STRyper: an open source macOS application for microsatellite genotyping

Jean Peccoud

Laboratoire Écologie et Biologie des Interactions, Équipe Écologie Évolution Symbiose, Université de Poitiers, UMR CNRS 7267, Bât. B31, 3 rue Jacques Fort, TSA 51106, 86073 Poitiers Cedex 9, France

jeanpeccoud@gmail.com

Running title: An application for microsatellite genotyping

Statements relating to ethics and integrity policies

- STRyper and its source code are available at https://github.com/jeanlain/STRyper/
- The author declares no conflict of interest.
- Part of this work was funded by intramural funds from the CNRS and the University of Poitiers, and by Agence Nationale de la Recherche grant ANR-20-CE02-0004 (SymChroSex).

Abstract

In the context of population genetics, microsatellite markers analyzed by capillary sequencing remain useful tools for quick genotyping and low-cost studies. This contrasts with the lack of a free application to analyze chromatograms for microsatellite genotyping that is designed for population geneticists. To fill this gap, I have developed STRyper, a macOS application whose source code is published under the General Public License. STRyper only uses macOS libraries, making it very lightweight, responsive, and behaving like a modern application. Its three-pane window enables easy management, searching and viewing of chromatograms imported from .fsa and. hid files, the creation of size standards and of microsatellite marker panels (including bins). The application has unique features allowing DNA ladder and genotype correction by drag-and-drop, and the management of variations in electrophoretic conditions. STRyper is available at https://github.com/jeanlain/STRyper.

Keywords: microsatellites, capillary electrophoresis, chromatograms, population genetics, graphical user interface

Introduction

More than a decade after the advent of next generation sequencing (NGS) technologies, microsatellites, also known as short tandem repeat (STR) loci, remain popular DNA markers. Microsatellite genotyping without DNA sequencing indeed offers a compelling money- and time-saving solution to assess gene flow, population history, structure and membership, ancestry, or the integrity of laboratory breeding lines, among other uses.

The reasonable cost per individual of microsatellite marker amplification and capillary-based electrophoresis contrasts with the cost of software solutions dedicated to the analysis of the resulting chromatogram files. As opposed to NGS data, which are generally analyzed with free command line tools, microsatellite genotyping requires inspecting fluorescence curves, hence scientific applications with a graphical user interface, which are rarely free. To various degrees, these applications are focused on human identification and forensics. They are therefore packed with features and safeguards that are of little relevance to molecular ecologists, which somewhat complicate their use, and which may come with a high price.

This is the case of GeneMapper by ThermoFisher Scientific, a commercial application running on the Windows operating system, and which remains, to my knowledge, the most widely used in our field for microsatellite genotyping. The price of a GeneMapper license may restrict its installation to a single computer per research laboratory. Geneious (http://www.geneious.com/) and its microsatellite analysis plugin represents a more affordable alternative. Still, the cost of a subscription to Geneious may deter population geneticists who do not need all the features that Geneious offers for the analysis of DNA sequences.

The Osiris software (Goor et al. 2011, Goor et al. 2021), by the National Institute of Health (https://www.ncbi.nlm.nih.gov/osiris/), stands out as being a free, cross-platform (Windows and macOS) tool for STR analysis. Yet, it is rarely used by population geneticists, as far as I know. The specialization of this tool in human identification makes it less suitable to the analysis of other species, from my experience with this application.

Population geneticists would therefore benefit from a free application enabling quick microsatellite genotyping of hundreds of samples. To meet this need, I present STRyper, an open-source, lightweight and user-friendly application that can analyze chromatogram files for STR genotyping. STRyper is published under the GNU General Public License v. 3 and its name is the contraction of “STR“ and “Genotyper”. The application and the codebase are available at https://github.com/jeanlain/STRyper.

