Published online April 14, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 7 ***e56***
doi:10.1093/nar/gkl185

Algorithm for automatic genotype calling of single nucleotide polymorphisms using the full course of TaqMan real-time data

A. Callegaro[1], R. Spinelli[2,3], L. Beltrame[2,3], S. Bicciato[1], L. Caristina[3], S. Censuales[4], G. De Bellis[3,5] and C. Battaglia[2,3,*]

[1]Department of Chemical Process Engineering, University of Padua, Padua, Italy, [2]Department of Sciences and Biomedical Technologies, University of Milan, Milan, Italy, [3]Array Technology Group, Interdisciplinary Center for Biomolecular Studies and Industrial Applications (CISI), University of Milan, Milan, Italy, [4]Centro Trasfusionale, Sacco Hospital, University of Milan, Milan, Italy and [5]Institute for Biomedical Technologies, National Research Council (CNR), Milan, Italy

Received October 1, 2005; Revised November 30, 2005; Accepted March 24, 2006

ABSTRACT

Single nucleotide polymorphisms (SNPs) are often determined using TaqMan real-time PCR assays (Applied Biosystems) and commercial software that assigns genotypes based on reporter probe signals at the end of amplification. Limitations to the large-scale application of this approach include the need for positive controls or operator intervention to set signal thresholds when one allele is rare. In the interest of optimizing real-time PCR genotyping, we developed an algorithm for automatic genotype calling based on the full course of real-time PCR data. Best cycle genotyping algorithm (BCGA), written in the open source language R, is based on the assumptions that classification depends on the time (cycle) of amplification and that it is possible to identify a best discriminating cycle for each SNP assay. The algorithm is unique in that it classifies samples according to the behavior of blanks (no DNA samples), which cluster with heterozygous samples. This method of classification eliminates the need for positive controls and permits accurate genotyping even in the absence of a genotype class, for example when one allele is rare. Here, we describe the algorithm and test its validity, compared to the standard end-point method and to DNA sequencing.

INTRODUCTION

The most common type of genetic diversity in the human genome is the single nucleotide polymorphism (SNP), which accounts for >90% of all sequence polymorphisms (1,2). SNPs have revolutionized human molecular genetics by providing a dense panel of genetic markers distributed across the entire genome. Although most SNPs do not affect gene function, they can be used to study population dynamics and evolution, to investigate the genetic basis of complex phenotypes and to develop diagnostic assays.

A number of techniques are now available for rapid SNP genotyping. A key requirement of a SNP genotyping method is that it distinguishes unequivocally between the allelic variants present in homozygous and heterozygous forms. The choice of technology depends on whether a few SNPs are to be screened in many individuals or many different SNPs are to be examined in a few individuals (3,4). Miniaturized assays, such as microarrays with oligonucleotide reagents immobilized on small surfaces, are frequently proposed for large-scale mutation analysis and high-throughput genotyping (5). Other high-throughput methods discriminate alleles by differential hybridization, primer extension, ligation and cleavage of an allele-specific probe (6,7).

A promising approach for a fully automated, large-scale SNP analysis is the 'homogeneous' assay, i.e. a single-phase assay without separation steps, permitting continual monitoring during amplification. The TaqMan assay (Applied Biosystems), originally designed for quantitative real-time PCR, is a homogeneous, single-step assay also used in determination of mutation status of DNA. The TaqMan SNP Genotyping Assay exploits the 5′-exonuclease activity of AmpliTaq Gold DNA polymerase to cleave a doubly

*To whom correspondence should be addressed at Array Technology Group (ATG), Department of Biomedical Sciences and Technologies and CISI, University of Milan, Milan, Italy. Tel: +39 02 50330425; Fax: +39 02 50330414: Email: cristina.battaglia@unimi.it

labeled probe hybridized to the SNP-containing sequence of ssDNA (1,8). Cleavage separates a 5′-fluorophore from a 3′-quencher (9) leading to detectable fluorescent signal. The use of two allele-specific probes carrying different fluorophores permits SNP determination in the same tube without any post-PCR processing. Genotype is determined from the ratio of intensities of the two fluorescent probes at the end of amplification. Thus, rather than taking advantage of the full set of real-time PCR data as in quantitative studies, only end-point data are used.

TaqMan SNP genotyping in a high-throughput, automated manner is facilitated by the use of validated Pre-made TaqMan® Genotyping assays, but Custom TaqMan® Assays may also be used (10,11). The results of the assay can be automatically determined by genotyping software provided with real-time thermal cyclers (e.g. IQ software of Bio-Rad, Sequence Detection Software of Applied Biosystems). Since these programs use autoscaling for generating the allele discriminating plot, caution must be taken when analyzing data for rare alleles. Therefore, accurate allele-calling requires manual intervention of an expert operator who must assess data quality, set fluorescent signal thresholds and decide the genotype. To avoid this problem, positive controls should be included for all genotypes (12). In alternative to analyzing end-point fluorescent data using commercial software or visual discrimination, a statistical clustering method called k-means can be used (10).

In the interest of optimizing real-time PCR genotyping, we have been investigating novel methods to analyze the fluorescent signals. In particular, we looked for data clustering algorithms to analyze the entire real-time PCR dataset, as well as statistical approaches to analyze these data clusters. Complex statistical methods have already been used to analyze large datasets from genotyping experiments. For example, Liu *et al*. (13) described the use of modified partitioning around medoids (MPAM) clustering algorithm to analyze high-density microarray data on SNPs, and Lovmar *et al*. (14) reported that tag-array minisequencing microarray data clusters can be distinguished on the basis of silhouette score. Here, we describe an algorithm for automatic genotype calling based on real-time PCR data and test its validity, compared to the standard end-point method and to DNA sequence data.

