BMC Genetics

SOFTWARE Open Access

Fragman: an R package for fragment analysis

Giovanny Covarrubias-Pazaran[1*], Luis Diaz-Garcia[1,2], Brandon Schlautman[1], Walter Salazar[1] and Juan Zalapa[1,3*]

Abstract

Background: Determination of microsatellite lengths or other DNA fragment types is an important initial component of many genetic studies such as mutation detection, linkage and quantitative trait loci (QTL) mapping, genetic diversity, pedigree analysis, and detection of heterozygosity. A handful of commercial and freely available software programs exist for fragment analysis; however, most of them are platform dependent and lack high-throughput applicability.

Results: We present the R package *Fragman* to serve as a freely available and platform independent resource for automatic scoring of DNA fragment lengths diversity panels and biparental populations. The program analyzes DNA fragment lengths generated in Applied Biosystems® (ABI) either manually or automatically by providing panels or bins. The package contains additional tools for converting the allele calls to GenAlEx, JoinMap® and OneMap software formats mainly used for genetic diversity and generating linkage maps in plant and animal populations. Easy plotting functions and multiplexing friendly capabilities are some of the strengths of this R package. Fragment analysis using a unique set of cranberry (*Vaccinium macrocarpon*) genotypes based on microsatellite markers is used to highlight the capabilities of *Fragman*.

Conclusion: *Fragman* is a valuable new tool for genetic analysis. The package produces equivalent results to other popular software for fragment analysis while possessing unique advantages and the possibility of automation for high-throughput experiments by exploiting the power of R.

Keywords: Fragment analysis, Genetic markers, R package, Least squares, Open source software

Background

Polymerase chain reaction (PCR)-based genetic markers such as microsatellite or simple sequence repeat (SSR), amplified fragment length polymorphism (AFLP), and single nucleotide polymorphism (SNP) markers have become essential in genetic research [1, 2]. Traditionally, fragment analysis has been the crucial first step for genetic research such as mutation detection, linkage and quantitative trait loci (QTL) mapping, genetic diversity and other genetic applications using molecular markers [2, 3]. During fragment analysis, fluorescently-labeled PCR amplified DNA fragments, which have been denatured by a chemical reagent (i.e., formamide), are injected along with an appropriate size standard into capillaries for electrophoresis size separation [4, 5]. Essentially, a high voltage charge applied to the PCR fragments promotes their migration along the capillaries from cathode to anode, where smaller fragments migrate faster than larger fragments [4]. When capillary electrophoresis is performed using ABI instrumentation (i.e., AB Genetic Analyzer, Applied Biosystems 3730), the size-separated PCR fragments are detected by reading their fluorescence intensity at different emission wavelengths and are recorded as FSA (.fsa) files, which can be used in downstream analyses [5]. FSA files are typically loaded into specialized licensed and platform dependent software which compare the capillary electrophoresis position and fluorescent intensity of the sample PCR fragments and the size standard to estimate the number of nucleotides in the sample. Here, we present a new freely available and

* Correspondence: cova_ruber@live.com.mx; jezalapa@wisc.edu
[1]Department of Horticulture, University of Wisconsin, Madison, WI, USA
Full list of author information is available at the end of the article

platform independent R package, *Fragman*, for performing fragment analysis. *Fragman* accurately and automatically scores DNA fragment lengths in diversity panels and biparental populations and transforms the observed lengths into formats required for further genetic analysis in other software such as GenAlEx, JoinMap and OneMap [6–8]. *Fragman* was compared with other fragment analysis software such as GeneMarker®, and we obtained similar genotyping results, but with superior automation and throughput scoring capabilities.

Implementation

The workflow of the program consists in the use of 5 basic steps: 1) Reading the data using the function *storing.inds*, which loads the FSA files and smooth the data; 2) matching the ladder with the function *ladder.info.attach*, which finds the correct peaks in the size-standard channel corresponding to the expected DNA sizes to fit a linear model in order to calibrate the samples and attaches such information to the R environment for subsequent use; 3) creating panels with the function *overview2,* which is used to generate bins of alleles by marker; 4) scoring peaks and assigning DNA sizes with the function *score.easy*, which locates the peaks provided in the panel and assigns the size in base pairs for each sample; and 5) Exporting to different formats with the functions *get.scores* and *jm.conv*.

In the special case where samples cannot be correctly matched by the automatic *ladder.info.attach* function because the samples are too noisy, the *ladder.corrector* function has been provided to allow the users to manually correct noisy samples. Also, an extra function named *overview,* allows the users to manually score the samples through regular functions available by default in R, such as the *locator* function.