Methods

The efficiency of an application designed for microsatellite genotyping mostly relies on its user interface (UI), as chromatograms must be visually checked and genotypes validated without a command line. However, underlying methods of fluorescence data analysis for automatic genotype calling will be described first and in greater details than design choices, which are not a scientific mater. Those will be covered in the results section which describe UI.

Fluorescence data analysis

The analysis of fluorescence data in chromatograms starts with the delineation of peaks, for which a simple algorithm was developed. This algorithm (detailed in the supplementary text) determines whether a fluorescence data point (a “scan”) is elevated enough, both relative to neighbor scans and in absolute level. Peak delineation serves as a basis to subtract baseline fluorescence level, which helps curve and peak interpretation. The method developed for this task adjusts the height (fluorescence level) of a curve such that the start and end point of each peak a placed at level zero (figure S1). Although this adjustment cannot be applied on signals that are too faint to contain meaningful peaks, it has the benefit of offering two baseline subtraction modes: one that preserves absolute peak height, and one that maintains relative peak elevation compared to the baseline (supplementary text and Figure S1).

In STR genotyping, it is crucial to determine whether a peak represents a DNA fragment or interference from another channel (“crosstalk”). The method developed for this task compares the position, shape and relative size of peaks between channels, accounting for saturation of the sequencer camera (supplementary text). While certain applications alter fluorescence data to correct for pull-up due to crosstalk (Goor et al. 2021), it was decided that flagging peaks resulting from crosstalk and leaving the source signal untouched was sufficient. The approach adopted simply ignores these peaks in automatic allele calling and size assignment of peaks found in the DNA ladder (detailed below).

Peak assignment

Two types of peaks in chromatograms must be assigned: those that correspond to DNA ladder fragments, in the context of sample sizing, and those corresponding to alleles in the context of genotyping.

The method used to detect DNA ladder fragments and assign them to sizes of a known size standard is based on relative peak positions and accounts for non-linear relationship between fragment size and migration speed (supplementary text). Peaks resulting from crosstalk or whose height are unusual compared to others are ignored. To account for non-linearity, a polynomial of the first, second, or third degree (depending on the user choice) is used to estimate fragment size, where the response variable is the size of a fragment specified in the size standard, and the explanatory variable is the scan numbers at the tip of the corresponding peak (representing migration speed). This principle is also implemented in other applications such as GeneMapper. Fitting is achieved via the Cholesky decomposition implemented in the Linear Algebra Package (https://netlib.org/lapack/). A score of sizing quality from 0 to 1 was developed, based on the residuals of the fitted model (differences between fragment sizes as defined in the size standard, and fragment sizes estimated by the model). This score involves computing the difference in residuals for every pair of adjacent peaks. If the maximum of this difference is large, the quality score is reduced (often to zero) to indicate a possible error in the size attributed to one or several peaks.

As for allele calling, automatic genotyping must account for two main biochemical processes producing DNA fragments of different lengths. One is the addition of a non-template nucleotide to the 3’ end of the new DNA strand by the DNA polymerase during PCR (Clark 1988). Because the added nucleotide is generally an adenosine, this process is referred to as “adenylation”. If adenylation affects only part of the replications, amplicons may differ in length by one nucleotide, generating two peaks. The other process is “splippage” during replication, causing indels in the repeated region (Hauge et al. 1993). These events result in amplicons whose length vary according to the size of the repeat, a pattern known as “stuttering”. These considerations served as a basis to develop a method for allele calling that first identifies peak clusters resulting from these processes (ignoring again peaks resulting from crosstalk). In each delineated cluster, the most intense peak is considered as that representing the allele. Estimation of peak intensity accounts for clipping due to saturation of the fluorescence signal, in that the width of the saturated region is used when peak height/area may not reflect the quantity of DNA material. Stutter and adenylation are treated internally by the application and are not communicated to the user, who is free to manually assign an allele to any peak. I did not consider such information as critical to researchers in our field, considering that STRyper is not designed for human identification and that its capabilities should not affect human lives.