MATERIALS AND METHODS

Peripheral blood was obtained from healthy blood donors (Sacco Hospital, Milan, Italy). Genomic DNA was extracted using QIAmp DNA Blood Mini Kit (Qiagen, Milan, Italy). DNA concentration was measured by absorbance at 260 nm using an ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

Real-time PCR SNP genotyping

Samples of genomic DNA were genotyped for 13 SNPs in 12 genes of major medical interest (Table 1). Genotyping was performed using Custom TaqMan® Assays for 5 SNPs and commercially available Pre-made TaqMan® Genotyping assays for 8 SNPs. For custom assays, File Builder software was used to select PCR primers and reporter probes labeled with a quencher, a minor groove binder (MGB) and either

Table 1. SNPs employed to test the BCGA using TaqMan real-time PCR

Polymorphism	dbSNP ID	Location	Chromosome	Chromosome position (bp)	Allele	Allelic frequency
ApoE −219	rs405509	Promoter	19	50100676	T	0.51
					G	0.49
IL-1 alpha −889	rs1800587	Promoter	2	113259191	C	0.70
					T	0.30
IL-1 beta −511	rs16944	Promoter	2	113311098	G	0.53
					A	0.47
IL-6 −174	rs1800795	Promoter	7	22539885	G	0.77
					C	0.23
TNF alpha −238	rs361525	Promoter	6	31651080	G	0.92
					A	0.08
TNF alpha −308	rs1800629	Promoter	6	31651010	G	0.91
					A	0.09
IFN-γ	rs2069727	(3′-UTR)	12	66834490	A	0.68
					G	0.32
CST3	rs2424574	Intron	20	23556980	C	0.51
					T	0.49
CACNA1D	rs723540	Intron	3	53730993	G	0.60
					C	0.40
AHRGEF3[a]	rs1500711	Intergenic region	3	56889194	T	0.66
					C	0.34
SERPINA3	rs4934	Coding region	14	94150556	G	0.54
					A	0.46
TF	rs3811656	Coding region	3	134957026	G	0.70
					A	0.30
LRRC2	rs1402152	Intron	3	46549422	T	0.67
					C	0.33

Data are from dbSNP database (http://www.ncbi.nih.gov/SNP).
[a]Since this SNP is intergenic, the closest gene in indicated.

6-carboxyfluorescein (FAM) or VIC (Applied Biosystems' propriety dye with $\lambda_{ex} = 488$ nm, $\lambda_{em} = 552$ nm); these sequences are reported in Supplementary Table 1. The specific Pre-made TaqMan® Genotyping assays used in this study were: tumor necrosis factor (TNF) alpha −238 (Applied Biosystems code C___2215707_10), IFN-γ (C___2683475_10), CST3 (C___9714066_10), CACNA1D (C___1692260_10), AHRGEF3 (C___1482133_1), SERPINA3 (C___2188895_10), TF (C___80379_10) and LRRC2 (C___115203_10).

Real-time TaqMan PCR was performed according to the manufacturer's standard PCR protocol. Briefly, 10 ng total DNA was mixed with the supplied 2× TaqMan Universal PCR Master Mix No AmpErase UNG and TaqMan Assay Mix to a final volume of 25 µl. Each sample underwent 45 amplification cycles on an iCycler real-time thermal cycler (Bio-Rad Laboratories, Hercules, CA). Fluorescent signals of the two probes were monitored throughout the entire amplification using iCycler IQ software (Bio-Rad). For the first 6 SNPs listed in Table 1, we tested a set of 32 samples and 2–3 blanks (no DNA); for each of the remaining 7 SNPs, we tested 12–48 different samples and one blank.

Allelic calls were first determined semi-automatically with the aid of IQ software. Briefly, the plots of fluorescent intensities per cycle for each reporter fluorophore were visually inspected to choose a baseline level, which was subtracted from each data point. The end-point of each normalized dataset was defined as cycle number 40, as suggested by the manufacturer. End-point fluorescent intensities of each probe were plotted in an allelic discrimination graph (VIC on abscissa, FAM on ordinate), and genomic 'clusters' were defined manually by sectioning the plots into quadrants with horizontal and vertical lines.

Best cycle genotyping algorithm (BCGA)

Genotype calling from real-time PCR data were performed using an algorithm, called BCGA, written in the open source language R (R Foundation for Statistical Computing, Vienna, Austria). BCGA is composed of five main steps (Figure 1):

(i) *Selection of best cycle*. The genotype analysis is based on the hypothesis that, during amplification, one cycle is most discriminating, not necessarily the end-point. Thus, we calculated the differences in probe intensities (VIC-FAM) at every cycle, and chose the cycle that had the greatest variance in intensity differences among samples (best cycle).

(ii) *Data clustering*. Working with data from the best cycle chosen in step 1, samples were grouped into clusters based on the fluorescent signal differences. This classification procedure was performed using partitioning around medoids (PAM) algorithm (15), which is included in the R Cluster Package. For SNP genotyping, 1–3 clusters are expected.

(iii) *Choice of best clustering model*. The best number of clusters for a particular SNP dataset, including the blanks, was chosen on the basis of average silhouette width (16), determined using R Cluster Package. Silhouette width, s(i), is defined as the average dissimilarity between a sample and all other specimens of its class, compared to all observations in the neighboring clusters. The best number of clusters is determined on the basis of the highest average silhouette width.

(iv) *Genotype calling*. Assignment of generic genotypes (AA, AB, BB) takes into consideration the behavior of blanks during the PCR time course. We observed that the difference in fluorescent signal (VIC-FAM) over time was relatively stable and near zero for blanks as well as for known heterozygotes. Thus, the algorithm identifies the heterozygous cluster as the one containing the blanks. Homozygous clusters are then defined on the basis of the predominance of one or the other fluorescent signal. This classification procedure, based on the behavior of blanks, can be used when one genotype class is absent from a set of samples, for example when one allele is rare. In principle, the method can also be applied in the presence of a unique homozygous genotype (e.g. a single cluster distinct from the one formed by blanks).

(v) *Quality control*. The quality of assignment of individual samples to clusters was determined on the basis of silhouette values (16). Individual samples with a large s(i), i.e. close to 1, are well clustered, with a small s(i) lie between two clusters, and with a negative s(i) are misclassified. Furthermore, the quality of the genotype assignments to clusters was determined from the average silhouette width over all samples: the larger the silhouette width, the higher quality the PAM classification.

The BCGA code and a help file are publicly available at http://www.dpci.unipd.it/Bioeng/Publications/BCGA.htm. Moreover, this page provides access to seven datasets that may be used to test the algorithm and the corresponding results in graphic format.

Validation of BCGA

BCGA genotype calls were validated by DNA sequencing of PCR-amplified DNA for the first 6 SNPs listed in Table 1. Sequencing was not considered necessary for the remaining seven validated Pre-made TaqMan® Genotyping assays; in these cases, we considered the genotype calls of the standard end-point procedure to be the reference value.

