BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 16 2011, pages 2306–2308
doi:10.1093/bioinformatics/btr383

Genetics and population analysis
Advance Access publication June 22, 2011

SNPman: a program for genotype calling using run data from TaqMan allelic discrimination

Martin Konopac[1,2], Petra Dusatkova[2,3] and Ondrej Cinek[2,3,*]

[1]Faculty of Biomedical Engineering, Czech Technical University in Prague, Kladno, [2]Department of Paediatrics, 2nd Faculty of Medicine, Charles University in Prague and [3]Department of Paediatrics, University Hospital Motol, CZ-15006 Prague, The Czech Republic

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: The SNPman program calls the genotypes of single nucleotide polymorphisms (SNP) from TaqMan allelic discrimination assays. It utilizes the fluorescence data collected over the whole PCR run, rather than relying on the end point fluorescence measurements that is the basis of the genotype calling process in most software solutions sold with the real-time instruments. This inspection of run data facilitates genotype calls in difficult sample sets, especially in those containing various concentrations of DNA or inhibitors, as indicated by results of a reanalysis of 3738 genotyping samples. The program works with data from three different widely used PCR instruments.

Availability: The compiled program is available online at http://sourceforge.net/projects/snpman/files/, along with its user documentation and demonstration data files. It is free of charge for non-commercial users.

Contact: Ondrej.Cinek@Lfmotol.cuni.cz

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on April 27, 2011; revised on June 13, 2011; accepted on June 17, 2011

1 INTRODUCTION

The TaqMan allelic discrimination is a very popular and widely used medium-throughput genotyping format: the assays use two short minor groove binder (MGB) hydrolysis probes, each annealing to one allelic variant of the genotyped single nucleotide polymorphism (SNP) amplified within a short polymerase chain reaction (PCR) product (Bustin, 2004; Chen and Sulivan, 2003, Komar *et al*., 2009, Sobrino, 2005). The TaqMan allelic discrimination is supported by most of real-time PCR platforms; however, their proprietary software solutions utilize mainly the end point analyses, largely overlooking the potential of fluorescence data collected over the whole course of a PCR run.

In PCR runs with samples of good quality and comparable quantity, the end point scatter plot suffices for genotype calling. Nevertheless, if the reaction runs in suboptimal conditions (e.g. cheaper alternatives to the brand polymerases are used, the quantity of starting DNA varies or the DNA contains inhibitors), the end point analyses may not effectively discriminate genotypes, because the groups in the scatter plot often stretch or blend together.

We therefore aimed to develop a program that would aid in TaqMan genotype calling across various platforms by enabling a detailed inspection of fluorescence data from the polymerase chain reaction (PCR) runs. In particular, we aimed for improving the genotype call rate in runs with unequally concentrated samples, in home-brew assays and in samples with variable presence of inhibitors.

2 METHODS

The program accepts PCR fluorescence data exported from SNP typing assays run on either of the Applied Biosystems 7300 (ABI7300), LightCycler LC480 (LC480) or Biorad CFX (Biorad CFX) instruments. It can analyze either 96 or 384-well plates. The generation of the import files is described in documentation downloadable along with the program.

After data import, an amplification graph is displayed (Supplementary Fig. S2, graph A): the graph plots the VIC fluorescence (for one allele) and the FAM fluorescence (for the other allele) against time (cycle), two curves for each well. The fluorescence can be shown in linear or logarithmic scale. Three baseline calculation options are available: (i) raw; (ii) horizontal, assuming no amplification-independent drift of signal; (iii) linear, which takes the linear regression equation from the baseline cycles, and thus can partly compensate for an amplification-independent drift. The cycle range from which the baseline is taken can be manually adjusted. This amplification graph has a crosshair widget: its centre as well as the line ends can be grabbed and moved using the mouse. The vertical line determines the current cycle that is viewed in the graph B (Supplementary Fig. S2, graph B plotting the VIC and FAM fluorescence against each other). The horizontal line is used for setting the fluorescence threshold that determines the threshold cycle (C_t) for every sample and allele; these threshold cycles are viewed in graph C and D.

The fluorescence graph (Supplementary Fig. S2, graph B) plots the VIC fluorescence (one allele, horizontal axis) against the FAM fluorescence (the alternative allele, vertical axis) at the time point set by the vertical line of the crosshair. The horizontal line of the crosshair (the threshold) is reflected here as a circle or ellipse quadrant. When the vertical line of the crosshair is moved along the course of the PCR, the points representing the samples fly from their origin at coordinates 0, 0 outwards. The angles under which they leave the bottom left corner of the graph reflects the genotype, which is useful for genotype assignment: the graph has three angle sectors that can be manually set and used in genotype calling.