Genotyping

Identifying alleles requires a user-defined range of expected allele sizes at a microsatellite marker. For a diploid individual, the number of different alleles detected within that range simply determine the sample’s genotype: homozygous if one allele is detected, heterozygous otherwise. Because this inference is invalid for polyploid markers, it was decided that only haploid and diploid markers could be defined in the application, constraining the number or alleles per locus to 2. While this limitation should not affect most users, the ability to annotate additional peaks was considered necessary to allow studying polyploid species, to indicate sample contamination or the existence of paralogs of the STR marker. Development of this feature however had to consider the nuisance of additional annotated peaks, which often require users to manually discard peaks during genotype inspection (at least in my experience with Genemapper up to version 4.1). This nuisance was easily avoided by the distinction of two types of peaks: those that are interpreted as alleles – whose number is limited to the ploidy of the marker – and others that represent additional DNA fragments. The latter are detected during genotype calling like alleles are (i.e., by identifying peak clusters) to avoid annotating peaks that may amount to noise, stuttering or adenylation. The relative height of peaks is used to categorize alleles (taller peaks) and additional peaks (shorter peaks). The annotation of additional peaks is optional and can be modified by the user without altering the sample's genotype. With this feature developed, it was decided that no genotype quality score needed to be computed. Barring some trained artificial intelligence (Taylor et al. 2016), which I could not realistically implement, I considered that no algorithm can yet usefully complement the visual assessment of chromatograms when it comes to the reliability of genotyping.

Another important aspect of allele calling is assignment of alleles to sizes in base pairs, which are integer numbers, as opposed to peak sizes estimated by fitting a model using DNA ladder fragments, as described previously. The binning approach used by other applications was adopted. If the size of an allele falls in a user-defined range (a so-called "bin"), the allele takes the bin name. By default, a bin name is the rounded size of its midpoint when it was first created, but it can be manually modified to any Unicode string.

Binning however doesn't account for variations in electrophoretic conditions that may differently affect molecular ladder fragments and amplicons. Such variations can shift the estimate size of alleles between runs, defeating the purpose of bins. Rather than managing multiple bin sets per marker (as in Genemapper), a new approach was adpoted to address this issue and considers that it is the sizes of alleles, not the position of bins, which should be corrected. The method thus corrects peak sizes using the formula $y = a + bx$, where x is the "uncorrected" size of a DNA fragment estimated by the fitted model mentioned earlier, y is the size that will be used for peaks found in the marker range, and a and b are constants. If there is no correction, a = 0 and b = 1. This approach assumes that the effect of varying electrophoresis conditions can be approximated by this linear combination. The implementation of this method does not require the user to compute nor enter any value and is described in Figure S2.

Development of the application

Developing an application with a complex UI greatly depends on the target operating system and development tools. These were dictated by my use of the Mac operating system (macOS) at work and at home, and by the fact that developing STRyper was a hobby project of an evolutionary biologist, not the effort of a team of professional developers. Being unencumbered by cross-platform development gave me the freedom to choose the right tools to program a UI that was intuitive, responsive and consistent with other macOS applications. STRyper was thus developed using Xcode and macOS object-oriented frameworks (https://developer.apple.com/library/archive/documentation/MacOSX/Conceptual/OSX_Technology_Overview/SystemFrameworks/SystemFrameworks.html). These frameworks include "Core Data" da, which is used to define and manage objects representing chromatograms, marker panels, bins, alleles, genotypes and size standards, and to save them in a persistent relational database. Development also relied on "AppKit" classes for most UI elements, "Core Graphics" functions to draw fluorescent curves, and "Core Animation" layers to accelerate compositing via the graphical processing unit (GPU) and to provide fluid animation for certain elements. The application was entirely written in objective-C, a superset of C. This language was required to use these frameworks (except Core Graphics, which is C-based) when the project started.