In the validation study, we determined the actual allelic compositions of 14 samples chosen on the basis of BCGA genotype calls in order to maximize the chances of having all alleles represented in 6 SNPs genotyping assays. To this end, we designed PCR primer sets to amplify the DNA containing 6 SNPs (Supplementary Table 2). For TNF alpha, the PCR primers permitted sequence validation also of the SNP at position −238, assayed with a Pre-made TaqMan® Genotyping assay. Forward and reverse primers were synthesized by Thermo Hybaid (Ulm, Germany). PCR was conducted in a final volume of 12.5 µl containing 200 ng genomic DNA, 1× PCR buffer, 1.2 mM dNTPs, 20 pmol each primer (one reaction per primer set) and 0.2 U Platinum *Taq* DNA Polymerase High Fidelity (Invitrogen, Carlsbad, CA). After an initial denaturation step at 95°C, amplification was carried out in 35 cycles consisting of 30 s at 94°C, 1 min at 55°C and 1 min at 68°C. A final, 7 min extension at 68°C ended each reaction.

Quality of PCR products was assessed by microcapillary electrophoresis (2100 Bioanalyzer, Agilent Technologies, Palo Alto, CA). PCR products were purified using the ExoSAP

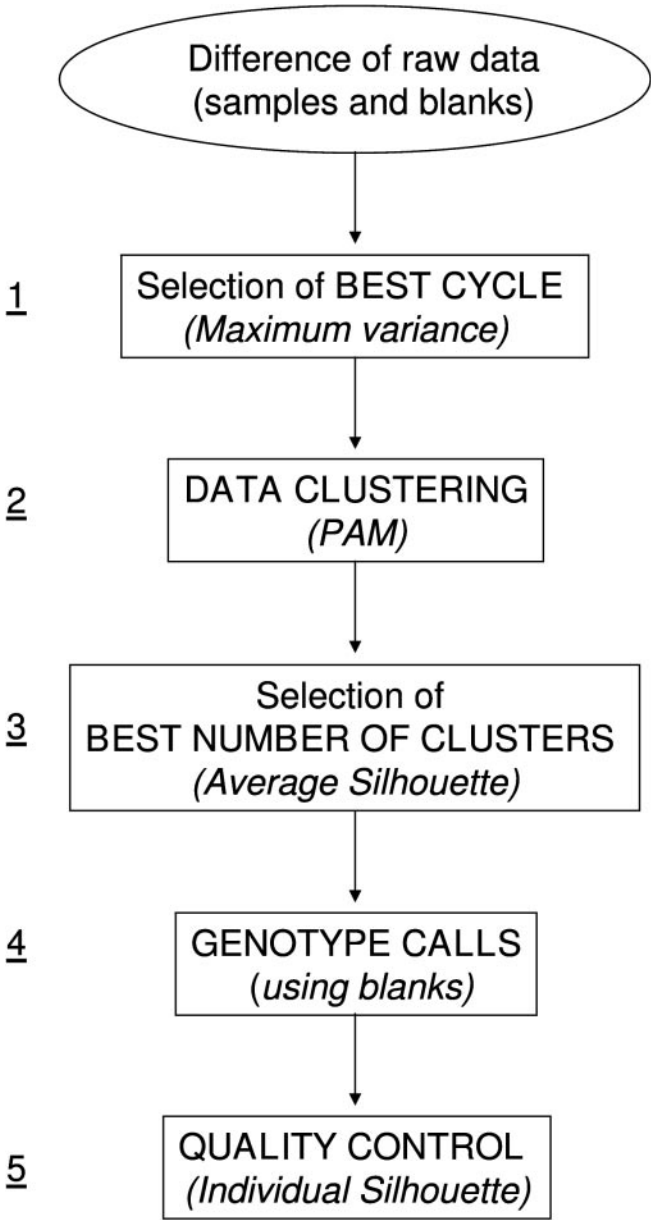

Figure 1. Five major steps of the novel BCGA. *PAM*, partitioning around medoids.

method (GE Healthcare, formerly Amersham Biosciences, Uppsala, Sweden) and sequenced with the DYEnamic ET Dye Terminator Kit (GE Healthcare) according to the manufacturer's instructions. Sequencing runs were performed on the MegaBACE 100 DNA Analysis System (GE Healthcare) according to the manufacturer's instructions, with injections at 3 kV for 40 s and sequencing runs at 6 kV for 180 min. The run results were analyzed using Sequence Analyzer software (GE Healthcare). Sequencing was carried out on forward and reverse strands.

RESULTS

The genotypes of healthy blood donors, regarding 13 SNPs in 12 genes, were assessed by a standard analysis of TaqMan end-point fluorescent data and by a novel BCGA. At standard end-point analysis of 32 subjects (Figure 2), clear identification of three tight clusters permitted easy assignment of allele-1 (AA), allele-2 (BB) and heterozygous (AB) genotypes for 3 SNPs: apolipoprotein E −219, interleukin (IL)-1 beta −511 and IL-6 −174. Only two clusters were observed for TNF alpha −238, while one broad cluster was observed for TNF alpha −308. In the analysis of IL-1 alpha −889, three genotype clusters were identified, but two samples (samples 17 and 22) in the heterozygote quadrant showed unique behaviors, casting doubts on the accuracy of the genotype assignment. Similar patterns were observed for the seven other SNPs studied (Supplementary Figure 1). These data exemplify the difficulties encountered in end-point PCR analysis of SNPs, precluding the use of automatic genotype calling software and necessitating operator intervention.

