Methods 50 (2010) S10–S14

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Application Note

ScreenClust: Advanced statistical software for supervised and unsupervised high resolution melting (HRM) analysis ☆

Valin Reja [a], Alister Kwok [b], Glenn Stone [c], Linsong Yang [b], Andreas Missel [d], Christoph Menzel [d,*], Brant Bassam [e]

[a] *Bio Republic, 14 Birriwa Street Greystanes, NSW 2145, Australia*
[b] *Corbett Research (a QIAGEN Company), 14 Hilly Street, Mortlake, NSW 2137, Australia*
[c] *CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW 1670, Australia*
[d] *QIAGEN GmbH, Qiagenstrasse 1, Hilden, Germany*
[e] *P.O. Box 408, Winton, Qld 4735, Australia*

ARTICLE INFO

Article history:
Accepted 5 February 2010
Available online 8 February 2010

Human genes:
Factor V Leiden

Keywords:
High resolution melting
Principal component analysis
k-Means
Cluster analysis
Linear discriminant analysis
Posterior class probabilities
Typicalities
Software
ScreenClust

ABSTRACT

Background: High resolution melting (HRM) is an emerging new method for interrogating and characterizing DNA samples. An important aspect of this technology is data analysis. Traditional HRM curves can be difficult to interpret and the method has been criticized for lack of statistical interrogation and arbitrary interpretation of results. *Methods:* Here we report the basic principles and first applications of a new statistical approach to HRM analysis addressing these concerns. Our method allows automated genotyping of unknown samples coupled with formal statistical information on the likelihood, if an unknown sample is of a known genotype (by discriminant analysis or "supervised learning"). It can also determine the assortment of alleles present (by cluster analysis or "unsupervised learning") without a priori knowledge of the genotypes present. *Conclusion:* The new algorithms provide highly sensitive and specific autocalling of genotypes from HRM data in both supervised an unsupervised analysis mode. The method is based on pure statistical interrogation of the data set with a high degree of standardization. The hypothesis-free unsupervised mode offers various possibilities for *de novo* HRM applications such as mutation discovery.

1. Introduction

High resolution melting (HRM) is a new method for monitoring DNA dissociation ("melting") kinetics. HRM is an entirely closed-tube procedure requiring only a generic DNA intercalation dye. As double-stranded DNA samples (for HRM analysis typically PCR products) dissociate with increasing temperature, dye is progressively released and fluorescence diminishes. Fluorescent measurements are collected at corresponding temperature increments and plotted as a "melt curve". Curve shape and position are characteristic of each sample allowing them to be compared and discriminated. Even a single base change between samples can be readily detected and identified [1,2].

HRM discriminates genotypes by comparing the relative position and shape of melt curves [2]. These changes reflect a sample's DNA sequence, random generation of heteroduplexes (i.e., mismatched strand duplexes that occur in samples containing more than one sequence variant), buffer conditions, and other reaction variables [1,2].

Enhanced and automated HRM data processing methods are needed, particularly for larger sample cohorts. Current HRM software plots variations in melt curve shape and position, however, the ability to statistically quantify differences is not supported. At present, the data analysis procedure typically uses melt curve fluorescence normalization followed by a simple subtraction (difference) plot generated from a known control sample. Although the method allows automated genotyping, no formal statistical information is provided to indicate the likelihood an unknown sample is of a known genotype (discriminant analysis or "super-

Abbreviations: HRM, high resolution melting; SNP, single nucleotide polymorphism; PC, principle component; LDA, linear discriminant analysis; NTC, no template control.

☆ This application note has been provided by Qiagen as supplemental educational material to this thematic special issue. This application note was sponsored by Qiagen and has not undergone a peer review process within Elsevier.

* Corresponding author. Address: QIAGEN Strasse 1, 40724 Hilden, Germany.
E-mail address: christoph.menzel@qiagen.com (C. Menzel).

1046-2023/$ - see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.ymeth.2010.02.006

vised learning") nor do current methods allow the number of alleles present to be determined (cluster analysis or "unsupervised learning"), useful in the discovery of new sequence variants.

Here we describe algorithms for reliable and relevant automated genotyping of HRM data in supervised and unsupervised mode using a set of advanced statistical methods such as principal component analysis in a software package we call *ScreenClust*.