Testing the application

Version 1.0 beta of the application was submitted to a "real world" test on newly generated chromatograms. I did not conduct extensive tests on previously published data because chromatograms to which I have access were analyzed with outdated versions of

Genemapper, making any comparative analysis meaningless, and because I considered more profitable to use STRyper for a new study (Vucić et al., in prep).

The chromatograms were obtained from 324 individuals of *Phoxinus lumaireul* (Teleostei, Cypriniformes), each amplified at two 6-plexes of microsatellite markers (Vucić et al. 2022) analyzed with an ABI 3200 sequencer using five fluorescent dyes. The test consisted in importing the 648 chromatograms into the application, applying the size standard (a Genescan 500-LIZ size standard from which sizes 35 and 250 were removed), checking the assignment of DNA ladder fragments, defining marker and bins, applying marker offsets if necessary, calling genotypes, visually checking called genotyped (correcting them if necessary) and exporting them as a text file. These tasks were performed on a Mac Studio equipped with an M1 Max processor.

Results

General characteristics

STRyper runs under macOS version 10.13 or higher. The application does not contain third-party libraries and does not require special installation steps. Its bundle contains binaries compiled for the X86 and arm64 architectures and weighs less than 15 Megabytes, including the user guide.

The application has one main window (Figure 1) composed of three panes; a design paradigm used by several database-management applications like email clients. The left collapsible sidebar is a hierarchical list of folders and subfolders containing samples (like mailboxes contain messages). Folder and samples can be organized freely my drag and drop. A middle pane shows the content of the selected folder (samples and associated genotypes) and comprises additional tabs to manage size standards and markers. The right pane shows the traces (fluorescent curves) of selected samples and genotypes, much like mail clients show the content of selected messages.

STRyper uses very few modal panels or dialogs to validate user actions and all actions that affect the database can be undone. Most are at a couple of clicks away or less as they do not require opening and closing windows. Drag and drop can be used throughout: from importing samples to applying size standards, markers, and to manually attributing alleles or size molecular ladder fragments to peaks.

STRyper has no concept of “projects” that must be saved and closed before opening another, neither does it require setting analyzes before viewing samples. STRyper can import FSA files (HID file support is experimental, as the HID format specifications are not public) containing data for 4 or 5 channels (fluorescent dyes). Samples are imported into folders and they can be moved or copied between folders at any time. A folder and all its content, including edited genotypes at microsatellite markers and custom size standards, can be archived and transferred between instances of the application. Upon importing an archived folder, any marker panel and size standard encoded in the archive is imported unless is it already in the database.

Since samples are not constrained to compartmentalized projects, the application provides search tools to find and gather samples from the whole database. Users can define various search criteria, including run date, sizing quality, well identifier, plate name, marker panel name, etc. Search results appear in “smart folders” which dynamically update their contents as new samples meet the search criteria. These smart folders behave like smart mailboxes.

Chromatogram viewing

Selecting a folder of the database shows all its samples, and associated genotypes if a panel of microsatellite markers have been applied to the samples. Samples can be filtered and sorted by various metadata items constituting columns that can be hidden and reordered. An inspector panel dynamically updates to show information about selected samples, including sizing information (Figure 2).

Chromatograms are displayed on the right pane. As the application fully supports the dark theme of macOS (version 10.14 or more recent), it can display fluorescent curves (“traces”) on a dark background to mitigate eye strain. Any region in which a peak statured the sequencer camera is shown behind curves as a rectangle whose color reflects the channel that likely caused saturation. Traces can be scrolled and zoomed in/out horizontally via trackpad gestures such as swipe, pinch and double tap, via the scroll wheel, or by dragging the mouse over horizontal rulers to define a size range. Dragging the mouse over the vertical ruler allows setting the fluorescence level at the top of the view, hence the vertical scale.

Zooming is animated, which helps users keep track of the range (in base pairs) that is displayed. On computers equipped with an Apple chipset (M1 or more recent), the drawing of curves is accelerated by the GPU and is therefore very efficient. In my tests on a laptop equipped with an M1 Pro chipset, hundreds of stacked fluorescence curves can be zoomed in and out without skipping a frame on the 120 Hz integrated high-resolution display.