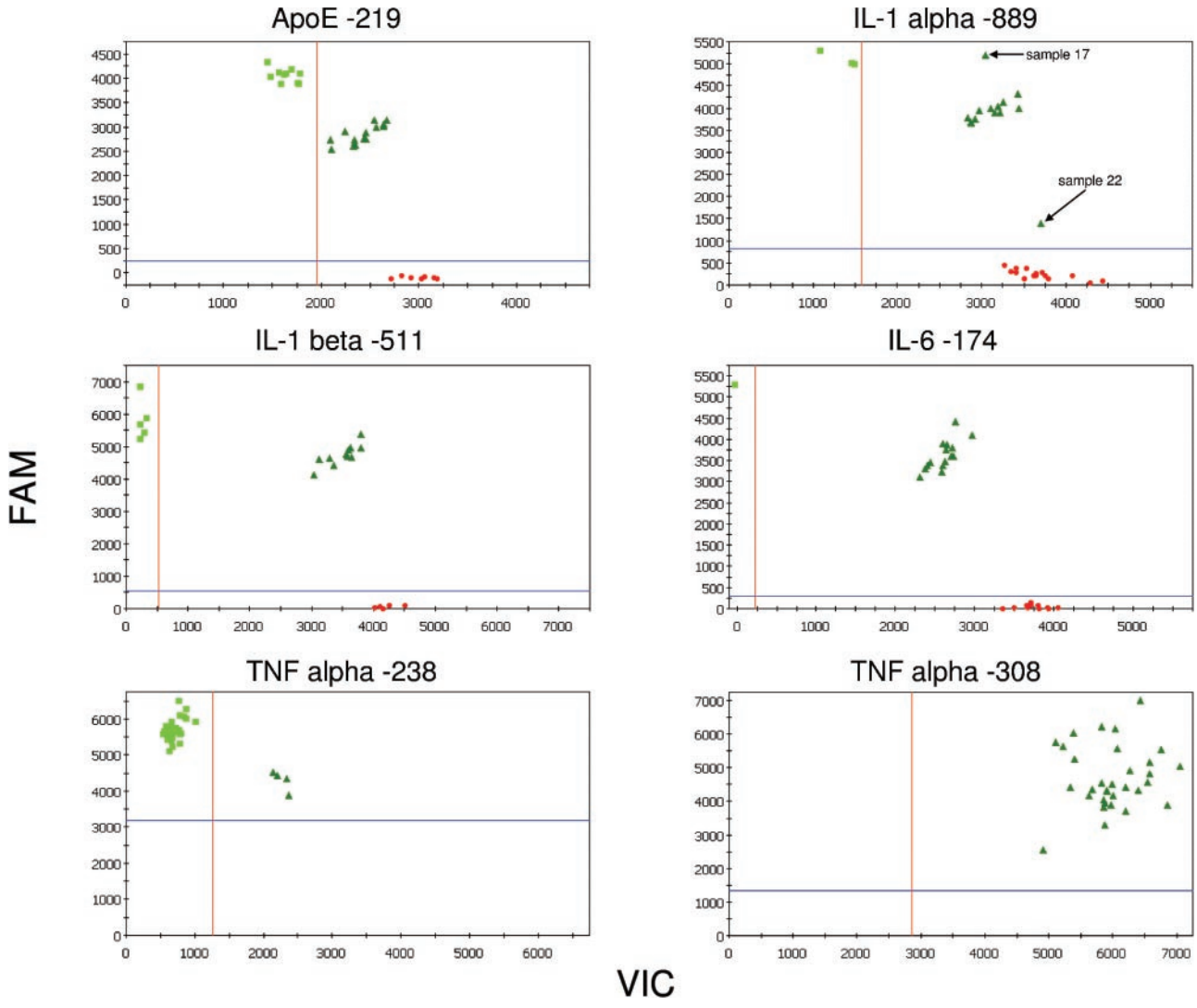

Figure 2. Allelic discrimination plots of end-point fluorescent TaqMan PCR data of 6 SNPs analyzed in 32 healthy blood donors. VIC fluorophore (*x*-axis) is associated with the probe for allele A, while FAM (*y*-axis) labels the allele B probe. The plots were generated using iCycler IQ software (Bio-Rad, Hercules, CA). For IL-1 alpha −889, outlying samples 17 and 22 are marked.

While the standard genotype analysis is based on the fluorescent intensities of the two reporter probes at an arbitrarily chosen PCR cycle near the end of amplification (usually cycle 40), our algorithm considers the time course of fluorescent signals throughout the amplification. Plots of FAM versus VIC fluorescent intensities (not baseline-subtracted) at each amplification cycle for 32 samples and 1–3 blanks (Figure 3) show the unique time courses of the signals: the probes may be easily distinguished at all time points, or their signals may first diverge and then converge toward the diagonal (e.g. TNF alpha −308). Nonetheless, for this latter SNP, the temporal traces permit distinction of clusters at earlier cycles. Of the two outlying samples seen at end-point analysis of IL-1 alpha −889, BCGA assigned sample 22 to the allele-1 cluster and sample 17 to the heterozygous cluster, although the fluorescent trace of 17 remained distinct (Figure 3). In all cases, blank samples (no DNA) had fluorescent signals lower than that of the first cycles of experimental samples and showed no relevant time trend in signal during amplification. Thus, blanks are all located in the lower left-hand corner of each graph. Similar results were observed for the remaining 7 SNPs (Supplementary Figure 2).

In BCGA, samples are classified according to the differences in their reporter probe signals. During the course of PCR (Figure 4 and Supplementary Figure 3), this difference for blanks is relatively stable and near zero because there is no probe degradation. This difference also remains near zero for heterozygous samples because the two probes are degraded to a similar extent. Therefore, classification according to signal differences places blanks and heterozygotes in the same cluster, and permits BCGA to begin genotype calling by identifying the heterozygous cluster as that containing the blanks.

BCGA is based on the hypothesis that best allelic discrimination is not necessarily at the end of PCR amplification. Thus, for each SNP analysis, we determined the PCR cycle with the greatest variance in the difference between the two fluorescent probe signals (best cycle). The best cycle for