Reading the files using the storing.inds function

This function is a wrapper function of *read.abif* from the *seqinr* package, which reads FSA files [9]. In addition, *storing.inds* extracts the channel information containing the fluorescent intensities from the DNA capillary electrophoresis up to any number of colors. Fourier frequency transformation (FFT) techniques are applied for smoothing the data in order to enhance the signal, pull up adjustments are performed to diminish channel to channel noise, and peak correction for saturated peaks over 8000 relative fluorescent units (RFUs) is performed as other licensed software do [10].

Matching the ladder with peaks found using the ladder.info.attach function

The user must supply a numeric vector containing the expected base pairs sizes of the ladder fragments comigrating with the sample DNA fragments during capillary electrophoresis. The program calculates the first derivative of the intensity vector for the channel of fluorescence containing the size standard, and finds the point where the slope approximates zero (i.e., $\frac{dy}{dx} = 0$: where y is the intensity with respect to the index position x) using the *rle* function from the *base* package [11]. An iterative procedure using least squares creates parallel models and model with the highest correlation is then selected. This procedure confidently finds the correct fluorescent peaks in all the FSA files to match them with the expected DNA sizes of the size standard, and finally uses a linear model of the form $y = X\beta + \varepsilon$ to assign a base pair value to each index of the intensity vector where y is the response defined as the expected DNA sizes for the ladder, X is the incidence matrix for fixed effects, β is the vector of fixed effects for the polynomial regression until the fifth order to account for the migration differential between DNA pieces of different sizes [12].

An example of how to create a vector with the expected sizes and size the samples is:

```
my.ladder <- c(120,125,129,150,175,
               200,225,250,275,300,325,350)
```

```
ladder.info.attach(stored=my.data,
                   ladder=my.ladder)
```

Panel generation by using the overview2 function

After matching the fragments in the size standard to their expected lengths, all sizing information must be loaded into the R environment. Subsequently, the best way to score samples is by creating panels across the capillary regions in the channels where the PCR products of interest were read [10]. Peaks can be easily visualized by using the *overview2* function, which generates a plot overlapping all the fluorescent signals for all loaded FSA files in order to manually select peaks for creating panels of allele bins.

The plot allows graphical assessment of zero slope peaks (i.e., alleles) present in the set of selected FSA files (i.e., individuals of a population). This plot can be used to create a panel of allele bins that are passed to the *score.easy* function by manually clicking on the desired zero slope peaks and obtaining the potential DNA sizes in the population using the *locator* function installed by default in R package *base* [11]. This function can create the panel as:

```
overview2(my.inds=my.plants,cols=1,
          ladder=my.ladder)
my.panel<-locator(type="p",pch=20,
                  col="red")$x
```

Scoring peaks and assigning DNA sizes with score.easy function

The core of the program relies on this function. The function uses information from the FSA files read by *storing.inds* and the size standard calibration information generated by *ladder.info.attach* to perform a zero slope peak search in the channels/fluorescent colors specified by the user and assigns size in base pairs for such peaks. In addition, a panel of zero slope peaks (i.e., alleles) supplied by the *overview2* function reduces the search to a subset of potential DNA fragment sizes.

The user can implement this function as follow:

```
ans<-score.easy(my.data,cols=1,
                 panel=my.panel,
                 ladder=my.ladder)
```

In this line of code the user provides the original data read in step 1 ("my.data"), chooses to score the first channel (cols = 1), provide the panel created in step 3 ("my.panel"), and includes the ladder specified in step 2 ("my.ladder").

Exporting to different formats with the functions get.scores and jm.conv

One of the strengths of the *Fragman* package is the capability to convert to other formats commonly used in genetic analysis such as JoinMap®, OneMap and GenAlEx. The implementation of such function is straight forward after the scoring step. To extract the results for a marker scored in data frame format as any other licensed software we can use:

```
newdata<-get.scores(ans)

newdataJM<-jm.conv(newdata)

write.csv(newdata,"GAdata.csv")

write.csv(newdataJM,"JMdata.csv")
```

where "newdata" is a user-friendly presentation of the data in a column format. The second line of code shows how the data in a column format is easily converted to JoinMap® format.

Results and discussion

To determine the accuracy of the software, we compared the analyses of 1000 raw FSA from four different cranberry populations (3 biparental mapping populations and 1 genetic diversity panel) in *Fragman* and the commercial software GeneMarker®.