Viewing options include automatic vertical scaling to the highest visible peaks, synchronizing of vertical scales and horizontal positions, showing/hiding bins and region of fluorescence saturation, and stacking curves from several samples or channels in the same view. Another viewing option allows users to identify peaks resulting from crosstalk by painting areas underneath these peaks with the color of the channel that was inferred to induce crosstalk. This option helps users avoid considering these peaks as alleles or DNA ladder fragments and makes clear why they were ignored during automatic genotyping.

To apply molecular ladders, STRyper comes with several widely used size standards, namely those from the GeneScan brand. Users can easily edit these size standards within the application and make their own. STRyper displays the trace of the molecular ladder like any other trace, letting users switch spontaneously between genotype and molecular ladder editing. Sizes attributed to molecular ladder fragment can be changed by dragging and dropping size labels onto peaks. Any change to the molecular ladder automatically updates the sizing of the sample without user validation.

Genotyping

Users can define their own panels of haploid or diploid microsatellite markers within STRyper and organize them into folders. Markers are defined by their fluorescent dye, ploidy, length of repeat motive, name, and the size range of their alleles. These attributes can be changed after a marker is created, except for the first two. Markers can be copied between panels. Users can export marker panels (which contain bins) to text files conforming to simple specifications. These text files can be imported back as marker panels.

A set of automatically named bins for a marker can be added by specifying the width and spacing of bins. The position and width of the whole be set can then be adjusted by clicking and dragging. Bins can also be added/removed/modified individually via click and drag.

These actions do not involve dedicated windows or panels, they can be performed at any time on the trace views where bins show (Figure 1, right pane).

All genotypes from displayed samples are listed in a table that can be sorted and filtered according to various criteria (including allele names and sizes). This table lets users quickly scan genotypes, as corresponding peaks and allele labels of the selected genotype(s) appear on the right-pane. Correcting errors in allele call typically takes a single step that does not require selecting the correct allele name from a list. Instead, users can simply drag the mouse from a peak to a bin, drag an allele label from one peak to another (Figure 3), or double-click a peak, which removes/attaches an allele from/to the peak. Double clicking allele labels lets users enter arbitrary allele names directly above peaks.

Genotypes and associated sample metadata can be exported as text files or simply copied from selected table rows to a text editor or a spreadsheet application.

Test results

Importation of the 648 chromatogram files took less than two seconds and application of the size standard took about one second. Ignoring electrophoresis failures than rendered certain samples unusable, assignments of peaks to ladder fragments required manual corrections in less than ten samples. In every case, a size was simply not assigned to any peak because the appropriate peak was too short compared to neighboring peaks. Overall, the verification of the DNA ladder for all samples took less than ten minutes.

Defining DNA markers and bins took a couple of minutes per marker. Automatic genotyping of all usable samples (3600 genotypes) was achieved in less than one second. Ignoring obvious PCR failures and apparent contamination leading to multiples peak per marker, visual inspection of the genotypes identified two main causes of genotyping errors. One involved peaks differing in size by just one nucleotide at a marker probably affected by a 1-bp indel. These peaks were wrongly interpreted as caused by adenylation, and one allele was inferred instead of two. Only comparisons between individuals (which the program does not perform during automatic genotyping) revealed that these individuals were likely heterozygous. The other common source of error was due to varying degrees of adenylation at certain markers. As a single allele causes two peaks (or more, due to stuttering), the one that is most intense (“taller”) depends on the degree of adenylation, which may vary between PCRs (the degree of stuttering, however, is rather constant). In this situation, one should always use the sorter, or always use the longer peak, as the one representing the allele at a given marker. However, the application always uses the taller peak. Other peak assignment errors where rare (affecting less than 1% of the genotypes) and mostly involved failures to identify cases of crosstalk, or extremely strong allele dropout due to a very large size difference between alleles (> 60 bp).

