Analyzing GoldenGate® Genotyping Data

Preliminary analysis of data sets produced using the GoldenGate genotyping assay leads to optimized call rates, ensuring data is of the highest quality.

Introduction

The Illumina GoldenGate Genotyping assay is a flexible, pre-optimized assay that provides researchers with the ability to design custom panels for high-quality, low- to mid-multiplex genotyping studies. The GoldenGate portfolio offers a wide range of applications across both BeadArray™ and VeraCode® technologies for highly robust analysis of from 48 to over 3,000 markers in a single reaction. Sample multiplexing is available with the GoldenGate Indexing™ assay.

This technical note describes recommended methods for optimizing call rates and evaluating assay performance, sample quality, and locus performance when analyzing data from the GoldenGate genotyping assay. Analysis begins with an overall evaluation of assay performance and determination of which samples, if any, require reprocessing or removal. Clustering should be done after inclusion of reprocessed samples and removal of failed or suboptimal samples, allowing for a more detailed evaluation of sample quality. Each locus can then be evaluated for editing or zeroing (excluding) to optimize call rates.

Controls

Sample-dependent, sample-independent, and contamination controls are all built into the GoldenGate assay (Table 1 and Figure 1). These controls provide a way to assess the overall performance of samples, reagents, equipment, and BeadChips. Analyzing control performance should be the first step in evaluating assay performance. They can be visualized in GenomeStudio® software through various dashboards.

During preliminary sample quality evaluation, any samples falling outside the expected performance parameters should be highlighted for additional analysis.

Locus Analysis and Reclustering

The GenomeStudio Genotyping Module uses a clustering algorithm that allows the user to define cluster positions for all loci on the samples in a project. A minimum of 100 samples should be contained in a project when reclustering to maximize statistical significance. Prior to reclustering, set the GenCall threshold to 0.25. Samples can then be reclustered by clicking the recluster button or selecting **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.

Special consideration must be given to mitochondrial and Y chromosome SNPs (Figure 2). The clustering algorithm does not automatically accommodate loci that lack heterozygous clusters, so manual editing is recommended. In some cases, reclustering is not helpful and the locus must be zeroed.

GenCall Score

Before evaluating the quality of SNP clusters, it is important to highlight samples that have poor performance in the genotyping assay. The GenCall score is a quality metric, ranging from 0–1, calculated

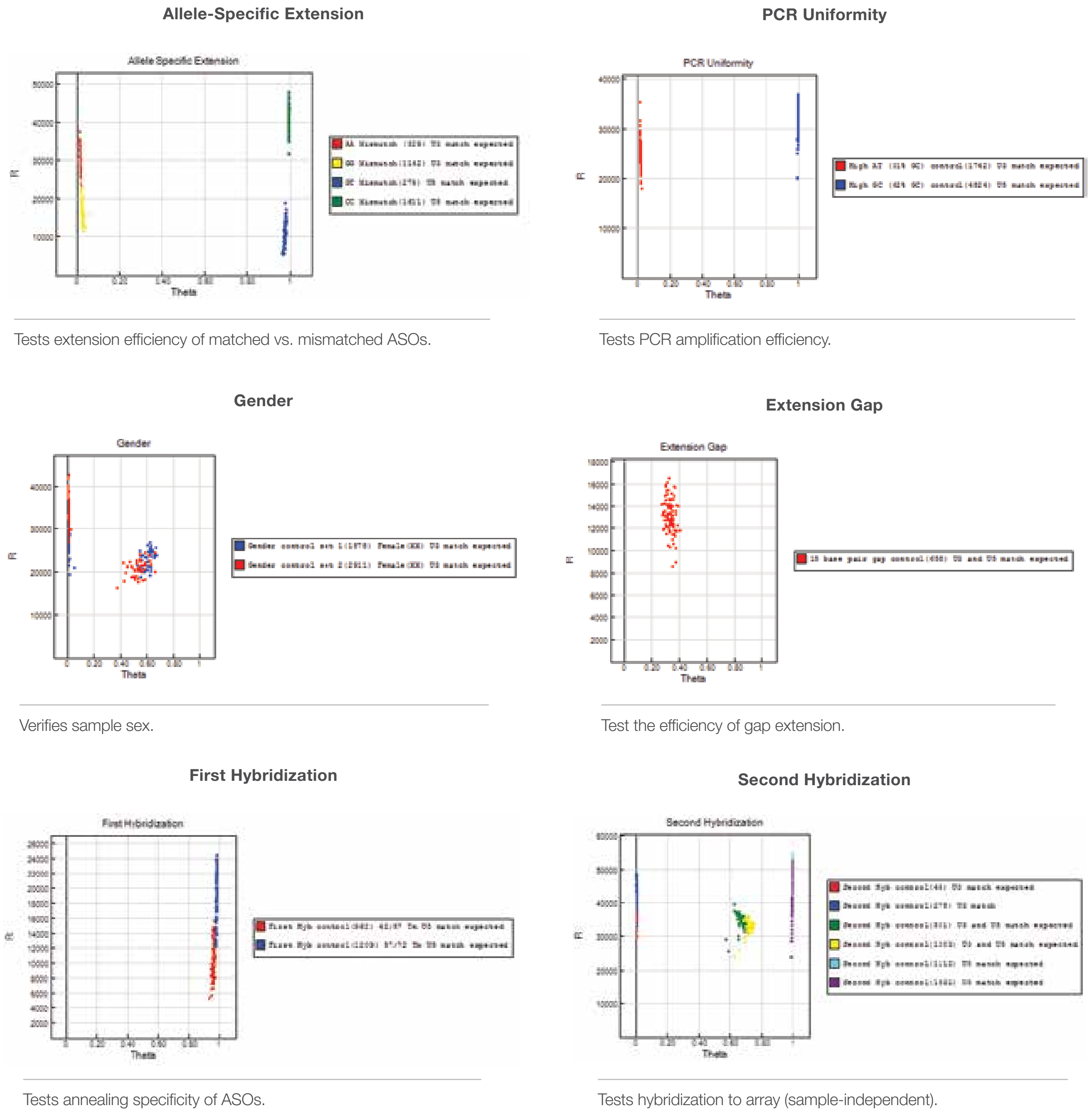
Table 1: Controls Provided in the GoldenGate Genotyping Assay*

