TECHNOLOGY SPOTLIGHT

Illumina GenCall Data Analysis Software

GenCall software algorithms for clustering, calling, and scoring genotypes

INTRODUCTION

Illumina's BeadStation and BeadLab genotyping solutions use the GenCall software application to automatically cluster, call genotypes, and assign confidence scores. The GenCall application incorporates a clustering algorithm (GenTrain) and a calling algorithm.Genotyping calls for a specific DNA are made by the calling algorithm, relying on information provided by the GenTrain clustering algorithm.

CLUSTERING ALGORITHM OVERVIEW

In a genotyping analysis, DNA from a population of several individuals is analyzed by a set of multiplexed arrays. The data for each multiplexed array is self-normalized using the information contained in that specific array. This normalization algorithm adjusts for nominal intensity variations observed in the two color channels, background differences between the channels, and possible crosstalk between the dyes. The behavior of each locus is then modeled using a custom clustering algorithm that incorporates several biological heuristics on SNP genotyping. In cases where fewer than three clusters are observed (e.g., due to low minor-allele frequency), locations and shapes of the missing clusters are estimated using neural networks. Depending on the shapes of the clusters and their relative distance to each other, a statistical score is devised (the GenTrain score). This score is designed to mimic evaluations made by a human expert's visual and cognitive systems. In addition, it has been evolved using the genotyping data from top and bottom strands (see "Score Validation"). This score is combined with several penalty terms (for example low intensity, mismatch between existing and predicted clusters) in order to make up the training ("GenTrain") score. The GenTrain score, along with the cluster positions and shapes for each SNP, is saved for use by the calling algorithm.

CALLING ALGORITHM OVERVIEW

To call genotypes for an individual's DNA, the calling algorithm takes the DNA's intensity values and the information generated by the clustering algorithm; subsequently, it then identifies to which cluster the data for any specific locus (of the DNA of interest) corresponds. The DNA data is first normalized (using the same procedure as for the clustering algorithm). The calling operation (classification) is performed using a Bayesian model. The score for each call (GenCall Score) is the product of the GenTrain Score and a data-to-model fit score. After scoring all the loci in the DNA of interest, the application computes a composite score for that DNA (DNA Score). Subsequently, the GenCall score of each locus for this DNA is further penalized by the DNA Score.

SCOPE OF THE GENCALL SCORE

The GenCall Score is not a probability, but a score, primarily designed as a means by which to rank and filter out failed genotypes, DNAs, and/or loci.[1] The sensitive region of the GenCall Score (i.e., the region with highest rate of change of accuracy versus GenCall Score) is between the values of 0.2 and 0.7 *(Figure 1)*. Scores below 0.2 generally indicate failed genotypes, while scores above 0.7 usually report well-behaving genotypes.

GenCall Scores may be averaged among DNAs and among loci for purposes of evaluating the quality of the genotyping within a particular DNA or locus. For example, we often evaluate "GC10" and "GC50" scores that are calculated by taking the 10th percentile and the 50th percentile (median) of the GenCall Scores for a certain locus, respectively. Using GC10 and GC50 Scores, a user may choose to fail particularly poor performing loci, for instance, by discarding loci with GC10 of 0.1 or lower. Also, a series of aggregate statistics (i.e., average) of the GC10 or GC50 scores for each DNA can be used to identify low-quality DNAs (for instance, a user may discard DNA samples with average GC10 scores of 0.2 or lower). The GenCall Score can also be used in situations where users have a minimum required call rate. This rate translates to making calls on a certain percentile of the data. Users can sort all their genotypes based on the GenCall Score, and then choose the top (Nth) percentile of interest for their study.

SCORE VALIDATION

The informative power of the GenCall score has been verified by top/bottom strand correlation studies.[1, 2] In these studies, each SNP is designed twice, once from the top strand, and once from the bottom strand. After completion of the assays and running GenCall, the data for the two strands are compared for similarity in the final calls. For a successful assay, one would not expect to have different calls (e.g., A/A and A/B) made on the two strands of a certain locus. Prior to comparing the calls, the "No Call" genotypes are identified and eliminated. A "No Call" implies that a valid call could not be made for that specific genotype. A user can define "No Calls" based on a threshold placed on the GenCall Scores. For a fixed threshold value, the application computes the strand-to-strand correlation and the call rate. Call rate is defined as the ratio of number of genotypes exceeding the threshold value to the total number of genotypes.