Discussion and conclusion

STRyper is focused on features that are relevant to population geneticists, who cannot afford spending as much time on an individual genotype as forensic researchers do. Indeed, since STRyper is not designed for human identification, it does not assume that allele calls are reviewed by several users. As a result, it does not record the history of manual corrections applied to genotypes (but still allows adding comments on genotypes). This feature does not seem useful to population geneticists as it would mostly clutter the user interface. Also, the number of peaks assigned to alleles never exceeds the marker's ploidy. Hence, users should rarely need to remove supplementary peaks to correct the genotype that was called. These attributes should make STRyper particularly adapted to the quick review hundreds or thousands of genotypes with a limited number of clicks, making this task less tedious.

To conclude, STRyper's strengths mostly rely in its simple and responsive interface. The current restriction of STRyper to macOS is partially balanced by its free nature (the cost of an entry-level Mac is less than that of paid applications used for microsatellite genotyping), its responsiveness, and its "native" feel, which is rare among scientific applications.

Acknowledgements

I thank Dr. Douglass Hoffman from the United States National Institute of Health for his advice on decoding HID files, and numerous colleagues for testing STRyper. I also thank M. Vucić and F. Grandjean for giving access to *Phoxinus* chromatogram files. Part of this work was funded by intramural funds from the CNRS and the University of Poitiers, and by Agence Nationale de la Recherche grant ANR-20-CE02-0004 (SymChroSex).

References

Clark, J. M. (1988). "Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases." Nucleic Acids Res **16**(20): 9677-9686.10.1093/nar/16.20.9677

Goor, R. M., L. Forman Neall, D. Hoffman and S. T. Sherry (2011). "A mathematical approach to the analysis of multiplex DNA profiles." Bull Math Biol **73**(8): 1909-1931.10.1007/s11538-010-9598-0

Goor, R. M., D. Hoffman and G. R. Riley (2021). "Novel Method for Accurately Assessing Pull-up Artifacts in STR Analysis." Forensic Science International: Genetics **51**.10.1016/j.fsigen.2020.102410

Hauge, X. Y. and M. Litt (1993). "A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR." Hum Mol Genet **2**(4): 411-415.10.1093/hmg/2.4.411

Taylor, D. and D. Powers (2016). "Teaching artificial intelligence to read electropherograms." Forensic Science International: Genetics **25**: 10-18.10.1016/j.fsigen.2016.07.013

Vucić, M., M. Jelić, G. Klobučar, D. Jelić, H. M. Gan, C. Austin, D. Guyonnet, I. Giraud, T. Becking and F. Grandjean (2022). "A new set of microsatellite markers for Phoxinus lumaireul senso lato, Phoxinus marsilii and Phoxinus krkae for population and molecular taxonomic studies." Journal of Fish Biology **101**(5): 1225-1234.https://doi.org/10.1111/jfb.15194