Control	Purpose
Allele-Specific Extension (human only, not applicable for custom GGI†)	Sample-dependent control that measures the extension efficiency of the properly matched and mismatched allele-specific oligos (ASO). Failures in this control could indicate either a processing failure at the Make ASE or Make MEL steps or poor DNA sample quality.
Gender (human only, not applicable for custom GGI†)	Sample-dependent control that verifies the sex of DNA Samples.
First Hybridization (human only, not applicable for custom GGI†)	Sample-dependent control that tests the specificity of annealing ASOs with different melting temperatures (Tm) to the same DNA locus. Observed effects to signal intensity and/or clustering could indicate either a general processing failure at the Make ASE step or poor DNA sample quality.
PCR Uniformity (human only, not applicable for custom GGI†)	Sample-dependent control that tests the PCR amplification efficiency for high AT and high GC regions of DNA. Low signal intensities could indicate either a problem with the PCR amplification process or low sample quality.
Extension Gap (human only, not applicable for custom GGI†)	Sample-dependent control that tests the efficiency of the 15-base extension from the 3'-end of the ASO to the 5'-end of the locus-specific oligo (LSO). Issues related to processing during the Make MEL step or poor DNA sample quality can impact the behavior of these controls.
Second Hybridization (applicable to all species and GoldenGate assays)	Sample-dependent control that tests the specificity of synthetic oligo targets to probes present on the beads. Issues related to hybridization of samples to BeadChips or VeraCode beads can lead to unexpected behavior of this control.

* For more information on controls, refer to the GoldenGate Genotyping assay guide

† GGI = GoldenGate Indexing assay

Figure 1: GoldenGate Assay Controls Displayed in the GenomeStudio Genotyping Module Controls Dashboard



Allele-Specific Extension

Tests extension efficiency of matched vs. mismatched ASOs.

PCR Uniformity

Tests PCR amplification efficiency.

Gender

Verifies sample sex.

Extension Gap

Test the efficiency of gap extension.

First Hybridization

Tests annealing specificity of ASOs.

Second Hybridization

Tests hybridization to array (sample-independent).

for each genotype (data point). GenCall scores generally decrease in value the further a sample data point is from the center of its cluster. Each SNP is evaluated based on the angle, dispersion, and overlap of clusters and intensity.

To identify problematic samples, create a scatter plot of the call rate as a function of sample number against a 10% GenCall score (10% GC or p10 GC) as a function of the sample call rate (Figure 3). Poorly performing samples—those with low sample call rates, low 10% GC scores, and are outliers with the main population—should be considered for reprocessing or exclusion from the project.

