should be zeroed.

Figure 21: Example of a SNP with Minor_Freq of 10



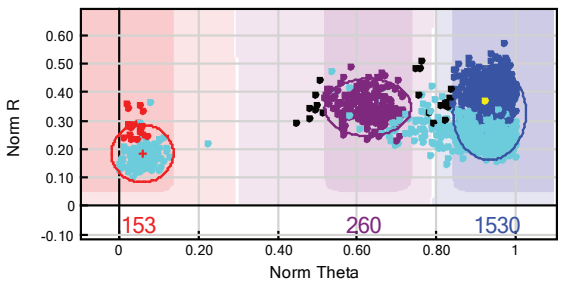
This SNP has been incorrectly clustered with a Minor_Freq of 0. In this case, clusters are too close together for genotypes to be called and the locus should be zeroed.

Figure 22: Example of X Chromosome SNP that Can Be Successfully Edited



The clustering algorithm called males AB at this X chromosome SNP. The males (light blue) that are called heterozygote should be manually edited to be included in the AA cluster.

Figure 23: Example of an Unsuccessful X Locus



Male AB and BB clusters (light blue) overlap and genotyping cannot be considered reliable.

follow the on-screen instructions. Filter the project data for excluded samples and non-zeroed SNPs, group, or other attributes. The Final Report Wizard also allows you to include SNP annotation and choose from among various formats to accommodate most genotyping analysis packages.

Summary

By identifying problematic samples and loci in a systematic manner, a project data set can be analyzed comprehensively. This takes full advantage of the robustness of Illumina Genotyping BeadChips in genotyping experiments. The sample workflow described in this document is a method recommended by Illumina to optimize final data quality. By editing loci that are not clustered or called correctly, collected data can be fully utilized. When editing is not possible for unsuccessful loci or samples, excluding them from the data set ensures that the remaining data are of the highest quality.

Appendix: DNA Report Column Definitions

#No_Calls: The total number of genotypes in each sample with a GenCall score below the no-call threshold as defined in the project options. Genotypes that are not called are shown on the GenomeStudio SNP Graph as black points falling outside of the darkly shaded regions.

#Calls: The total number of genotypes in each sample with a GenCall score above the no-call threshold.

Call_Freq: Call_Freq is equal to #Calls/(#No_Calls + #Calls). Call_Freq is equivalent to Call Rate in the GenomeStudio Samples Table.

A/A_Freq: For each sample, the number of AA genotype calls divided by #Calls.

A/B_Freq: For each sample, the number of AB genotype calls divided by #Calls.

B/B_Freq: For each sample, the number of BB genotype calls divided by #Calls.

R: The normalized R-value of a SNP in a specific sample.

AA R Mean: Mean of the normalized R-values for the AA genotypes.

AB R Mean: Mean of the normalized R-values for the AB genotypes.

BB R Mean: Mean of the normalized R-values for the BB genotypes.

Minor_Freq: If the number of AA calls is less than the number of BB calls for a sample, the frequency for the minor allele A is: [(2 • AA) + AB] divided by [2 • (AA+AB+BB)] across all called loci for that sample.

GenCall Score: This score is a quality metric that indicates the reliability of the genotypes called. A GenCall score value is calculated for every genotype and can range from 0.0 to 1.0. GenCall scores are calculated using information from the sample clustering algorithm. Each SNP is evaluated based on the angle of the clusters, dispersion of the clusters, overlap between clusters, and intensity. Genotypes with lower GenCall scores are located furthest from the center of a cluster and have lower reliability. There is no global interpretation of a GenCall score as it depends on the clustering of samples at each SNP. Clustering is affected by many different variables, including the quality of the samples and loci.

50%_GC_Score: For each sample, this represents the 50th percentile of the distribution of GenCall scores across all called genotypes. For SNPs across all samples, this is referred to as the 50%_GC_Score. For samples across all loci, it is referred to as p50GC in the Samples Table.

10%_GC_Score: For each sample, this represents the 10th percentile of the distribution of GenCall scores across all called genotypes. For SNPs across all samples, this is referred to as the 10%_GC_Score. For samples across all loci, it is referred to as p10GC in the Samples Table. Note: Call frequency, 50% GenCall score, and 10% GenCall score are useful metrics for evaluating the quality and performance of DNA samples in an experiment.

0/1: GenomeStudio calculates a threshold from the distribution of 10%_GC_Score values across all samples in the DNA report. A '1' is assigned to samples whose 10%_GC_Score is at or above this threshold. A '0' is assigned to samples whose 10%_GC_Score is below this threshold. The equation defining this threshold is 0.85 • 90th percentile of 10%_GC_Score values for all samples in DNA Report (i.e., 0.85 • 90th percentile of column K in the DNA report).

Additional Information

Learn more about GenomeStudio and other analysis software products from Illumina at www.illumina.com/software/genomestudio_software.ilmn.

References

1. Performing Clustering in the GenomeStudio Polyploid Clustering Module, supportres.illumina.com/documents/downloads/software/genomestudio/genomestudio_polyploid_clustering_15044760_a.pdf.

Illumina, Inc. • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 970-2007-005 Current as of 31 January 2014

illumina®