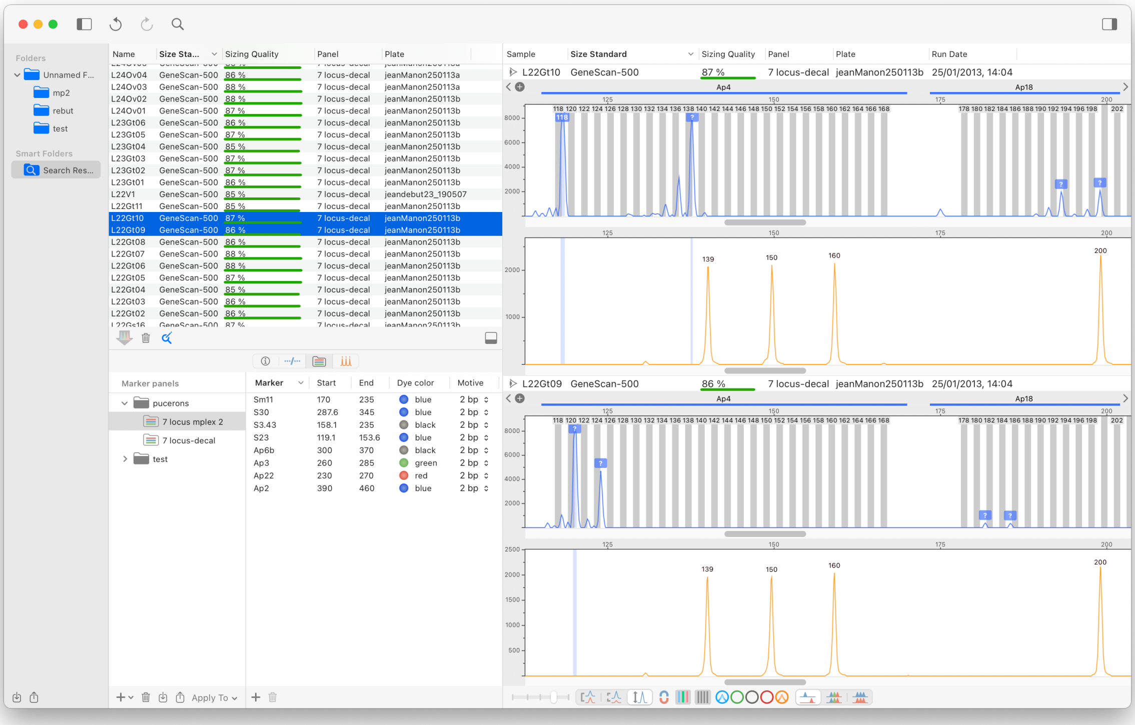
Data accessibility

STRyper and its source code are available at https://github.com/jeanlain/STRyper/

Author Contributions

JP developed the application and wrote the paper.