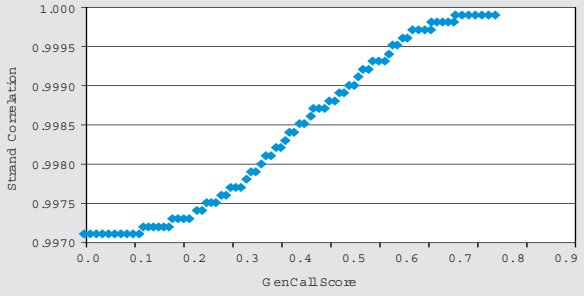
$$CallRate = \frac{NumGenotypes\ (score \geq threshold)}{NumGenotypes}$$

The strand-to-strand correlation is a proxy for the accuracy (hereinafter referred to as accuracy) and is defined as the ratio of the number of matching alleles to the total number of alleles.

$$Accuracy = \frac{NumAlleles\ (Matched)}{NumAlleles}$$

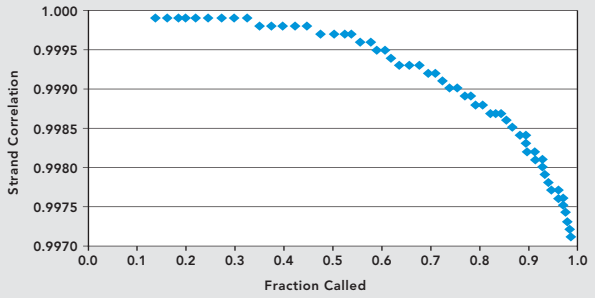
As the threshold value is changed, different measures of accuracy and call rate are obtained, resulting in a curve. Figure 2 illustrates a strand correlation experiment evaluating the effectiveness of the GenCall Score. Each point represents strand correlation coefficient at a particular threshold applied by the GenCall Score. As one increases the threshold, the call rate decreases, while the accuracy increases. This increase of accuracy (as a function of threshold) demonstrates the informative power of the GenCall Score.

FIGURE 1: RELATIONSHIP BETWEEN STRAND CORRELATION (ACCURACY) AND GENCALL SCORE



Correlation of assay quality, as estimated by GenCall Score, with cumulative accuracy (strand correlation) of SNP calls. The data are progressively excluded by GenCall Score, beginning at a GenCall Score of zero where all the data are evaluated. A total of 355 DNA's were assayed on both strands at 288 SNP loci (408,960 total allele calls). Accuracy is measured by the correlation of alleles determined on the two DNA strands.

FIGURE 2: RELATIONSHIP BETWEEN STRAND CORRELATION (ACCURACY) AND CALL RATE



The same set of 408,960 allele calls from Figure 1 is evaluated for call rate (fraction call) and accuracy (strand correlation). Each point represents the mentioned values at a particular threshold applied to GenCall Score. For this study, one can select a call rate of 99% to obtain 99.7% accuracy, whereas a call rate of 90% yields 99.8% accuracy.

ADDITIONAL INFORMATION

To learn more about Illumina technology, products, and services, visit our website or contact us at the address below:

Illumina, Inc.
Customer Solutions
9885 Towne Centre Drive
San Diego, CA 92121-1975
1.800.809.4566 (toll free)
1.858.202.4566 (outside the U.S.)
techsupport@illumina.com
www.illumina.com

REFERENCES

(1) Oliphant, A, Barker, D.L., Stuelpnagel, J.R., Chee, M.S. BeadArray™ Technology: Enabling an Accurate, Cost-Effective Approach to High-Throughput Genotyping. *SNPs; Discovery of Markers for Disease (supplement to Biotechniques)*, June 2002.

(2) Fan, J.B., Oliphant, A., Shen, R., Kermani, B.G., Garcia, F., Gunderson, K.L., Hansen, M.S., Steemers, F., Butler, S.L., Deloukas, P., Galver, L., Hunt, S., McBride, C., Bibikova, M., Rubano, T., Chen, J., Wickham, E., Doucet, D., Chang, W., Campbell, D., Zhang, B., Kruglyak, S., Bentley, D., Haas, J., Rigault, P., Zhou, L., Stuelpnagel, J. and Chee, M.S., Highly Parallel SNP Genotyping. *Cold Spring Harbor Symposia on Quantitative Biology, Volume LXVIII*, 69-78, January 2004. © 2003 Cold Spring Harbor Laboratory Press.

FOR RESEARCH USE ONLY

illumina®
making sense out of life