Loading and cleaning of the data

Initially, we started a project by loading the data into R using the function *storing.inds* [9]. The function extracted fluorescent intensity information from all channels/fluorescent colors creating a data frame that was smoothed by applying a Fourier transform using only the top 40 % of lowest frequencies (Fig. 1). A pull up correction was then applied to each channel to decrease channel to channel noise (Fig. 1). First, all channels were added to identify capillary regions containing zero slope peaks. Then, a window was set across the capillary regions containing zero slope peaks to identify the channel where each peak had

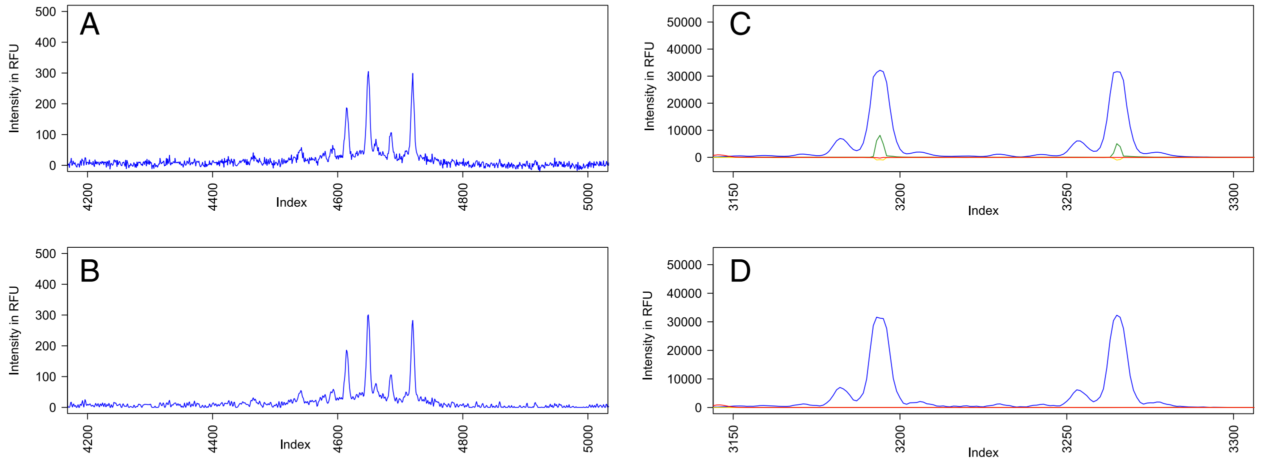

Fig. 1 Effect of Fourier transformation on smoothing and pull up correction during fragment data analysis using *Fragman*. **a** Raw data extracted from the FSA file and (**b**) data treated after Fourier transformation. Raw peak prior to (**c**) and after (**d**) pull up correction to decrease noise between channels

the largest intensity and then subtract the noise caused by such peaks in other channels.

Testing the correlation of the ladder sizing

Size standard calibration in *Fragman* in the channel containing the size standard relies on detection of all possible combinations of zero slope peaks that surpass an initial fluorescent threshold. Subsequently, the program conducts an iterative procedure and selects the combination with the highest correlation with expected size standard fragment sizes. In order to assess the accuracy of size standard zero slope peak selection and sizing in *Fragman*, two different size standards were tested, ROX375 (GeneScan™) and LIZ500 (GeneScan™), and size standard calibration was performed in a diversity panel using 24 microsatellite markers in 2 channels (1000 FSA files total) using FAM (blue) and HEX (green) fluorescent dyes. Size standard calibration performance in *Fragman* was compared to GeneMarker®. First, we validated the *Fragman* size standard calibration function*, ladder.info.attach*, by extracting the correlation observed between the expected size standard fragment lengths and the selected size standard zero slope peaks determined for 1000 FSA files. We found that they yielded an average correlation of 0.99951, indicating that the expected fragment lengths and those observed using *storing.inds* matched (Fig. 2). Using both ROX375 (GeneScan™) and LIZ500 (GeneScan™), we found a perfect match between the size standard zero slope peaks and their predicted fragment length in base pairs when compared with commercial software GeneMarker® sizing [10] (Fig. 3). Similar to commercial software, *Fragman* had problems detecting the correct combination of ladder peaks when the ladder's relative fluorescent units (RFUs) was lower than 150, which dramatically increased the number of peaks in the channel containing the size standard due to noise, making it impossible to compute such a large number of zero slope peak combinations. To deal with low fluorescent signals in the channel containing the size standard, *Fragman* uses an exhaustive sampling strategy of 15,000 random peaks in the size standard channel to find the best combination of possible size standard zero slope peaks. In our tests, this strategy resulted in a correct solution about 94 % of the time.

Panel validation

Identical DNA zero slope peaks (i.e., alleles), even though exactly the same fragment length, do not necessarily occur at the exact same base pair index in two different FSA files due to differential migration of fragments during capillary electrophoresis. This problem can lead to different allele calls between FSA files when in reality they are the same allele. Therefore, most commercial software such as GeneMarker® and GeneMapper® have an option for creating allele panels with scoring windows that account for differential migration in order to make size fragment scoring faster and more accurate. For example, creating a panel for an allele of 200 bp in length with a zero slope index of 200.3 bp and a window of 0.5 bp will allow every peak in that range to have the same allele call. We implemented a similar approach for creating panels of allele bins with in the *overview2* function by overlapping the curves of our samples to determine the appropriate window size for each allele bin. This approach allows *Fragman* to correct for differential migration and to perform genotypic calls more efficiently (Fig. 4). The *locator* function from the R *base* package is implemented in *overview2* to allow the user to click on the desired zero slope peaks (alleles) for allele bin creation. This was an effective means to extract a vector of user-supplied alleles to create panels to be used in *score.easy* and also making panel construction much simpler and faster than other methods implemented in the licensed fragment analysis software.

Scoring and multiplexing

Incorporation of multiple fluorescence dyes into DNA during PCR amplification using M13 allows researchers to reduce the cost of fragment analysis by pooling PCR products from multiple markers into the same capillary, and then later separating the fragments by

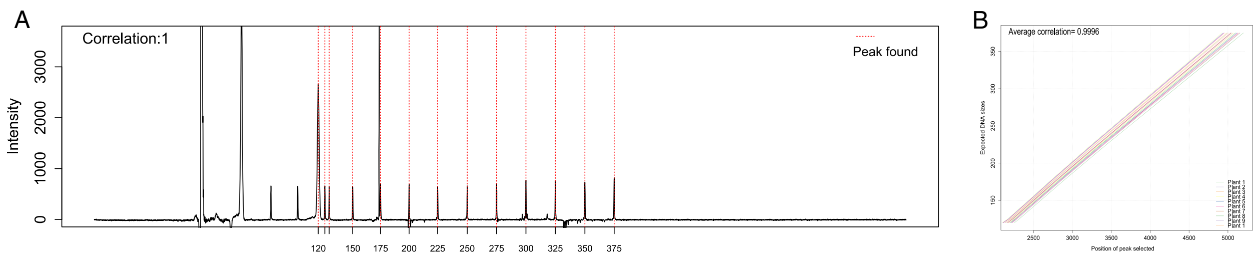

Fig. 2 Size standard calibration features of *Fragman*. In (**a**) An example of zero slope peaks selected in the 4th (*red*) channel, showing a correlation greater than 0.9999 with the expected size standard fragment lengths selecting the right peaks. In (**b**) correlation between expected DNA fragment in the Rox375 size standard (y-axis) and observed zero slope peaks for the size standard selected by *Fragman* in 10 random FSA files (x-axis)