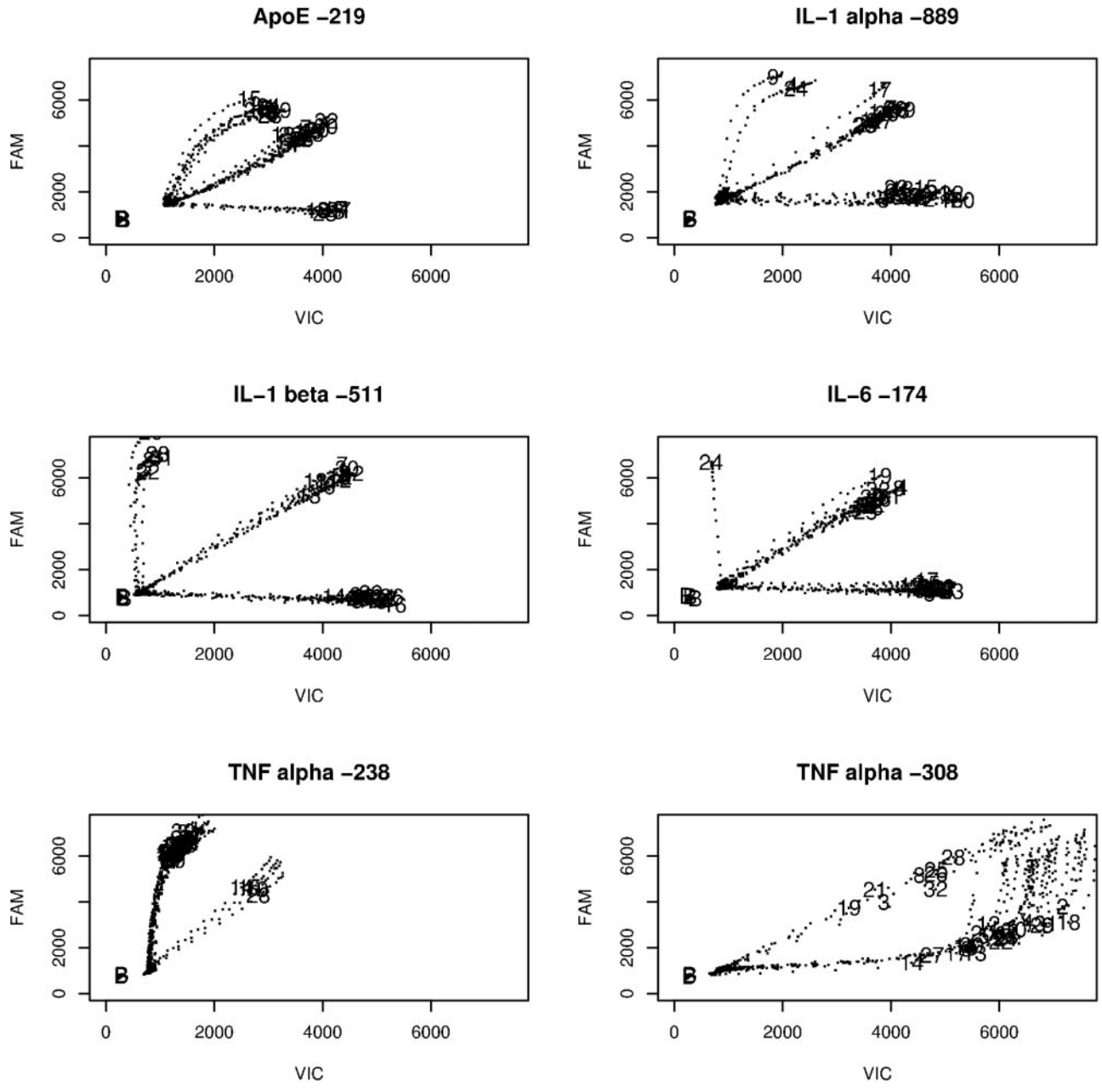

Figure 3. Real-time PCR traces of VIC- and FAM-labeled reported probe intensities at each cycle, for 6 SNPs analyzed in 32 subjects. Traces are labeled with sample numbers at the best cycle. Blank samples (B) are clustered at the lower left-hand corner of each plot (low fluorescent signal intensity without temporal changes).

13 SNPs ranged between 34 and 44 (Table 2). These data confirm our hypothesis that an arbitrarily chosen PCR cycle (e.g. 40) is not necessarily the best for distinguishing TaqMan reporter probe signals.

Using reporter probe signals at the best cycle, we employed PAM to group the samples and blanks into 2 or 3 clusters. The best classification model was determined on basis of the highest average silhouette width for all samples, including blanks. The best classification contained two clusters for 3 SNPs, while a three cluster model was chosen for 10 SNPs (Table 2). For IL-6 −174, one cluster contained a single sample. Average silhouette width ranged from 0.71 (AHRGEF3) to 0.93 (SERPINA3). Silhouette bar plots of individual s(i) values illustrate the classification of samples and blanks into two or three clusters (Figure 5 and Supplementary Figure 4); for the cluster containing only one sample, calculation of silhouette width was mathematically not possible.

To make the genotype calls, we first identified the heterozygous cluster as the one containing the blanks, given that differences in reporter probe signals for heterozygotes are comparable with those for blank samples, irrespective of the actual signal intensities. When multiple blanks were used, they were always assigned to the same cluster (Figure 5). The genotype calls of the other 1–2 clusters per SNP analysis were made on the basis of the predominating reporter probe at the best PCR cycle. The number of samples for each cluster (excluding blanks) and the specific genotype