2. Materials and methods

2.1. HRM data sets

All analyses were performed with a prototype version of the *ScreenClust* software package (Rotor-Gene *ScreenClust* HRM Software, QIAGEN, Hilden, Germany). Three HRM data sets were used to investigate the new algorithms. All of them were generated on a Rotor-Gene Q 5plex HRM instrument (QIAGEN, Hilden, Germany) using 25 μL reaction volumes. These included: (A) 32 known replicates for alleles of the human factor V Leiden (G1691A) polymorphism and a SYBR Green based master mix, 300 nM each primer and 25 ng template DNA. (B) Five replicates of each allele of a synthetic and challenging Class IV (A to T) SNP template, run with EvaGreen fluorescent intercalating dye (Biotium, Hayward, USA), 300 nM each primer and 20 ng template. (C) Three replicates of each of five allele ratios of the factor V Leiden (G1691A) polymorphism (percentage mutation to wild type; 2.5%, 5%, 10%, and 50%) along with the wild type and mutation controls, run with 1× SYTO-9 fluorescent green intercalating dye (Invitrogen, Carlsbad, USA), 300 nM each primer and 25 ng template.

2.2. Analysis workflow

The workflow of the analysis procedure with all processing steps is depicted in Fig. 1. The following sections detail the steps performed.

2.3. Normalization

Any HRM analysis requires the normalization of the start and end fluorescence, as differences in raw fluorescence can be induced by various factors in the PCR and HRM process, e.g., different amounts of total DNA present for amplicons and template. In this study, HRM curves were normalized using two different methods in order to determine which method provides the best representation of known genotypes particularly for unsupervised cluster analysis.

The first method applies curve scaling to a line of best fit such that the highest fluorescence value was equal to 100 and the lowest to zero. A region prior to and following melt curve transition is selected to calculate average fluorescence and slope of the curve and applied in the normalization.

The second method fits an idealized model of a double-stranded melt curve using Levenberg–Marquardt least-squares estimation of non-linear parameters [3]. The idealized model was adapted from the ideal melt curve described by Azbel [4] amended by parameters describing background noise and fluorescence changes due to temperature, which are typically observed in HRM melt curves (see Supplemental material for details of the fitting functions).

Both methods of normalization were compared using two HRM data sets containing various types of alleles of differing fragment length and complexity of sequence.

2.4. Data processing and principal component analysis

To accentuate differences between individual samples, normalized melt curves are first differentiated in *ScreenClust*. Following this, a residual plot is generated by subtracting all the differentiated curves by the composite median of all curves (see Fig. 1). The residual plot is used as the data basis for a form of principal component analysis to extract a set of features for each curve. Principal component analysis selects the linear combination of the data vector that shows the most variation among the samples as the