Evaluating and Editing SNP Cluster Positions

To identify loci that need to be manually edited or zeroed (excluded), evaluate newly reclustered SNPs using the metrics listed in the SNP Table. These metrics are based on all samples for each locus, providing overall performance information for each locus. To find loci that might need to be edited or removed, start by determining hard cutoffs and grey zones. To do so, sort data one column at a time, exploring values at the extremes of the ranges. The hard cutoff should be defined as the level, below or above, at which the majority of loci are unsuccessful and should be zeroed. The grey zone should be defined to contain loci that are 80–90% successful and can be improved by manual editing. The upper limit (or lower limit) of the grey zone is the point at which all loci are successful. SNPs falling in the grey zone should be either zeroed or manually edited by moving cluster positions. Hard cutoffs and grey zones may not transfer between projects since they are highly dependent on sample quality and generally should be determined separately for each project.

Carefully consider the choice to manually edit loci. Any changes made should be consistent with principles of population genetics to prevent the introduction of subjective bias into the data set. The following procedure describes a method for evaluating the quality of newly created cluster positions by sequentially sorting the SNP Table by various column statistics. For each step, it may be helpful to determine and record hard cutoffs and grey zone thresholds.

Cluster Separation

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. SNPs with overlapping clusters should be zeroed (Figure 4).

SNP Call Frequency

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, such as a chromosomal deletion, in a subset of samples (Figure 5).

AB Mean for Intensity (R) and Theta (T)

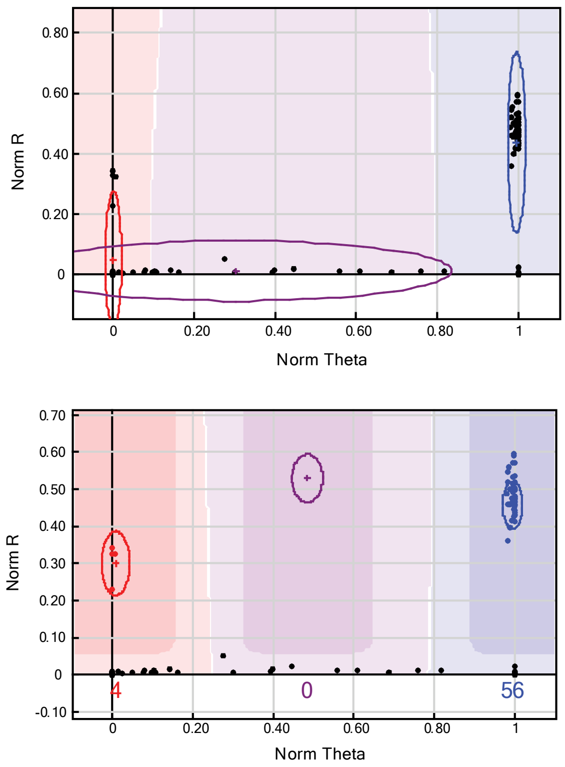
Sort the SNP Table by AB R Mean, the mean normalized intensity (R) of the heterozygote cluster. This metric helps identify SNPs with low intensity data and has values increasing from 0. Evaluate SNPs from low to high AB R Mean and zero any SNPs with intensities too low for genotypes to be called reliably (Figure 6).

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; otherwise, zero the locus.

Updating and Exporting Clusters

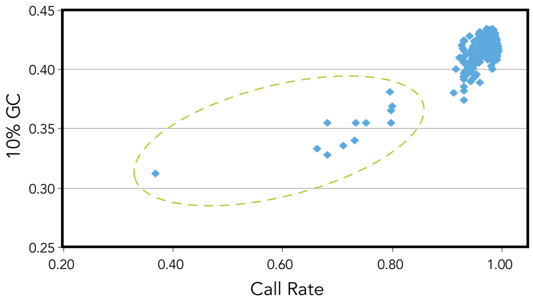
After completing the above data analysis and editing SNP clusters, calculate sample statistics by clicking the calculator button in the Samples Table toolbar. If necessary, also update sample reproducibility and heritability statistics by selecting **Analysis | Update Heritability | Reproducibility Errors.** At this point, the GenomeStudio project is final. The newly created cluster file (*.egt) can be exported from the GenomeStudio project by selecting **File | Export Cluster Positions.**

Figure 2: A Successfully Edited Chromosome SNP



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

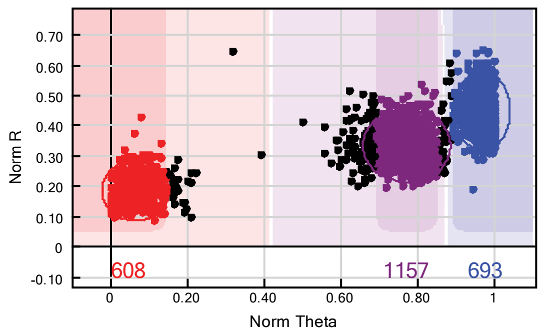
Figure 3: Scatter Plot of 10% GC Score Compared to Call Rates of a Sample Set