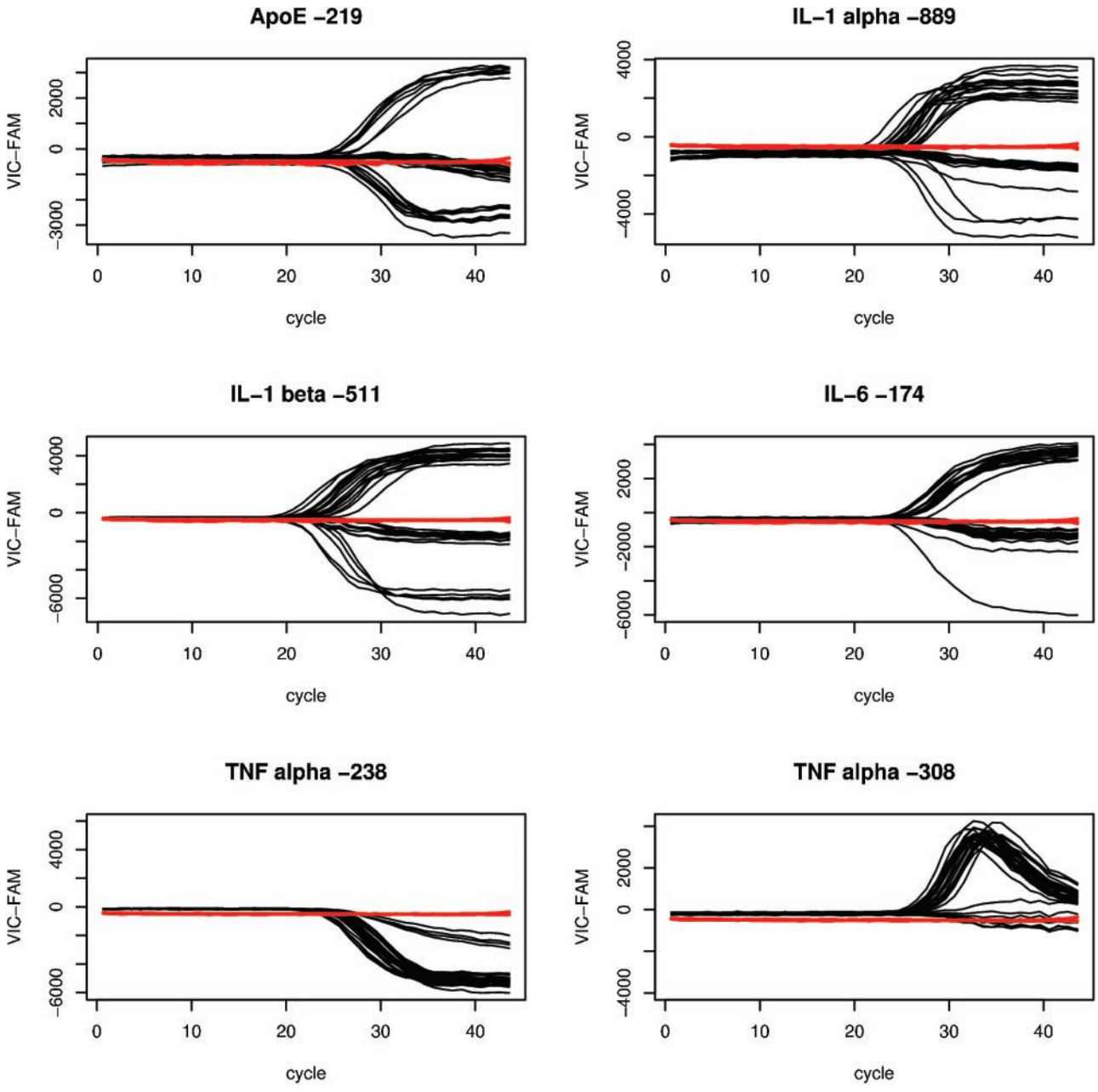

Figure 4. Real-time PCR traces of differences in reporter probe signals, for 6 SNPs analyzed in 32 subjects, demonstrate the co-segregation of blanks with heterozygous samples. Traces for blank samples are shown in red.

assignments to each cluster are indicated in Figure 5 and Supplementary Figure 4.

Quality of assignment of individual samples to clusters was determined on the basis of silhouette values. Almost all samples had s(i) > 0.65, indicating good classification. One exception was sample 17, in the analysis of IL-1 alpha −889, which had s(i) = 0.299; this sample was considered an outlier at end-point analysis (Figure 2) and was visually distinct from the heterozygous cluster in the plot of fluorescent traces (Figure 3). Poor classification was also observed in AHRGEF3 assay for four samples with s(i) values between 0.36 and 0.61 (Supplementary Figure 4). These samples did not form tight clusters at end-point analysis (Supplementary Figure 1). Finally, for sample 24 in IL-6 −174 analysis, a value of s(i) could not be determined because it was the only sample in its cluster.

The accuracy of genotype calls made by standard end-point analysis and by BCGA was determined by direct sequencing of PCR-amplified genomic DNA for 14 selected subjects and for the first 6 SNPs, including all 5 SNPs tested with Custom TaqMan® assays. Overall, end-point analysis with operator intervention correctly called 72 of 84 analyses (6 SNPs in 14 subjects), while BCGA correctly called all 84 analyses (Table 3). Most errors during standard analysis regarded TNF alpha −308. For the remaining 7 SNPs, there was exact agreement in genotype call between end-point analysis

Table 2. Best cycle, cluster number and average silhouette width for 13 SNPs assayed with BCGA (including blanks)

Polymorphism	Best cycle	Clusters, n	Average s(i)
ApoE −219	42	3	0.87
IL-1 alpha −889	38	3	0.83
IL-1 beta −511	43	3	0.89
IL-6 −174	44	3	0.88
TNF alpha −238	36	2	0.89
TNF alpha −308	34	2	0.91
IFN-γ	43	3	0.88
CST3	44	3	0.92
CACNA1D	44	2	0.91
AHRGEF3	44	3	0.71
SERPINA3	39	3	0.93
TF	44	3	0.91
LRRC2	44	3	0.91

and BCGA except for one of the poorly classified samples in AHRGEF3 assay (data not shown).

DISCUSSION

BCGA permits accurate genotype calling by taking advantage of the full course of real-time PCR data. BCGA is based on the assumptions that classification depends on the time (cycle) of amplification and that it is possible to identify a best discriminating cycle for each SNP assay. The algorithm is unique in that it classifies data clusters according to the behavior of blanks (no DNA samples), thereby eliminating the need for positive controls and permitting accurate genotyping even in the absence of a genotype class, for example when one allele is rare. The method can also be applied in the presence of a unique genotype, provided that is not the heterozygous one. Finally, the algorithm works