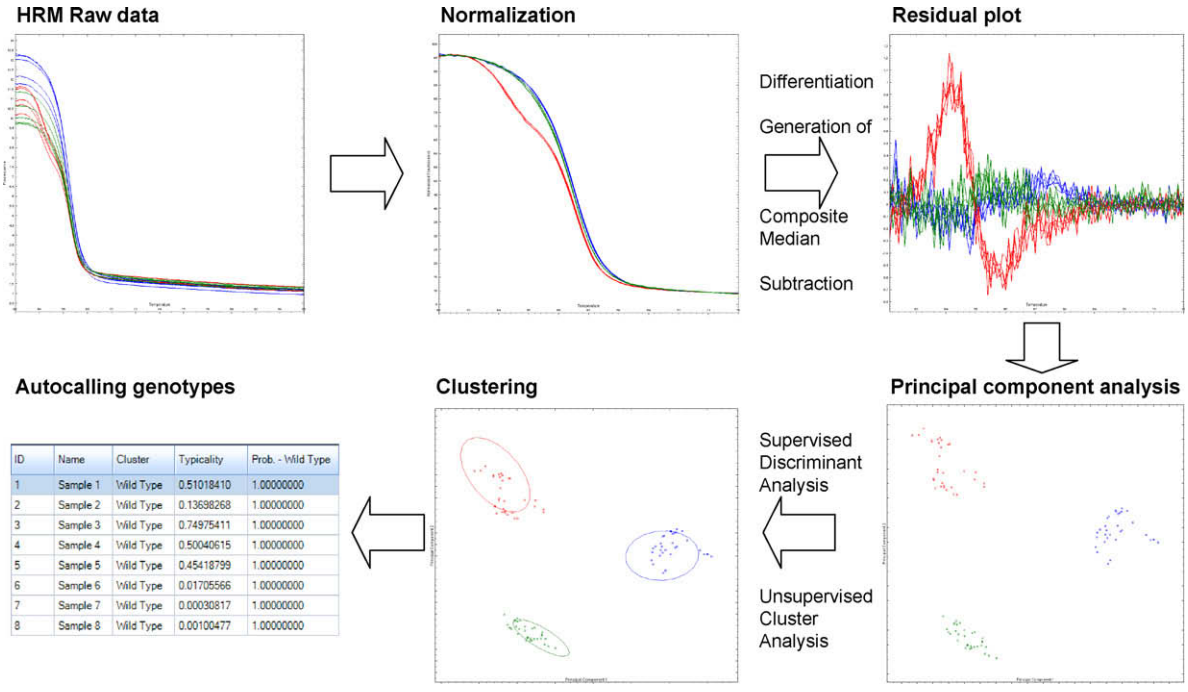

Fig. 1. Processing and analysis workflow in the *ScreenClust* software (see text for details).

first principal component (PC) [5]. Subsequent PCs account for as much of the remaining variability as possible. For most HRM data sets only up to the first three PCs are required in our experience as most of the remaining variance in the other PCs will be attributed to uninteresting variation or noise (data not shown). The selection of the appropriate number of PCs to use is described below.

2.5. Supervised discriminant analysis of data with known controls

Classification of unknown samples into known groups using known samples as controls was achieved using linear discriminant analysis (LDA). LDA arises as the optimal classifier when the data have a multivariate normal distribution, and each class has the same covariance matrix but differing means [6]. Using the control samples, LDA calculates a cluster distribution. The center of the cluster is set as the mean of the controls.

Unknown samples are allocated to a cluster based on their proximity to the mean points of the controls (see Supplemental material for more mathematical details). When the number of controls in each group is $\geqslant 2$, then LDA can be applied formally. However, it would be desirable to have more known samples than this (i.e., $\geqslant 4$ controls per group).

If only one control is provided per cluster, a nearest neighbor calculation is used by allocating the samples to a cluster based on their proximity to the control.

The choice of PC dimensionality using a supervised data set is achieved using a cross-validation to find the number of PCs that produces the lowest number of misclassifications. Cross-validation involves leaving a known sample, and using the remainder to build a classifier. The left out observation is then classified and any error counted. This process is repeated for two and three PCs and the dimensionality with the lowest error rate chosen.

2.6. Unsupervised cluster analysis without known controls

The aim of unsupervised cluster analysis is to find *de novo* data groups without a priori knowledge on the number and kind of genotypes present in the data set. To achieve such a hypothesis free analysis we selected the method of *k*-means cluster analysis. *k*-Means undergoes an iterative process where clusters are generated by choosing *k* random cluster centers and allocating samples to clusters. The ideal cluster configuration for *k* clusters is the one where the within cluster sum of squares is minimized [7].

Alone, *k*-means is unable to determine the number of clusters; rather it defines clusters based on a given number of clusters. To determine a choice of the number of clusters we combine *k*-means with the Gap statistic [8]. The idea is to look at a measure of cluster quality and compare it to that of a simulated data set known to not have any real clusters (see Supplemental material for more details on the implementation of *k*-means and the gap statistics in *ScreenClust*).

2.7. Posterior class probabilities and typicalities

Posterior class probabilities are the probabilities that each sample is a member of each group assuming that the sample is a member of one of the groups. An unknown sample would be allocated to the group with the largest posterior class probability.

Posterior class probabilities tell us which group a sample is most likely a member of, given that it is a member of at least one of them.

An additional typicality index tells us how consistent a sample is within its own group, i.e., the typicality measures how well a sample fits into its assigned cluster (for details of the calculation of posterior probabilities and typicalities see Supplemental material).