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

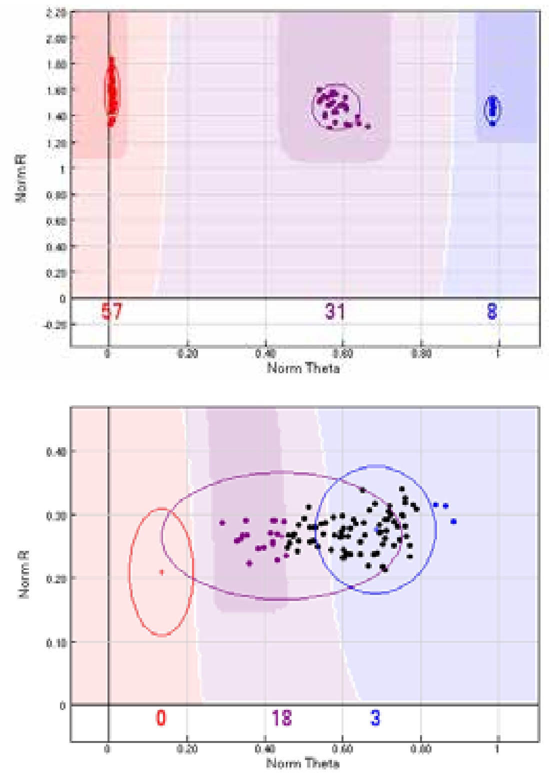
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCA
AATCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACC
AACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTA
TTAACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGAT
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCA
AAGATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGA
AACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACA

Figure 4: Example of an Unsuccessful Locus



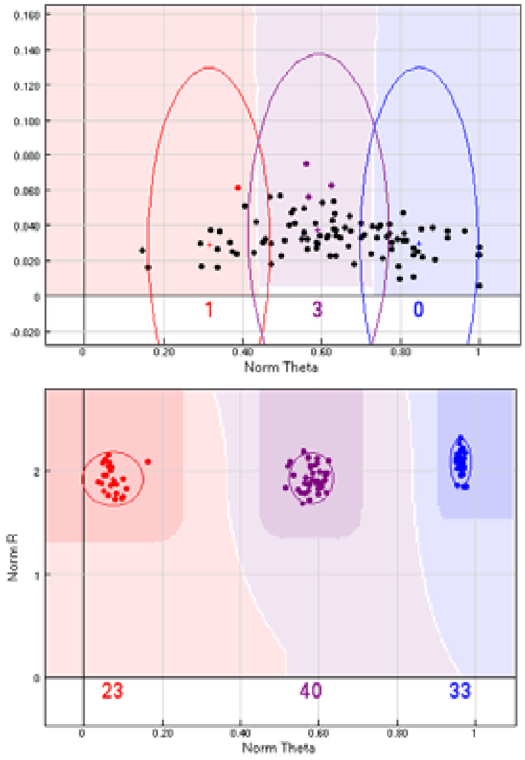
AB and BB clusters overlap, and genotyping cannot be considered reliable. This SNP should be zeroed.

Figure 5: Example of Loci with High and Low SNP Call Frequency



Call frequency (call freq) values range from 0-1. Upper panel: a locus with a call freq closer to 0; Lower panel: a locus with a call freq of 1.

Figure 6: Example of Loci with High and Low AB R Mean



Zero SNPs with low AB R Mean that may have intensities too low for reliable genotyping (upper panel). Lower panel: locus with high AB R Mean.

Final Report

The Final Report Wizard lets you export the genotyping data from GenomeStudio software for use in downstream analysis applications. To run, select **Analysis | Reports | Report Wizard** and follow the on-screen instructions to filter the project data for excluded samples, non-zeroed SNPs, by group, or by other attributes. The Final Report Wizard allows you to include SNP annotation and choose from a variety of formats to accommodate most genotyping analysis packages.

Summary

By identifying problematic samples and loci in a systematic manner, you can ensure optimal final data quality from the GoldenGate genotyping assay. Editing loci that are not clustered or called correctly, enables full use of collected data. 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.

Additional Information

Visit www.illumina.com or contact us at the address below to learn more about Illumina's DNA Analysis and Software products.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY