Improved Genotype Clustering with GenTrain 3.0

The GenTrain 3.0 clustering algorithm offers improvements for both sample intensity normalization and data clustering.

Introduction

The GenTrain clustering algorithm is an integral part of GenomeStudio® Analysis Software. To meet the escalating demands of Infinium® BeadChips with increasing marker densities, GenTrain 2.0 updates employ a 3-step clustering procedure to minimize erroneously clustered loci and deliver cleaner data sets. GenTrain 3.0, the successor to GenTrain 2.0, optimizes GenTrain 2.0 algorithms for the Infinium XT production-scale genotyping solution to enable accurate genotyping with efficient and robust clustering. Although GenTrain 3.0 is automatically included in the GenomeStudio 2.0 Software update, GenTrain 2.0 can still be selected for backward compatibility with previous experiments.

Overview of Data Clustering Method

Infinium assays produce 2-color readouts, with intensity values for each of the 2 colors (designated A and B) that convey information about the allelic ratio at a single genomic locus. For diploid organisms, biallelic loci are expected to exhibit 3 clusters (AA, AB, and BB). The GenTrain 2.0 algorithm simplifies the clustering process by transforming the A and B values into new values labeled θ (theta) and R. This transformation requires fewer parameters for proper characterization of clusters, which GenTrain 2.0 identifies through the 3-step clustering process. The GenTrain 3.0 update maintains this method, with improvements indicated in subsequent sections of this technical note.

1. **Preliminary Clustering:** GenTrain 2.0 performs a preliminary clustering to group samples with similar θ values with no consideration for the total number of clusters found.
2. **Secondary Clustering:** GenTrain 2.0 proposes a series of models to describe the observed θ and R values. The mean and standard deviations of θ and R are computed and used as estimates of the parameters of each cluster.
3. **Final Clustering:** GenTrain 2.0 scores each proposed model, taking into account the compactness of each cluster, the spread between clusters, and the probability of observing the sample assignment under Hardy-Weinberg equilibrium. The model with the highest score is used for genotype calling.

GenTrain 3.0 Improvements

The GenTrain 3.0 algorithm includes improvements for various aspects of data processing and analysis, including sample intensity normalization, clustering, and more.

Sample Normalization Performance for Infinium I Loci

In the sample intensity normalization process, specific groups of loci are normalized together in "normalization bins." Due to differences in probe design, Infinium I loci (2 probes per locus) and Infinium II loci (1 probe per locus) are normalized in separate bins. If the number of loci in a normalization bin is small (< 192 loci), the normalization process can be negatively impacted. With the low bead pool complexity supported on the Infinium XT platform, the occurrence of small normalization bins may be more prevalent, especially with normalization bins consisting of Infinium I loci. With the GenTrain 2.0 algorithm, small normalization bin size negatively impacts the normalization of intensity data for the given locus (Figure 1A).

The GenTrain 3.0 algorithm improves the normalization of small bins by taking advantage of the special nature of Infinium I assay data, where the signal intensity for both alleles originates in the same color channel. This affords the possibility to fit a normalization model with only 2 free parameters, instead of 6. When applied to the same data mishandled by GenTrain 2.0, GenTrain 3.0 improves the performance of the intensity normalization and generates tight clusters (Figure 1B). The GenTrain 3.0 algorithm applies the improved normalization model for any normalization bin containing fewer than 192 Infinium I loci.

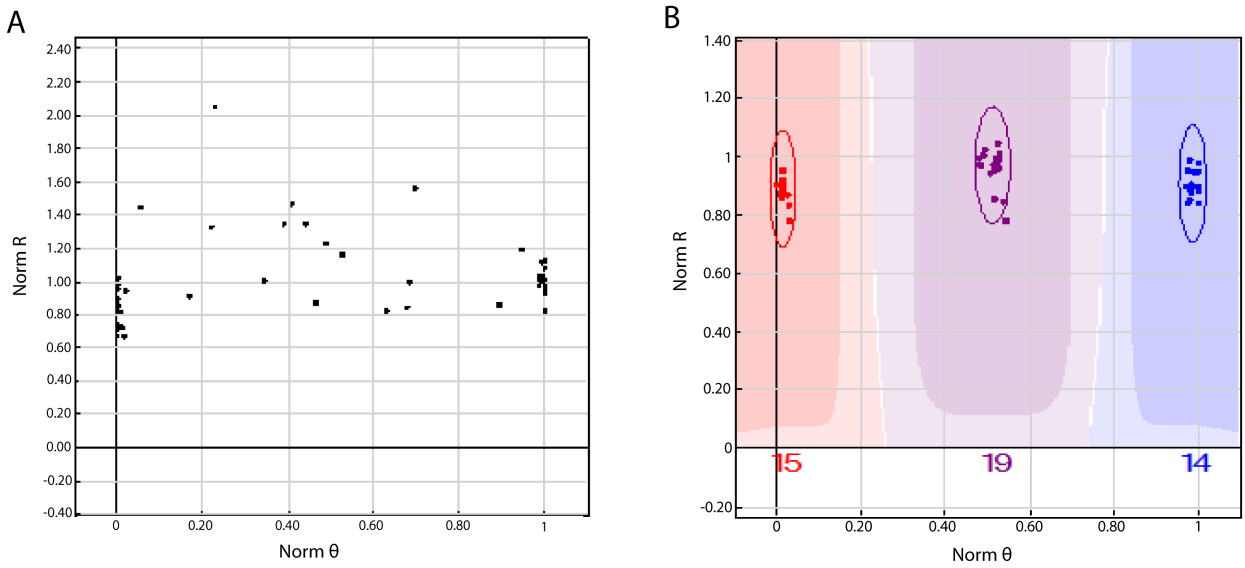

Figure 1: Comparison of Sample Normalization for a Small Number of Loci Between GenTrain 2.0 and GenTrain 3.0—A normalization bin with a few loci negatively impacts clustering in the GenTrain 2.0 algorithm (A). Improvements to sample normalization for normalization bins with low numbers of loci in GenTrain 3.0 allow for proper clustering (B).

Data Clustering Performance

In the final clustering step, the GenTrain 2.0 algorithm scores many possible clustering models of the data. Improvements in Infinium XT technology can reduce variability within a genotype cluster. As a result, the center of homozygous clusters can be much closer to (or on top of) the axis in the Norm R vs Norm θ display. The GenTrain 2.0 algorithm assigns a strong negative penalty to this clustering configuration (Figure 2A).

In the GenTrain 3.0 algorithm, the magnitude of this negative contribution is limited to successfully cluster loci from Infinium XT data (Figure 2B). However, changes to the GenTrain 3.0 algorithm are designed to ensure consistent clustering results compared to the previous approach in GenTrain 2.0.