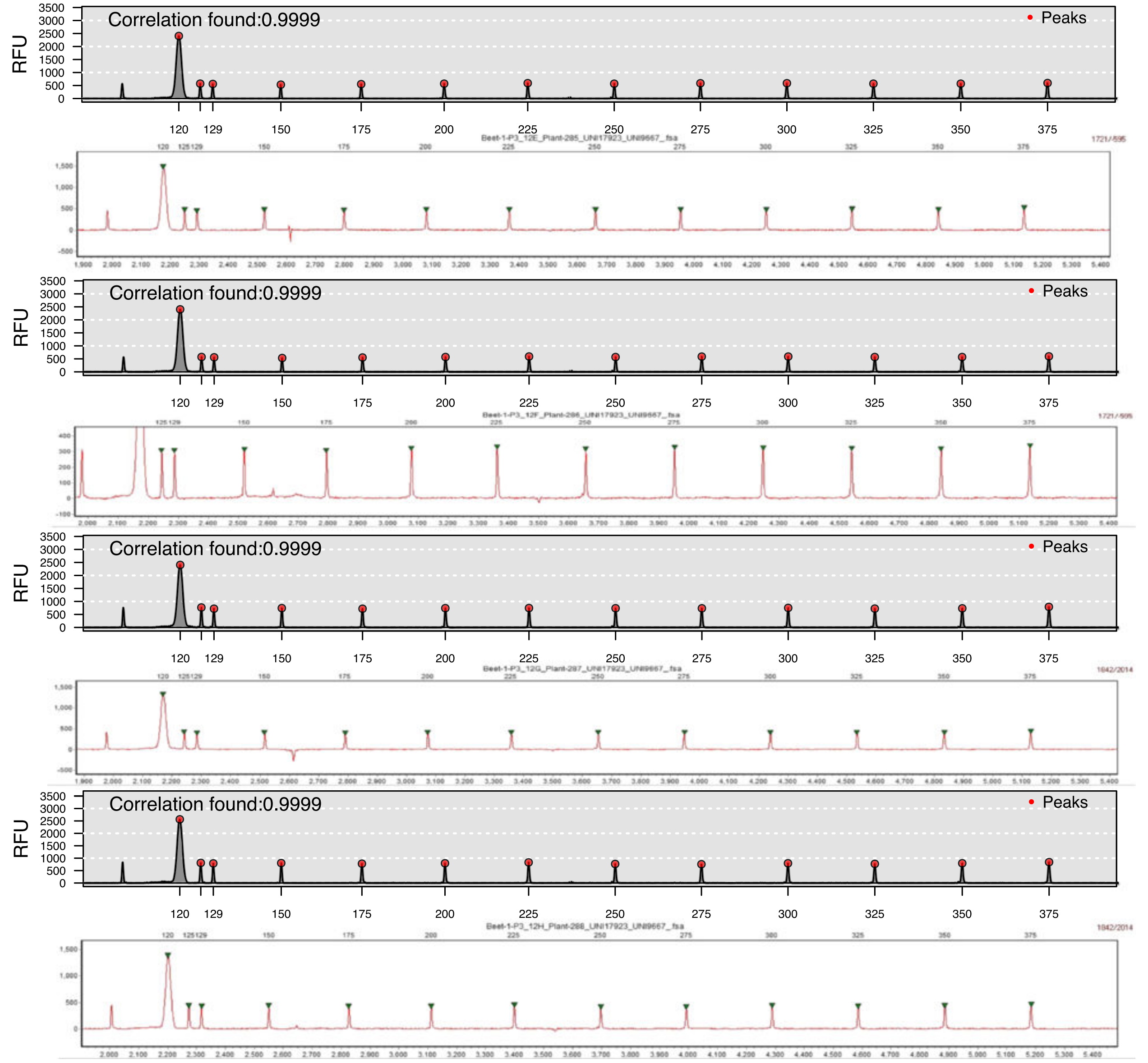

Fig. 3 Ladder sizing comparison. Comparison between the size standard calibration capabilities of the *Fragman* package (odd positions to the bottom) and the licensed GeneMarker® software (even positions to the bottom) for the same 4 samples. In x axis the base pair size is displayed whereas the intensity is plotted in the y axis. Samples scored with dotted red lines correspond to *Fragman* whereas samples scored with green dots correspond to GeneMarker®

marker into multiple channels based on their fluorescence at unique emission wavelengths [13]. Furthermore, multiple markers can be combined into a single channel as long as there is no overlap in the allele ranges of the designed panels. We tested the ability of the *Fragman* package to deal with any number of dyes and markers per channel compared to commercial software used for fragment analysis. When using the *score.easy* function in *Fragman* and GeneMarker® to determine the fragment lengths of several markers in 1000 individual FSA files derived from 3 biparental populations and a diversity population, we found that up 98 % of the samples were scored correctly using *Fragman*, which implements the functions *left.cond* and *right.cond*, whereas only 85 % where accurately scored using the commercial software GeneMarker® prior to manually deleting or adjusting miscalled peaks (Fig. 5).

Fragman analysis software was designed to select the best alleles based on the ploidy of the organism; however, this process tends to be highly error-prone due to differential patterns of PCR amplification that lead