2.8. ScreenClust software

Raw HRM data from the Rotor-Gene operating software are imported into *ScreenClust* software directly using the .rex file format, with syntactic analysis (parsing) to ensure the correct data is analyzed. The overall analysis procedure is guided by a software wizard and default values allowing, if desired, the standardized generation of a genotyping result with only one operator selection for "supervised" or "unsupervised" mode. All no template controls (NTC) and unnamed samples are automatically removed from the analysis as the lack of melt curve features affects the normalization. Following sample selection, the user can select between analyzing supervised or unsupervised data sets.

Choosing "supervised" allows the appropriate controls for each genotype to be selected. If more than two controls were used for each group, LDA is used and the appropriate number of PC determined via the cross-validation function. Having only one control would activate the nearest neighbor classification of samples into groups.

Selecting "unsupervised" enables the *k*-means clustering and Gap statistic algorithms, with the software selecting the most appropriate cluster number and PC number to use.

The clusters are graphically plotted using the loading scores of each PC for all samples (e.g., PC1 vs. PC2) for both supervised and unsupervised methods. An ellipsoid representing cluster covariances following classification is also drawn. Unknown samples are classified into each cluster group with posterior probabilities and typicalities being calculated for all samples.

3. Results

3.1. Evaluation of normalization methods

Using the normalization process of fluorescence scaling to a line of best fit we observed that the data retained most of the curve features post normalization whereas the idealized Levenberg–Marquardt fitting algorithm resulted in the characteristic heteroduplex curve double inflection to disappear (see Supplemental material). The clustering and calling of unknown samples and clusters was nevertheless effective for most data sets and both normalization methods. However, for one of our data sets containing difficult to resolve allelic ratios, down to as low as 2.5%, of the factor V Leiden polymorphism, loss of curve features following normalization using the idealized model algorithm did result in incorrect cluster-

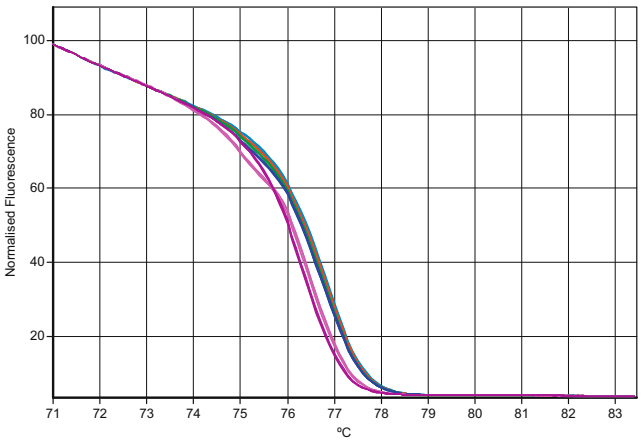

Fig. 2A. Normalized HRM curves of four allelic ratios of the factor V Leiden (G1691A) polymorphism (from 5% to 100%) as well as wild type and heterozygous samples using the fluorescence scaling to a line of best fit normalization. The curves have minimal curve shape topography, especially for allelic ratios of less then 10%.

ing and calling of pseudo-unknowns using unsupervised analysis (data not shown). Using the fluorescence scaling to a line of best fit model for this data set, all clusters and samples were called correctly with high posterior probabilities and typicalities in unsupervised mode (Fig. 2A, B).

3.2. Class IV SNP genotyping

ScreenClust was capable of clustering all three genotypes of the synthetic Class IV SNP (A to T) polymorphism data set calling all pseudo-unknowns correctly into their respective genotypes using the unsupervised analysis feature (Fig. 3A). When the supervised analysis feature was applied and the first three samples of each genotype selected as controls, the software again was capable of correctly calling all the pseudo-unknowns (see auto-calling results table in the Supplemental material). We compared the *ScreenClust* supervised result for the Class IV SNP data set to a result obtained using the HRM analysis module of the Rotor-Gene operating software (version 1.7.9.4). The first sample of each genotype was selected as the control and the percentage confidence limit was set to 70%. Three out of 12 samples could not be correctly called using the HRM module within the Rotor-Gene operating software when a 70% confidence percentage was applied (see Fig. 3B and the comparison table of autocalling results in the Supplemental material).