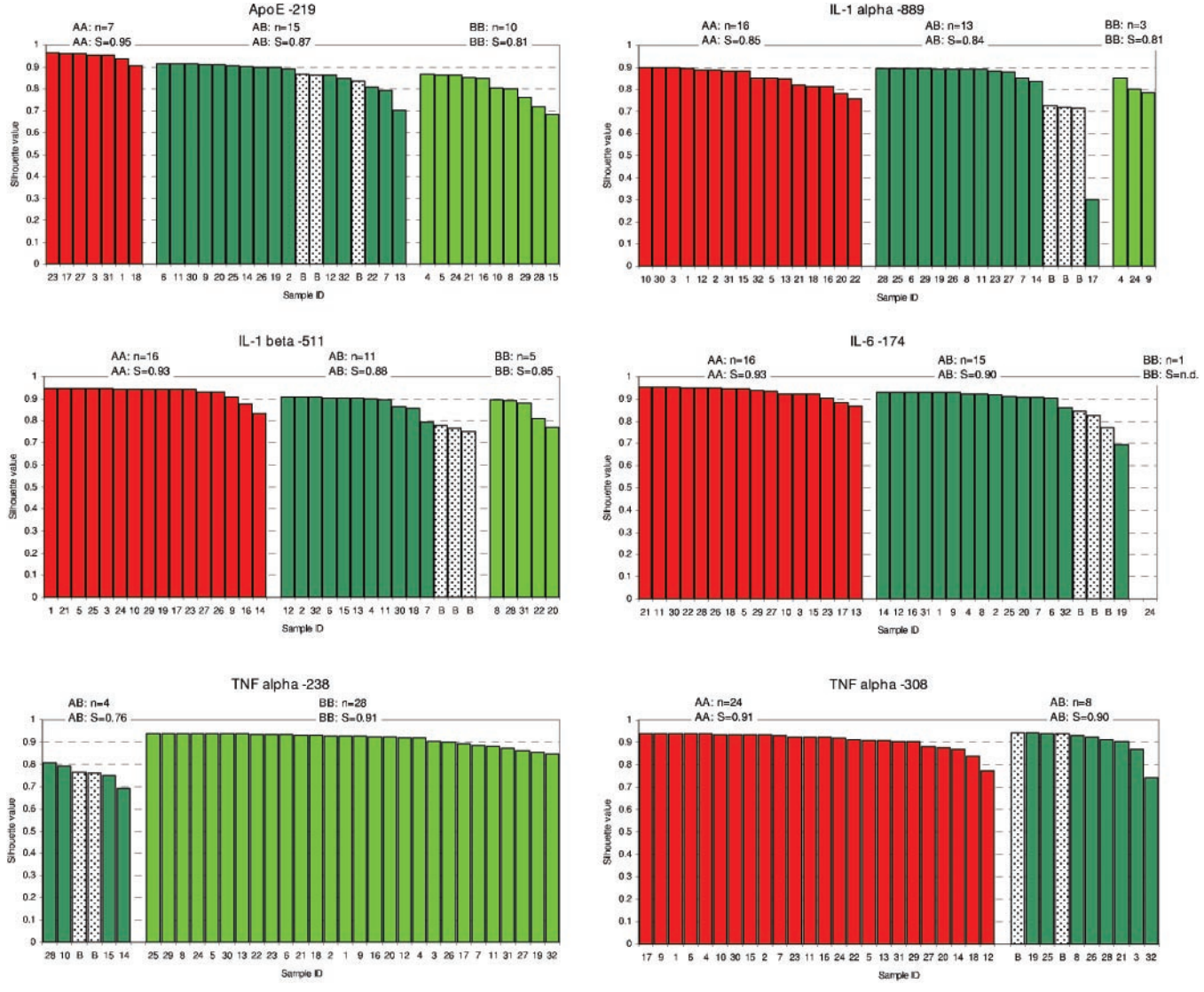

Figure 5. Silhouette width s(i) histograms for 32 samples and 2–3 blanks (speckled bars), classified according to partitioning around medoids (PAM) (13). For each SNP analyzed, the plots report the number of samples, n, assigned to each generic genotype cluster (excluding blanks), and the mean silhouette width, S, for each cluster.

Table 3. Genotype determination of 6 SNPs in 14 subjects, by direct sequencing of both strands, and comparison with results from end-point fluorescent PCR analysis and the novel BCGA (real-time fluorescent PCR)

SNP	Subjects, n			Correct calls, n	
	AA	AB	BB	End-point	BCGA
ApoE −219	3	4	7	14	14
IL-1 alpha −889	7	5	2	13	14
IL-1 beta −511	5	5	4	14	14
IL-6 −174	7	6	1[a]	14	14
TNF alpha −238	0	2	12	14	14
TNF alpha −308	11	3	0	3	14

[a]This subject corresponds to sample 24, the only BB homozygote found for this SNP; thus despite an incalculable silhouette width, the genotype calls at end-point and BCGA analysis were correct.

on raw fluorescent probe data and does not require normalization or background subtraction.

These characteristics of BCGA improve upon standard methods of SNP determination with TaqMan assays, in which semi-automated genotype assignments are based on the ratio between the fluorescent intensities of VIC and FAM after normalization. A high VIC/FAM ratio represents a homozygous allele-1 sample (cleaved VIC, intact FAM probes) and vice versa a low VIC/FAM ratio corresponds to a homozygous allele-2 sample; a ratio close to unity represents the heterozygotic state (17). This theoretical behavior is observed with validated Pre-made TaqMan® Genotyping assays, which are optimized in a multistep protocol including the selection of probes and primers lacking homology to other sequences in the human genome (11). To study SNPs for which validated assays are not available, *ad hoc* Custom TaqMan® assays may be requested, as we did in the present study. In these cases, even though the best probes and primers had been selected using the manufacturer's proprietary algorithm, it is possible to observe atypical behavior of a few samples (e.g. the outliers in IL-1 alpha −889) or of the entire population (e.g. TNF alpha −308). This non-classic behavior makes it difficult to establish universal thresholds for genotype assignments. Thus, to obtain valid information using the current end-point data analysis, it is necessary to optimize the structure of the allele discriminating probes (18,19) or to develop specific genotyping methods (20). However, the need for individual assay optimization hinders large-scale SNP screening, which requires universal methods and tools (3). A computational method like BCGA that lets investigators obtain accurate genotype information from Custom TaqMan® assays, even in light of the unique and sometimes problematic behavior of SNPs, represents a useful contribution to automatic SNP genotyping. A disadvantage of this approach using the full course of real-time data rather than end-point data are that a real-time PCR instrument is required rather than just a fluorescent reader.

Our hypothesis that during real-time PCR there is a best cycle in which to analyze data were confirmed by the traces of VIC versus FAM fluorescent intensities, which diverge up to three distinct profiles starting at about PCR cycle 25 and then tend to converge back towards the central, diagonal profile at later cycles. Whether the clusters reach the diagonal or simply turn inward depends on the number of PCR cycles performed and on characteristics of the individual SNPs. This behavior is explained by the release of both fluorophores due to non-specific degradation of both probes. Although standard amplification reactions are performed up to 35 cycles, the universal thermal cycling protocol of TaqMan Genotyping Assays recommends that real-time PCR be carried out to 45 cycles, with end-point reading typically at cycle 40. In our study, the best cycles were essentially the latest cycles before the phenomenon of non-specific probe degradation raised the signal-to-noise ratio; in the extreme case (TNF alpha −308), probe degradation resulted in a non-interpretable, single broad cluster at cycle 40. Selection of the best cycle allowed us to obtain good genotyping data even from a suboptimal assay.

BCGA is the first method to assign genotypes based on an analysis of the full course of TaqMan real-time data. An algorithm reported by Ranade *et al*. (10) used end-point data to assign samples to a predetermined number of groups; a quality score was then calculated for each genotype assuming a bivariate normal distribution of fluorescent values. BCGA makes use of PAM, a more robust version of k-means introduced by Kaufman and Rousseeuw (15), to cluster samples at the best discriminating cycle as determined by the analysis of the entire time course. Moreover, quality of PAM classification is assessed by the parameter silhouette width, developed for this purpose by one of the authors of PAM (16). Quality assessment according to silhouette width was also used by Liu *et al*. (13) and Lovmar *et al*. (14) in the analysis of microarray data. Both these groups observed that silhouette widths above 0.65 corresponded to visually good classification, and Lovmar *et al*. (14) proposed a more conservative cutoff of 0.7. Our use of silhouette to assess PAM classification is in agreement with these studies, because s(i) exceeded 0.65 for all analyses, with six exceptions. One sample in the analysis of IL-1 alpha −889 gave $s(i) = 0.299$, indicating poor classification; we attribute this low reading to experimental error. For another sample in the analysis of IL-6, s(i) was undetermined because it was the only sample in a cluster. Finally, four samples in the assay of AHRGEF3 had low s(i), possibly due to their anomalous, delayed signals during the course of PCR.