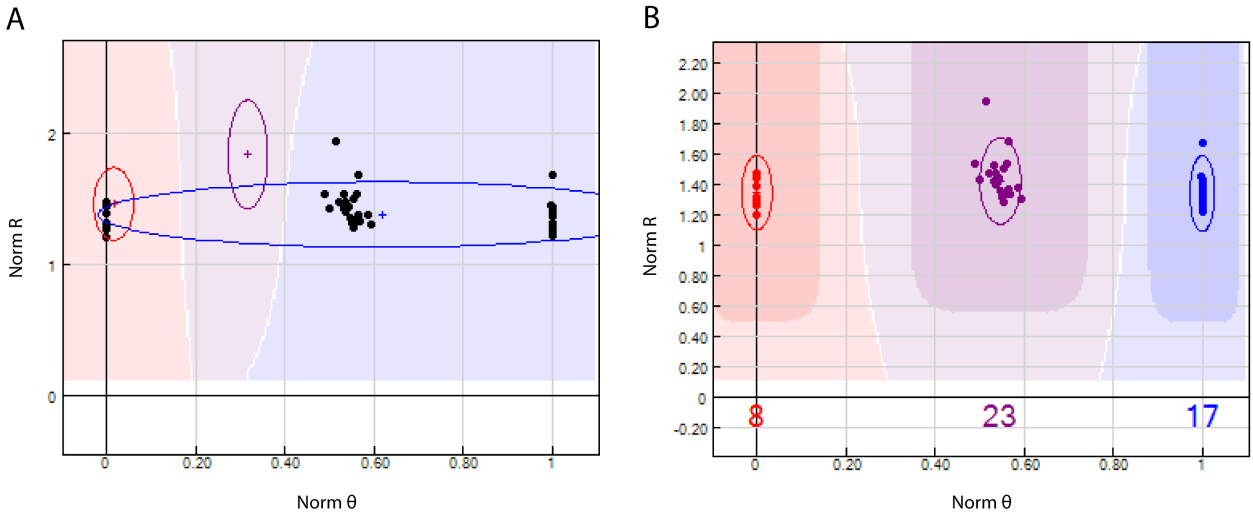

Figure 2: Comparison of Data Clustering in GenTrain 2.0 and GenTrain 3.0—The GenTrain 2.0 algorithm penalizes a homozygous cluster that falls on the Norm R axis (A, red circle). The GenTrain 3.0 algorithm does not penalize this configuration and properly identifies all 3 clusters (B).

Proper Handling of Samples with No Intensity Data

Part of the intensity normalization process involves an estimate of those samples that belong to the AA and BB genotyping clusters. In the GenTrain 2.0 algorithm, samples with 0 intensity in both channels are inappropriately assigned to the BB cluster for the purposes of estimating the average of BB sample intensities. GenTrain 3.0 was designed to ignore these points when estimating the average intensity of the BB cluster. This only applies to normalization bins with fewer than 192 loci.

Performance

The GenTrain 3.0 algorithm does not dramatically impact call rate for high-complexity arrays, except for the special cases noted in this technical note. To demonstrate this consistency, a data set of 10,000 samples from a high-complexity human BeadChip with approximately 700,000 loci was processed with GenTrain 2.0 and GenTrain 3.0. Call rates were the same with both algorithms (Table 1). The manifest for this BeadChip was then modified to simulate a low-complexity array with only 60 loci. On the same set of 10,000 samples, call rates using the GenTrain 3.0 algorithm were slightly improved compared to the GenTrain 2.0 algorithm (Table 1). Of the 5449 samples with a difference in call rate between the 2 algorithms, GenTrain 3.0 improved calls in 96% of those cases. Concordance was also high between GenTrain 2.0 and GenTrain 3.0 when comparing analysis results on the Infinium OmniExpress-24 and Infinium Omni2.5-8 BeadChips (Table 2).

Table 1: Comparison of Call Rates Between GenTrain 2.0 and GenTrain 3.0

	GenTrain 2.0	GenTrain 3.0
Call Rate – 700,000 loci	99.45%	99.45%
Call Rate – 60 loci	98.23%	99.70%

Table 2: Comparison of GenTrain 2.0 and GenTrain 3.0 on 2 Infinium BeadChip Products

OmniExpress-24 v1.2 (A1)	GenTrain 2.0	GenTrain 3.0
Call Rate	99.84%	99.84%
LogR Dev	0.093	0.093
Reproducibility	100%	> 99.99%
Concordance between GenTrain 2.0 and GenTrain 3.0	> 99.99%	
Omni2.5-8 v1.3 (A1)	**GenTrain 2.0**	**GenTrain 3.0**
Call Rate	99.86%	99.86%
LogR Dev	0.114	0.109
Reproducibility	> 99.99%	> 99.99%
Concordance between GenTrain 2.0 and GenTrain 3.0	> 99.99%	

Summary

GenTrain 3.0 updates and replaces the GenTrain 2.0 clustering algorithm as the default algorithm in the GenomeStudio 2.0 Software update. GenTrain 3.0 optimizes sample normalization and clustering to improve accurate genotype calling with Infinium XT data sets. Call rate performance in GenTrain 3.0 is maintained for high-complexity arrays and improved over GenTrain 2.0 for low-complexity Infinium XT arrays.

Learn More

To learn more about Illumina genotyping solutions, visit www.illumina.com/applications/genotyping.html.

References

1. Illumina (2009) Improved cluster generation with Gentrain2. (www.illumina.com/documents/products/technotes/technote_gentrain2.pdf).
2. Illumina (2016) Infinium iSelect custom genotyping assays. (www.illumina.com/documents/products/technotes/technote_iselect_design.pdf).

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

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

illumina®