The threshold cycle graph (Supplementary Fig. 2, graph C) plots the threshold cycles (C_t) of the two probes against each other (the horizontal C_t is of the VIC probe, the vertical of the FAM probe). If the sample does not reach the amplification, its threshold cycle is arbitrarily set to the end of PCR (45 in the run shown in Supplementary Fig. 2, graph A) to retain the dot in the graph for allele calling.

The difference threshold cycle graph (Supplementary Fig. 2, graph D) is a modification of the graph C containing the identical data. While the vertical

*To whom correspondence should be addressed.

axis shows the FAM C_t for the samples, the vertical axis shows the difference between the VIC and FAM C_t. This helps in discrimination of the genotypes, since samples of identical genotype are expected to cluster vertically, having a similar distance between the VIC and FAM C_t, regardless of their DNA concentration.

The graphs B–D possess features allowing calling the genotypes of the individual points or to their groups. Groups of samples can be selected using the mouse, and in the graph B the genotype can be also assigned using the three angle sectors. Every assignment is logged and the last is recorded into the result file. The PCR plate layout tabulated below the graphs allows selecting a subset of samples for detailed inspection. After the genotypes are assigned, the resulting file can be exported into a text file. The set of analyzed plates (along with the analysis settings as threshold or cycle, and with genotype assignments) can also be saved as a project, and later retrieved.

3 RESULTS

To test whether this algorithm aids in calling alleles that could not be distinguished using proprietary software of the three vendors, we reanalyzed run files that contained genotypes unresolvable by the proprietary software according to the manufacturer's instructions. In particular, we concentrated on selecting PCR runs with DNA of poor quality, sample sets of uneven DNA concentration, samples containing varying amounts of inhibitors and runs containing alternative master mix chemistry.

We reanalyzed 21 files from ABI7300 with 1814 samples, 6 files from LC480 with 1560 samples and 4 files from Biorad CFX with 479 samples. The files were selected based on previously encountered problems with genotype calling using the proprietary software of the real-time PCR machines, mostly due to strong variations in DNA quantity and quality. The primers and probes for the assays were produced by Applied Biosystems, and were targeted mostly to SNPs within—and adjacent to—the *NOD2/CARD15*, *CTLA4*, *IFIH1*, *GCK* and *PTPN22* genes typed in our previous studies (e.g. Hradsky *et al.*, 2008). The DNA samples were extracted from human whole blood or saliva using various techniques and archived up to 12 years thereafter. The reanalysis was done independently on the original genotype calls from the instruments' software. Both the original analysis and the reanalysis were done by two independent evaluators.

The result of the above reanalysis is summarized in Table 1: in total, genotypes of 372/3738 samples could not be resolved by the programs sold with the instruments. Of these 372 unresolved genotypes, 116 (31%) were then called using the SNPman. Of the genotypes that were called by the instruments' software, the SNPman generated five genotypes discordant with the previous calls, and eight samples could not be resolved by SNPman although the instrument's analysis yielded a genotype. We did not analyze differences across the three instrument platforms as the success rate of the genotyping heavily depended on the quality and quantity of DNA in the various projects, rather than on the type of the instrument.

Figure 1 shows an example of a run that entirely failed in the end point analysis but was successfully sorted into genotypes using our program (this run was not included in the overall analysis). Here, the blending of signals from unevenly concentrated samples is visible toward the end of PCR, which effectively hampered end point analyses using fluorescence data. Note that the graphs C and D aided in genotype calling independently of these fluorescence data, and that moving the vertical line of the crosshair enabled better resolution of the samples in their early amplification.