The few poorly classified samples in our study illustrate the advantages of using silhouette width, rather than visual inspection of plots, in data quality control. Mathematical evaluation of classification quality facilitates the identification of samples that were not accurately assessed and thus need to be retested, reducing the need for manual editing and the chance of human error. Moreover, the use of mathematical quality checking, as provided by silhouette width or probability score of k-means, may help identify samples with allelic imbalance due to DNA duplications. Allelic balance must be confirmed by quantifying copy number of each allele using other methods (21,22). A false suggestion of allelic imbalance may come from differences in the maximum fluorescent intensities of the two fluorophores upon release from reporter probes, possibly due to differences in their chemical environments. We observed this behavior for LRRC2 Pre-made TaqMan® Genotyping assays, for which the maximum fluorescent signal of one probe exceeded three times that of the other probe (Supplementary Figure 5). In such cases, and when at least two clusters are present, classification based on raw data may be incorrect and scaling is necessary. To handle

this problem, BCGA performs a quality check on the ratio of the probes' maximum fluorescent signals and flags situations where this ratio exceeds three; this permits the operator to intervene by applying an optional scaling step, which should be sufficient for sample classification. In this case, scaling permitted excellent classification—with s(i) > 0.8 for all samples—and genotype calls were in exact agreement with those of the standard method.

We believe that the BCGA is a valid computational method for analyzing TaqMan real-time PCR data, because it provides accurate genotype calls even for samples that exhibit atypical behavior during standard TaqMan genotype assay. BCGA is especially useful for SNP genotyping when the minor allele is rare or absent, since the clustering step does not require the presence of the three expected genotype classes. Moreover, the ability of the algorithm to identify the best discriminating cycle in which to assign genotypes overcomes the problematic behavior of some SNPs observed with Custom TaqMan® assays. Compared to currently available software, this open source algorithm is a substantive and important improvement in both the implementation and accuracy of SNP discrimination.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Valerie Matarese for editing and critically reviewing this manuscript. This work was supported by grants from the Italian Ministry of University and Research: COFIN 2002064481; FIRB RBNE01TZZ8 and by funds from Center for Biomolecular Studies and Industrial Applications (CISI), University of Milan. Funding to pay the Open Access publication charges for this article was provided by FIRB RBNE01TZZ8.

REFERENCES

1. Livak,K.J., Marmaro,J. and Todd,J.A. (1995) Towards fully automated genome-wide polymorphism screening. *Nature Genet*., **9**, 341–342.
2. Collins,F.S., Brooks,L.D. and Chakravarti,A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res*., **8**, 1229–1231.
3. Gut,I.G. (2001) Automation in genotyping of single nucleotide polymorphisms. *Hum. Mutat*., **17**, 475–492.
4. Syvanen,A.C. (2005) Toward genome-wide SNP genotyping. *Nature Genet*., **37**, S5–S10.
5. Wang,D.G., Fan,J.B., Siao,C.J., Berno,A., Young,P., Sapolsky,R., Ghandour,G., Perkins,N., Winchester,E., Spencer,J. *et al*. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.
6. Syvanen,A.C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Rev. Genet*., **2**, 930–942.
7. Twyman,R.M. and Primrose,S.B. (2003) Techniques patents for SNP genotyping. *Pharmacogenomics*, **4**, 67–79.
8. Lee,L.G., Connell,C.R. and Bloch,W. (1993) Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Res*., **21**, 3761–3766.
9. Kutyavin,I.V., Afonina,I.A., Mills,A., Gorn,V.V., Lukhtanov,E.A., Belousov,E.S., Singer,M.J., Walburger,D.K., Lokhov,S.G., Gall,A.A. *et al*. (2000) 3′-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Res*., **28**, 655–661.
10. Ranade,K., Chang,M.S., Ting,C.T., Pei,D., Hsiao,C.F., Olivier,M., Pesich,R., Hebert,J., Chen,Y.D., Dzau,V.J. *et al*. (2001) High-throughput genotyping with single nucleotide polymorphisms. *Genome Res*., **11**, 1262–1268.
11. De la Vega,F.M., Lazaruk,K.D., Rhodes,M.D. and Wenz,M.H. (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. *Mutat. Res*., **573**, 111–135.
12. Johnson,V.J., Yucesoy,B. and Luster,M.I. (2004) Genotyping of single nucleotide polymorphisms in cytokine genes using real-time PCR allelic discrimination technology. *Cytokine*, **27**, 135–141.
13. Liu,W.M., Di,X., Yang,G., Matsuzaki,H., Huang,J., Mei,R., Ryder,T.B., Webster,T.A., Dong,S., Liu,G. *et al*. (2003) Algorithms for large-scale genotyping microarrays. *Bioinformatics*, **19**, 2397–2403.
14. Lovmar,L. and Syvanen,A.C. (2005) Genotyping single-nucleotide polymorphisms by minisequencing using tag arrays. *Meth. Mol. Med*., **114**, 79–92.
15. Kaufman,L. and Rousseeuw,P.J. (1990) Partitioning Around Medoids (Program PAM). In *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, NY, pp.68–127.
16. Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*., **20**, 53–65.
17. Latif,S., Bauer-Sardina,I., Ranade,K., Livak,K.J. and Kwok,P.Y. (2001) Fluorescence polarization in homogeneous nucleic acid analysis II: 5′-nuclease assay. *Genome Res*., **11**, 436–440.
18. Cheng,J., Zhang,Y. and Li,Q. (2004) Real-time PCR genotyping using displacing probes. *Nucleic Acids Res*., **32**, e61.
19. Hultin,E., Kaller,M., Ahmadian,A. and Lundeberg,J. (2005) Competitive enzymatic reaction to control allele-specific extensions. *Nucleic Acids Res*., **33**, e48.
20. Consolandi,C., Frosini,A., Pera,C., Ferrara,G.B., Bordoni,R., Castiglioni,B., Rizzi,E., Mezzelani,A., Bernardi,L.R., De Bellis,G. *et al*. (2004) Polymorphism analysis within the HLA-A locus by universal oligonucleotide array. *Hum. Mutat*., **24**, 428–434.
21. Wong,K.K., Tsang,Y.T., Shen,J., Cheng,R.S., Chang,Y.M., Man,T.K. and Lau,C.C. (2004) Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. *Nucleic Acids Res*., **32**, e69.
22. Brasch-Andersen,C., Christiansen,L., Tan,Q., Haagerup,A., Vestbo,J. and Kruse,T.A. (2004) Possible gene dosage effect of glutathione-*S*-transferases on atopic asthma: using real-time PCR for quantification of GSTM1 and GSTT1 gene copy numbers. *Hum. Mutat*., **24**, 208–214.