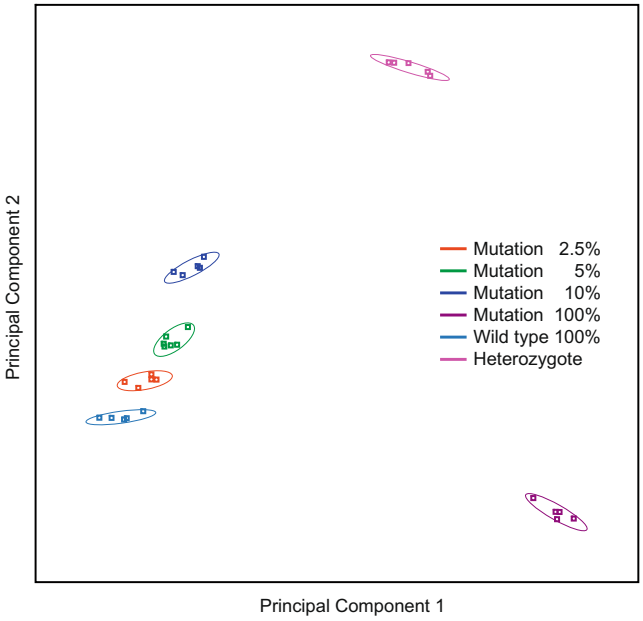

Fig. 2B. Unsupervised *ScreenClust* analysis of the factor V Leiden (G16191A) mutation at various allelic ratios. Using the fluorescence scaling to a line of best fit normalization resulted in all allelic ratios being detected.

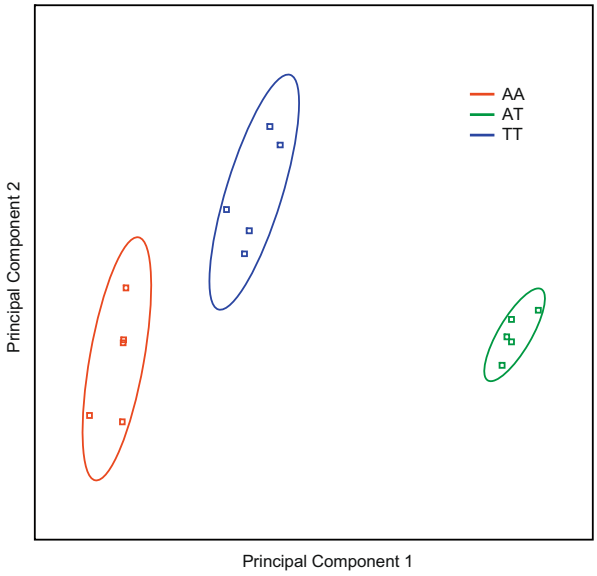

Fig. 3A. Unsupervised *ScreenClust* analysis of a class IV (A to T) SNP. All pseudo-unknowns for the AA (red), TT (blue) and AT (green) were correctly clustered and called into the respective genotypes. Clusters are highlighted by ellipsoids that represent the covariance of the classified samples in a two PC dimension plot.

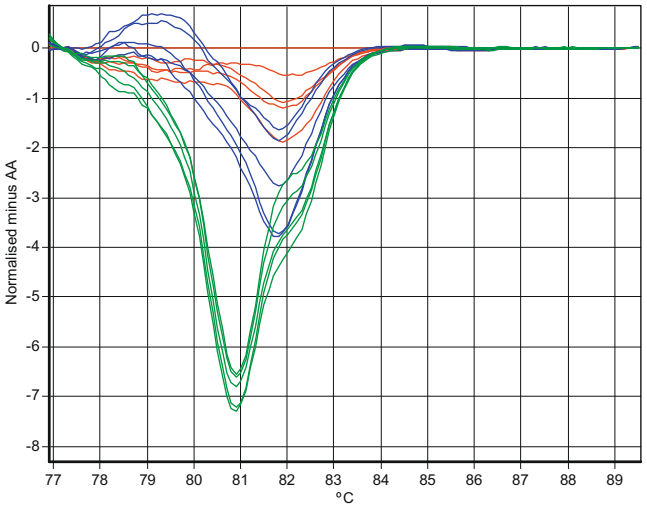

Fig. 3B. Standard HRM subtraction plot of the class IV SNP provided by the Rotor-Gene Q version 1.7 software with the first AA sample selected as the control. A number of the TT allele samples (blue) could not be differentiated from the AA genotype (red). Only the heterozygote AT allele samples were all called correctly.