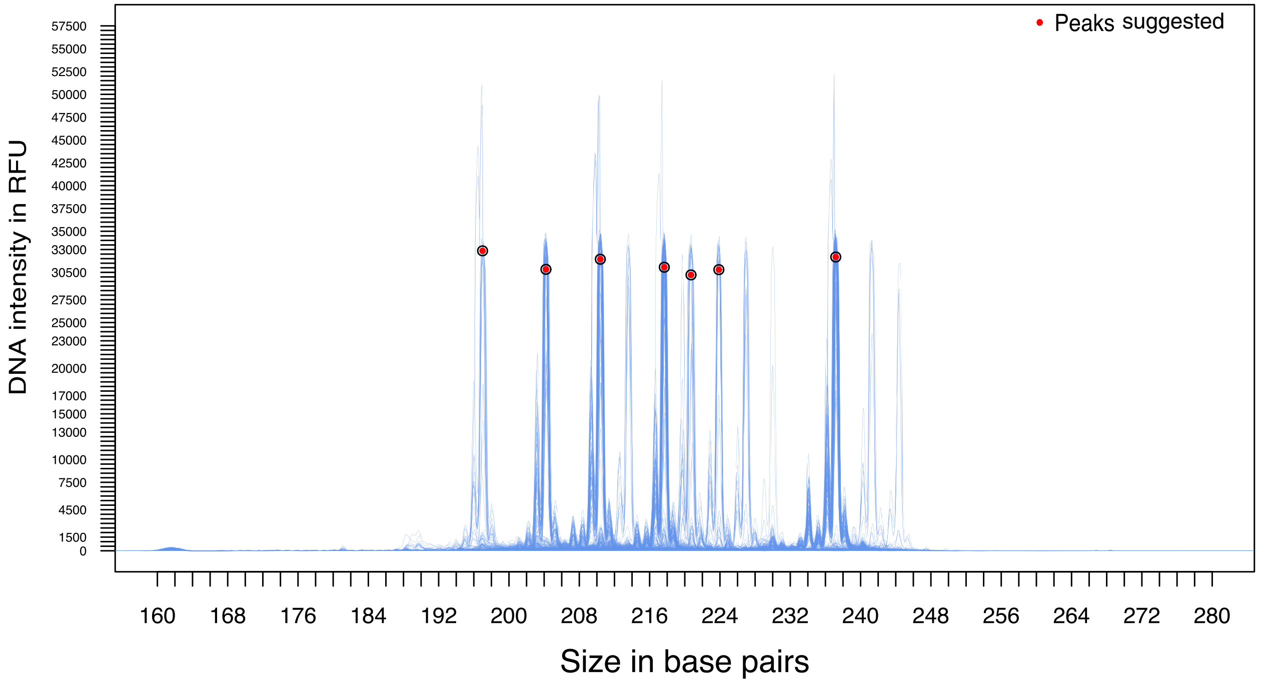

Fig. 4 Panel construction in *Fragman* package. Visual output of the *overview2* function implemented in *Fragman* to create panels of potential alleles found in a population. The function overlaps the fluorescent intensities in all the loaded FSA files for the selected channel. Here, the fluorescent emission intensity of FAM labeled DNA fragments in 288 diploid plants from a diversity panel are plotted to show the allelic variability for this genetic marker. The function by default suggests and returns all peaks with minor allele frequency (MAF) > 0.05 along with a plot which can be used interactively with the use of the *locator* function to manually select alleles

to background peaks and stuttering resulting from incomplete 3' nucleotide addition [5]. For polyploid organisms, *Fragman* will call all zero slope peaks above a selected threshold when users are scoring non-diploid organisms. Additional functions were added to *Fragman* for formatting the fragment lengths measured in *score.easy* according to the required inputs for the common linkage mapping and diversity analysis software, JoinMap® [6], OneMap [8] and GenAlEx [7]. These functions allow users to avoid time-consuming manual conversion of data types between software, and the same functions can also be used to convert SNP calls from by genotyping by sequencing (GBS) to JoinMap® and OneMap as well [14].

Comparing Fragman with other existing fragment analysis software

Fragman offers a full fragment analysis pipeline comparable to GeneMarker® consisting of three main steps: 1) FSA files are read and sized according to a ladder, 2) data is scored using bins established by the user, and 3) data is exported as an excel file to be used for linkage mapping or other genetic analyses. To our knowledge, there is no other free software package available that performs all the same services as *Fragman*. The only other freely available, platform independent software comparable to *Fragman* is Peak Studio, written in Java, but during our tests it often crashed with large sample numbers and it did not allow for automation and is no longer being updated [15]. Other R packages that could be comparable to *Fragman* include Genomatic (unpublished), Biostrings (unpublished), and MsatAllele [16], but they all rely on licensed software such as GeneMapper® or GeneMarker® or the windows dependent STRAND [17] (Additional file 1). Thus, these packages independently do not provide a full fragment analysis pipeline, which includes sizing, peak calling, and data export to other software for genetic analyses. *Fragman* is a complete fragment analysis, software package, which provides most if not all of the services available in licensed software while also implementing several new previously unavailable and automated functions useful in downstream genetic analyses.

