EM algorithm for gene copy number estimation using TaqMan® assays

Catalin Barbacioru, Kelly Li, Caifu Chen, and Raymond Samaha
Applied Biosystems, 850 Lincoln Centre Dr., Foster City, CA 94404

Abstract

Multiple studies have discovered an abundance of submicroscopic copy number variation of DNA segments ranging from kilobases (kb) to megabases (Mb) in size [1,2]. Recently, TaqMan® assays have been developed for detection of genetic variation at gene level using primers and probes designed for genomic DNA sequences. In this study, we present an algorithm for gene copy number estimation from TaqMan® assays based on EM algorithm for mixtures of normal distributions.

Introduction

Recently, TaqMan® Gene Copy Number Assays have been developed for detection of genetic variation at gene level using primers and probes designed for genomic DNA sequences. Each well is duplexed with two assays. The FAM™ dye-based assay is designed to detect the genes-of-interest and the VIC® dye-based assay is for the reference gene, RNase P. The difference between FAM and VIC measurements (dCT) is indicative of the relative abundance of the gene-of-interest against 2 copies per diploid genome regardless of the status of the gene-of-interest. In this study, we present an algorithm for gene copy number estimation from TaqMan® Gene Copy Number Assays based on EM algorithm for mixtures of normal distributions [3].

Statistical considerations

There are 3 independent types of errors in the CT measurements: (1) Sample error: the same in replicated wells of a given sample, and the same in FAM and VIC measurements; (2) Pipetting error: the same in FAM and VIC measurement from the same well; (3) Technical error: independent in FAM and VIC. Therefore, the FAM and VIC measurements from sample i = 1, . . . ,M, having n copies of the gene-of-interest, for replicate j = 1, . . . ,R_i are

$$CT_{FAM,i,j} = \mu_{FAM} + C_n + S_i + P_{i,j} + e_{i,j,FAM}$$
$$CT_{VIC,i,j} = \mu_{VIC} + S_i + P_{i,j} + e_{i,j,VIC}$$
$$\Delta CT_{i,j} = \mu + C_n + e_{i,j}$$

where S is sample error, P is pipetting error, e is technical error, μ_{FAM}, μ_{VIC} are CT measurements for 2 copies samples of the gene of interest and RNaseP respectively, C_n is the relative difference to a 2 copies sample, $\mu = \mu_{FAM} - \mu_{VIC}$, and $e = e_{FAM} - e_{VIC}$. For simplicity we can assume that the two assays have the same efficiency, and therefore $C_n = \log_2(n) - 1$. Assuming that technical errors are independent and normally distributed, then ΔCT measurements can be modeled as a mixture of normal distributions.

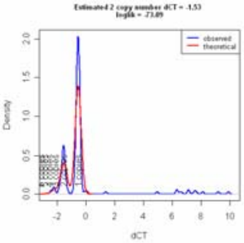
Algorithm

For each sample the difference between ΔCT of individual wells and the sample average ΔCT represents the technical error and it comes from the same normal distribution for all samples. Therefore these measurements can be used to estimate the variance of the technical error. Using this estimate, outliers are identified (95% confidence) and removed from further analysis. The parameter set $\Theta = (\mu, \sigma^2, \pi_0, . . . , \pi_N)$, contains the mean dCT value for two copies samples, technical variation of the instrument, and the mixture coefficients. The maximum-likelihood estimates (MLEs) of the parameter set Θ, are obtained using the Expectation Maximization algorithm (EM) for mixture of normal distributions [3]. The model depends on unobserved samples copy number of the gene-of-interest.

Results

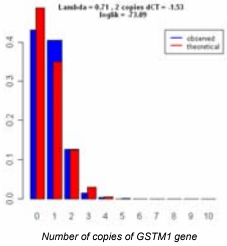
The algorithm was tested on 270 individual samples from International HAPMAP Project representing 4 different populations, 15 samples being duplicated. Five important drug metabolism genes, CYP2D6, CYP2E1, CYP2A6, GSTM1 and GSTT1, with four replicates on different plates were run for each assay.

Figure 1. Estimated model using EM algorithm



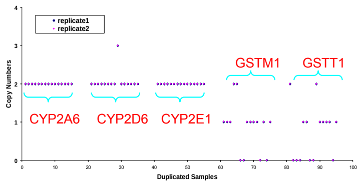
Multimodal behavior of observed dCT distribution (blue) indicates differences in copy number of GSTM1 across samples. EM algorithm estimates mean dCT for different copy number samples and technical variance. EM estimates of the parameters are used to represent (red) the model chosen for the data.

Figure 2. Copy number distribution



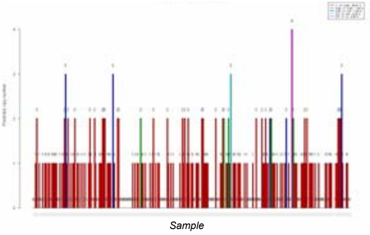
For each of the 270 individual samples the number of copies of GSTM1 is predicted. Distribution of predicted copy number (blue) is presented against Poisson distribution (red) of parameter λ = total number of copies / 270. Such similarities were observed in different populations for several assays.

Figure 3. Sample Duplication Consistency



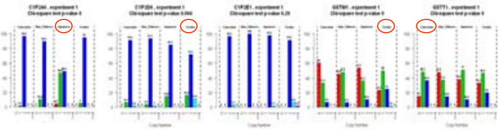
Fifteen samples, covering all 4 populations, from the 270 unique individuals are spotted twice on each plate. Sample duplicate consistency of copy number call is evaluated for each of the 5 assays. Perfect consistency of sample duplicates is observed in all the population and copy numbers (0, 1, 2, and 3).

Figure 4. Predicted copy number and prediction confidence



GSTM1 gene is used as an example for copy number prediction of individual samples. For each prediction, confidence level is calculated from the mixture model estimated by EM algorithm. In this case samples are predicted to have between zero and four copies of GSTM1 gene, with a confidence of at least 99% for 95% of the samples.

Figure 5. Copy Number Frequencies in Each Population



Copy number frequencies in all four populations for all five genes are calculated. All five genes except CYP2E1 show significant differences in copy number frequencies among the four populations when Chi-square test is used. Red circles indicate the population that is different from others.

Under current protocols, we are capable of distinguishing up to 8 copies of the gene of interest with at least 95% confidence, assuming 100% efficiency of the FAM™ dye-based assay, when 4 sample replicates are used.

CONCLUSIONS

In this study, we present an algorithm for gene copy number estimation from TaqMan® Gene Copy Number Assays based on EM algorithm for mixtures of normal distributions. The copy number analysis for these genes show perfect consistency for sample duplicates. Copy number variation (from 0 to 4) is observed for all 5 genes. Significant differences between population are revealed.

This algorithm is implemented into the R package TaqGCN and will be released as part of Bioconductor.

REFERENCES

[1] Iafrate, A. J. *et al. Detection of large-scale variation in the human genome*, Nature Genet. 36, 949–951 (2004)

[2] Feuk, L., Carson, A. R. & Scherer, S. W. *Structural variation in the human genome*, Nature Rev. Genet. 7, 85–97 (2006)

[3] Robert Hogg, Joseph McKean and Allen Craig. *Introduction to Mathematical Statistics*, Upper Saddle River, NJ: Pearson Prentice Hall, 2005

ACKNOWLEDGEMENTS

We thank Kashif A. Haque and Robert A. Welch from Core Genotyping Facility, Division of Cancer Epidemiology and Genetics, SAIC Frederick, National Cancer Institute for providing the samples and their support.