4. Discussion

4.1. ScreenClust software was capable of detecting, clustering and correctly calling genotypes from various HRM data sets

Comparisons between two different normalization techniques demonstrated that the scaling of fluorescence using a line of best fit was more appropriate than the idealized model melt curve algorithm for HRM curves particularly for smaller variations in allele composition (<5%). In applications such as the search for somatic mutations, discrepancies in detecting smaller allelic ratios can be problematic. Therefore, we implemented the methods of scaling of fluorescence to a line of best fit normalization for all other investigations.

Other known software packages for HRM analysis apply a "temperature shift" to the knee at the end of the melt transition [9], essentially normalizing the *X*-axis temperature data in addition to the *Y*-axis fluorescence data. This type of data manipulation eliminates much of the information content of the HRM curve, with only the overall shape of the curve left intact. We have, therefore, intentionally avoided any type of temperature-shifting normalization in order to provide the algorithms with the broadest unbiased information available.

In summary, *ScreenClust* is capable of observing complex genotypes with high sensitivity and specificity. We believe the *ScreenClust* algorithms are outperforming the current generation of software programs for HRM analysis with respect to the discriminating power for genotypes as well as the statistical interrogation

and interpretation of the sample set. In this context, it should be noted that the new algorithms offer a completely orthogonal approach for HRM analysis software for independent validation and verification of HRM assays developed with the standard HRM analysis approach. Furthermore, it is the first software that allows for the detection and statistical analysis of unsupervised HRM data sets. This feature is highly advantageous to investigators attempting to discover new polymorphisms. However, the sensitive algorithms may also easily find and cluster artifacts as individual pseudo-genotypes such as deviations in the master mix compositions (unpublished results). This, on one hand, allows monitoring the quality of the HRM procedure with every run, but, on the other hand, emphasizes the need also for highly standardized and reliable reaction conditions for the applied reagents and the melt analyzer for a successful HRM analysis.

The unsupervised mode is also the method of choice, if not for all putative genotypes in the data set controls are available. The partial set of controls is then employed as pseudo-unknowns in unsupervised mode. These controls and the corresponding samples of the same genotype will form a cluster whereas a new polymorphism will separate in another cluster.

Here we have shown first examples of the *ScreenClust* algorithms for SNP genotyping, respectively, mutation discovery, but we believe that the approach also offers interesting possibilities for other HRM applications such as the detailed analysis of insertions and deletions, pathogen detection but also methylation analysis.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ymeth.2010.02.006.

References

[1] K.M. Ririe, R.P. Rasmussen, C.T. Wittwer, Anal. Biochem. 245 (1997) 154–160.
[2] C.T. Wittwer, G.H. Reed, C.N. Grundry, J.G. Vandersteen, R.J. Pryor, Clin. Chem. 49 (2003) 853–860.
[3] D.W. Marquardt, J. Soc. Ind. Appl. Math. 11 (1963) 431–441.
[4] M. Azbel, Proc. Natl. Acad. Sci. USA 76 (1979) 101–105.
[5] I. Jolliffe, Principle Component Analysis, second ed., Springer, New York, 2002.
[6] J. Ye, T. Li, T. Xiong, R. Janardan, IEEE/ACM Trans. Comput. Biol. Bioinform. 1 (2004) 181–190.
[7] J.A. Hartigan, M.A. Wong, Appl. Stat. 28 (1979) 1000–1008.
[8] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J.R. Stat. Soc. 63 (2001) 411–423.
[9] M.G. Herrmann, J.D. Durtschi, L.K. Bromley, C.T. Wittwer, K.V. Voelkerding, Amplicon DNA melting analysis for mutation scanning and genotyping: cross-platform comparison of instruments and dyes, Clin. Chem. 52 (2006) 494–503.