Conclusions

We have developed an R package with the ability to perform efficient and accurate fragment analysis by taking advantage of the power of R [11], which provides extra graphical and high-throughput capabilities for high dimensional projects.

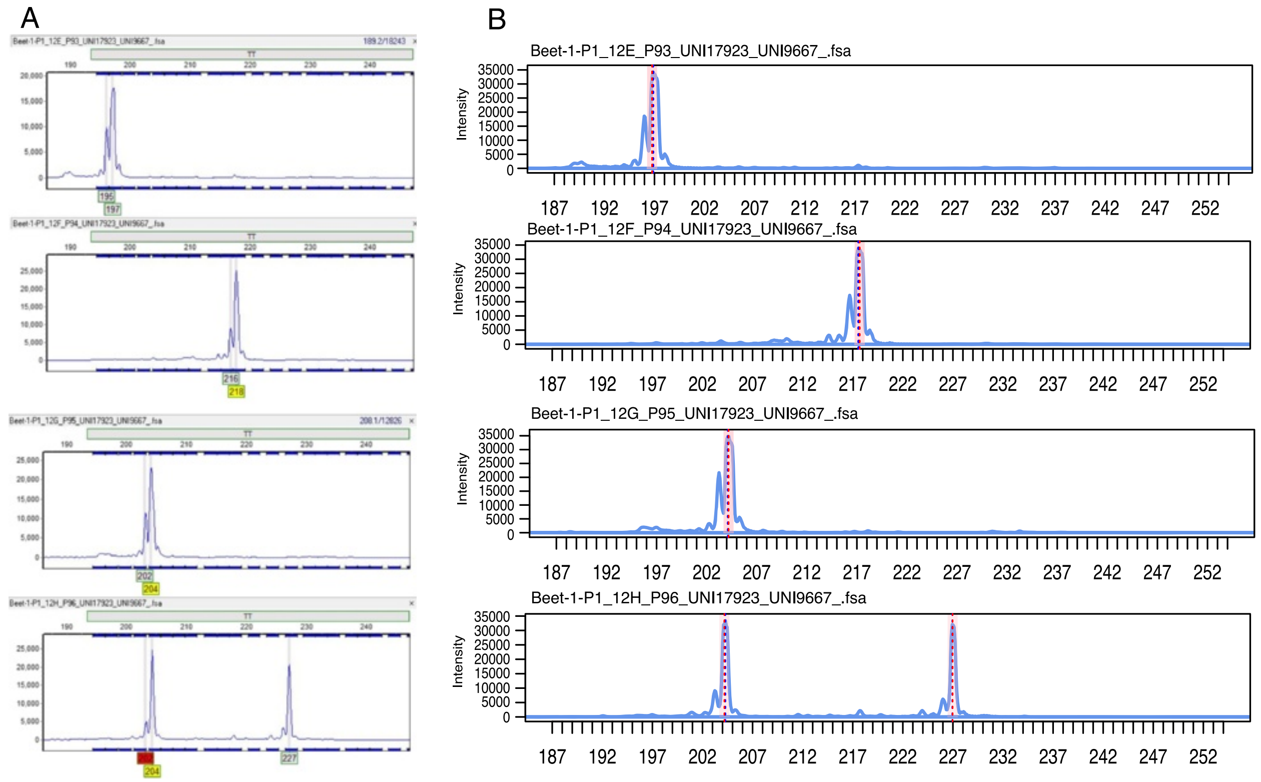

Fig. 5 Scoring comparison between *Fragman* and GeneMarker®. **a** Example of GeneMarker® scoring compared to (**b**) *Fragman* scoring (zero slope peak selection) in four random FSA files, showing the ability of *Fragman* to better differentiate between real and noisy peaks compared with GeneMarker®, which usually picks noisy peaks as real

Major features of *Fragman* are:

1. The ability to directly load FSA files for fragment analysis
2. The ability to work with any DNA standard (ladder)
3. The ability to automatically and simultaneously analyze multiple samples
4. Flexibility in the scoring preference (i.e. manual or automatic scoring)
5. Accurate automatic scoring of diversity and biparental populations
6. Additional tools for analyzing and converting the data into other formats

Availability and requirements

Project name: *Fragman*: an R package for fragment analysis

Project home page: https://cran.r-project.org and http://cggl.horticulture.wisc.edu

Operating system(s): Platform independent

Programming language: R

Other requirements: R > 2.0

License: GPL-3

Ethics (and consent to participate)

Not applicable.

Consent to publish

Not applicable.