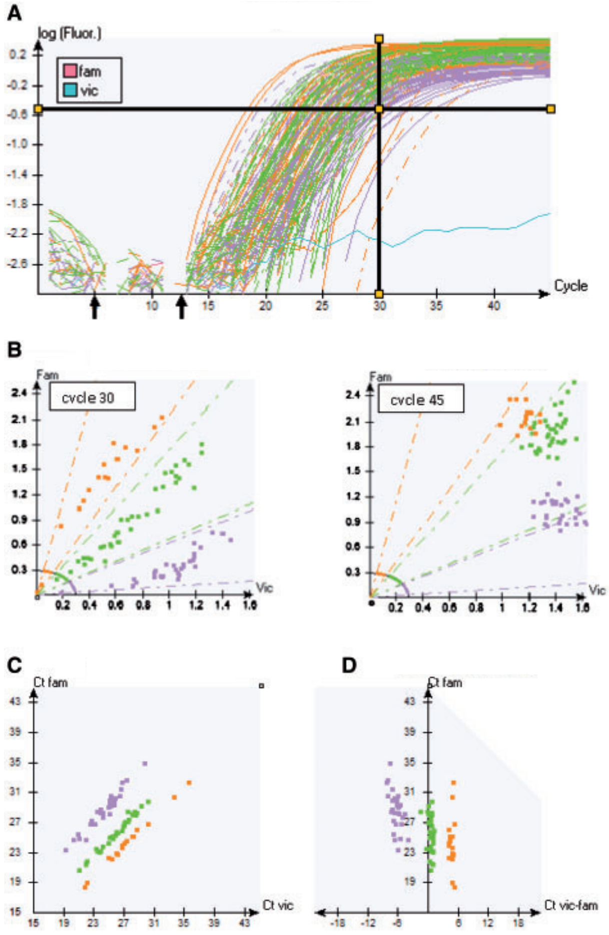

Fig. 1. Typical PCR run whose genotypes could not be called using the end point analysis. (**A**) The amplification plot with a wide spread of the curves indicating the uneven DNA concentrations. (**B**) The fluorescence graph. (Left) Cycle 30: the groups in the fluorescence graph are well separated, yet a significant proportion of the samples has not yet amplified. (Right) Cycle 45: all samples have been amplified, but the signals suffer from a significant cross-reactivity that blends the groups in the fluorescence graph together. (**C**) Graphs utilizing the threshold cycle. The genotypes can be successfully called using either graphs C and D. (**D**) Difference in C_t (color codes show the allele calls. Orange, FAM allele homozygotes, green, heterozygotes, blue, VIC allele homozygotes).

4 CONCLUSIONS

Our software brings deeper insight into the course of the fluorescence over the whole SNP-typing PCR run. This may add important information, and substantially help correctly call genotypes in suboptimal samples, e.g. with uneven concentration and quality of DNA. The main advantage of our system is the parallel inspection of four views on the amplification, the interactive character of the analysis that is directed using the crosshair widget and the data input from three different PCR machines. As our program requires no installation and provides a very simple user interface

Table 1. Results of the repeated analyses of difficult PCR runs, using our program, and comparison to the original end point analyses

	ABI7300	LC480	Biorad CFX	Total
Runs reanalyzed	21	6	4	31
Samples reanalyzed	**1699**	**1560**	**479**	**3738**
Concordant genotypes by both methods	1424 (83.8)	1423 (91.2)	390 (81.4)	3237 (86.6)
Unresolvable by either method	181 (10.7)	117 (7.5)	74 (15.4)	372 (10.0)
Unresolvable by the end point analysis, but successfully called by the SNPman program	90 (5.3)	15 (0.96)	11 (2.2)	116 (3.1)
Discordant genotypes between methods	1 (0.06)	2 (0.13)	2 (0.42)	5 (0.13)
Called by the end point analysis but unresolvable by the SNPman program	3 (0.18)	3 (0.19)	2 (0.42)	8 (0.21)

Data are *n* (%).

and export options, it may complement the existing proprietary software solutions and increase the productivity of low- to medium-throughput genotyping.

ACKNOWLEDGEMENTS

We thank Dr Jan Štěpánek from the Faculty of Mathematics and Physics, Charles University in Prague, for an invaluable input on Perl programming strategies and on algorithms.

Funding: Czech Ministry of Education (grant MSM0021620814).

Conflict of Interest: none declared.

REFERENCES

Bustin,S.A. (2004) *A-Z of Quantitative PCR*, 1st edn. University Line, La Jolla, USA.

Chen,P. and Sulivan,P. (2003) Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *Pharmacogenomics J.*, **3**, 77–96.

Hradsky,O. *et al.* (2008) Variants of CARD15, TNFA and PTPN22 and susceptibility to Crohn's disease in the Czech population: high frequency of the CARD15 1007fs. *Tissue Antigens*, **71**, 538–547.

Komar,A.A. *et al.* (2009) *Single Nucleotide Polymorphisms: Methods in Molecular Biology*, 2nd edn. Humana Press, New York, NY, USA.

Sobrino,B. (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Sci. Int.*, **154**, 181–194.