Additional file

Additional file 1: Features of other computer software available to analyze fragment data. (DOCX 126 kb)

Abbreviations

Bp: base pair; *DNA*: Deoxyribonucleic acid; FAM, HEX, NED, ROX and PET: dyes used for DNA sequencing. All four dyes can be excited at a single wavelength (488 nm), but emit at distinctly different wavelengths; *FSA*: File associated with Applied Biosystems, usually output of DNA fragment analysis. A type of FASTA file format; MAF: minor allele frequency; *M13*: universal primer for multiplexing; *SNP*: single nucleotide polymorphisms; *SSR*: simple sequence repeat.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GCP, LDG, BS and JZ conceived the software. GCP wrote the software. GCP, WS and LDG conducted the experiments and analyzed the data. GCP, BS and JZ wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

JZ and BS want to express their gratitude through 1Co10:31. We thank to the anonymous reviewers who helped to enhance the quality of this paper.

Funding

This project was supported by USDA-ARS (project no. 3655-21220-001-00 provided to JZ and SS); WI-DATCP (SCBG Project #14-002); National Science Foundation (DBI-1228280); Ocean Spray Cranberries, Inc.; Wisconsin Cranberry Growers Association; Cranberry Institute; BS was supported by the Frank B. Koller Cranberry Fellowship Fund for Graduate Students; GCP and LDG were supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT, Mexico).

Author details

[1]Department of Horticulture, University of Wisconsin, Madison, WI, USA. [2]Instituto Nacional de Investigaciones Forestales, Agricolas, y Pecuarias, Campo Experimental Pabellon, Aguascalientes, Mexico. [3]USDA-ARS, Vegetable Crops Research Unit, University of Wisconsin, Madison, WI, USA.

Received: 14 October 2015 Accepted: 7 April 2016
Published online: 21 April 2016

References

1. Yang X, Xu Y, Shah T, Li H, Han Z, Li J, Yan J. Comparison of SSRs and SNPs in assessment of genetic relatedness in maize. Genetica. 2011;139:1045–54.
2. Van Inghelandt D, Melchinger AE, Lebreton C, Stich B. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet. 2010;120:1289–99.
3. Bonin A, Ehrich D, Manel S. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Mol Ecol. 2007;16:3737–58.
4. Grossman P, Colburn J. Capillary electrophoresis: Theory and practice. 2nd ed. Boca Raton: CRC Press; 1997.
5. Applied Biosystems: DNA Fragment Analysis by Capillary Electrophoresis. User Guide. Life technologies 2014, 211 pp. https://tools.thermofisher.com/content/sfs/manuals/4474504.pdf. Accessed August 2015.
6. Ooijen Van JW: JoinMap 4®. Users guide, Kyazma. 2006.
7. Peakall R, Smouse PE. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics. 2012;28:2537–9.
8. Margarido GR, Souza P, Garcia F. OneMap: Software for Genetic Mapping in Outcrossing Species. Hereditas. 2007;144(3):78–9. doi:10.1111/j.2007.0018-0661.02000.x.
9. Charif D, Lobry JR. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: Bastolla U, Porto M, Roman E, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations. New York: Springer Verlag; 2007. p. 207–32.
10. Softgenetics: Genemarker® the biologist friendly software. User Guide, Softgenetics 2012.
11. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2015. http://www.R-project.org/.
12. Chambers JM, Hastie TJ. Statistical Models in S: Linear models. 1st ed. Pacific Grove: Wadsworth & Brooks/Cole; 1992.
13. Schuelke M. An economic method for the fluorescent labeling of PCR fragments: A poor man's approach to genotyping for research and high-throughput diagnostics. Nature Biotech. 2000;18:233–4.
14. Elshire RJ, Glaubitz JC, Sun Q, Poland J, Kawamoto K, Buckler ES, Mitchell SE. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS One. 2011;6(5):e19379. doi:10.1371/journal.pone.0019379.
15. McCafferty J, Reid R, Spencer M, Hamp T, Fodor A. Peak Studio: a tool for the visualization and analysis of fragment analysis files. Environ Microbiol Rep. 2012;4:556–61.
16. Alberto F. MsatAllele_1.0: An R Package to Visualize the Binning of Microsatellite Alleles. J Hered. 2009;100(3):394–7.
17. Toonen RJ, Hughes S. Increased Throughput for Fragment Analysis on ABI Prism 377 Automated Sequencer Using a Membrane Comb and STRand Software. Biotechniques. 2001;31:1320–4.