Figures

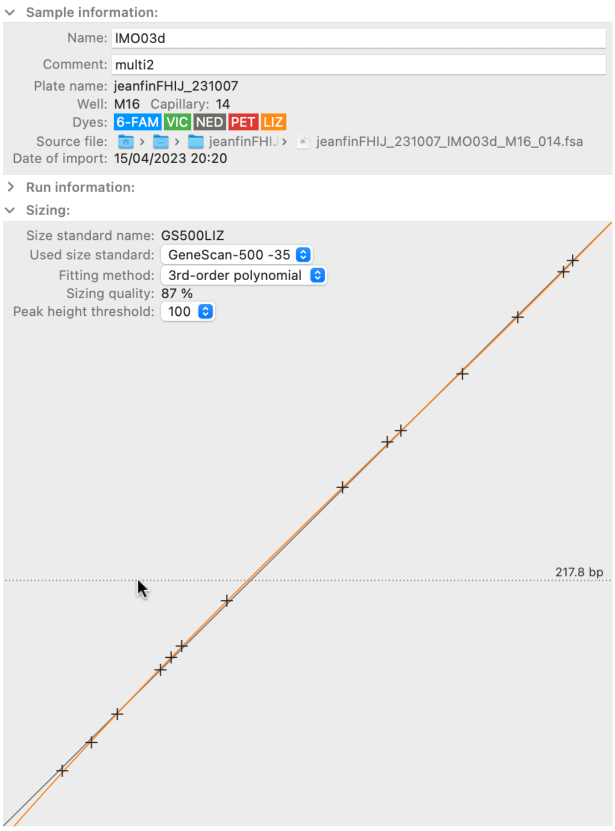


Name	Size Sta...	Sizing Quality	Panel	Plate
L24Ov04	GeneScan-500	86 %	7 locus-decal	jeanManon250113a
L24Ov03	GeneScan-500	88 %	7 locus-decal	jeanManon250113a
L24Ov02	GeneScan-500	88 %	7 locus-decal	jeanManon250113b
L24Ov01	GeneScan-500	87 %	7 locus-decal	jeanManon250113b
L23Gt06	GeneScan-500	86 %	7 locus-decal	jeanManon250113b
L23Gt05	GeneScan-500	87 %	7 locus-decal	jeanManon250113b
L23Gt04	GeneScan-500	85 %	7 locus-decal	jeanManon250113b
L23Gt03	GeneScan-500	87 %	7 locus-decal	jeanManon250113b
L23Gt02	GeneScan-500	87 %	7 locus-decal	jeanManon250113b
L23Gt01	GeneScan-500	86 %	7 locus-decal	jeanManon250113b
L22V1	GeneScan-500	85 %	7 locus-decal	jeandebut23_190507
L22Gt11	GeneScan-500	85 %	7 locus-decal	jeanManon250113b
L22Gt10	GeneScan-500	87 %	7 locus-decal	jeanManon250113b
L22Gt09	GeneScan-500	86 %	7 locus-decal	jeanManon250113b
L22Gt08	GeneScan-500	86 %	7 locus-decal	jeanManon250113b
L22Gt07	GeneScan-500	88 %	7 locus-decal	jeanManon250113b
L22Gt06	GeneScan-500	88 %	7 locus-decal	jeanManon250113b
L22Gt05	GeneScan-500	87 %	7 locus-decal	jeanManon250113b
L22Gt04	GeneScan-500	85 %	7 locus-decal	jeanManon250113b
L22Gt03	GeneScan-500	86 %	7 locus-decal	jeanManon250113b
L22Gt02	GeneScan-500	86 %	7 locus-decal	jeanManon250113b
L22Gs16	GeneScan-500	87 %	7 locus-decal	jeanManon250113b

Marker	Start	End	Dye color	Motive
Sm11	170	235	● blue	2 bp
S30	287.6	345	● blue	2 bp
S3.43	158.1	235	● black	2 bp
S23	119.1	153.6	● blue	2 bp
Ap6b	300	370	● black	2 bp
Ap3	260	285	● green	2 bp
Ap22	230	270	● red	2 bp
Ap2	390	460	● blue	2 bp

Sample	Size Standard	Sizing Quality	Panel	Plate	Run Date
L22Gt10	GeneScan-500	87 %	7 locus-decal	jeanManon250113b	25/01/2013, 14:04
L22Gt09	GeneScan-500	86 %	7 locus-decal	jeanManon250113b	25/01/2013, 14:04

Figure 1. The main window of STRyper. The left pane contains the list of folders and smart folders (search results) containing samples. The middle pane is a split view comprising a top pane listing the samples of the selected folder. Its bottom pane has four tabs, which are from left to right: an inspector showing data on selected samples (Figure 2), a table of genotypes from the samples shown on the top pane, the marker library (currently shown) and the size standard library. The right pane shows the traces of selected samples, in a scrollable view that can display thousands of traces. The blue channel currently shows traces for two diploid DNA markers that contain bins shown as vertical grey rectangles. The orange channel shows the molecular ladder.



Sample information:

Name: IMO03d

Comment: multi2

Plate name: jeanfinFHIJ_231007

Well: M16 Capillary: 14

Dyes: 6-FAM VIC NED PET LIZ

Source file: › › jeanfinFHI.. › jeanfinFHIJ_231007_IMO03d_M16_014.fsa

Date of import: 15/04/2023 20:20

Run information:

Sizing:

Size standard name: GS500LIZ

Used size standard: GeneScan-500 -35

Fitting method: 3rd-order polynomial

Sizing quality: 87 %

Peak height threshold: 100

Figure 2. The sample inspector of STRyper. This panel with three collapsible sections dynamically updates to display information on samples that are selected in the sample table (Figure 1). The plot at the bottom shows the relationship between the time at which DNA fragments of the molecular ladder (black crosses) were detected by the sequencer camera (the X axis) and their estimated sizes in base pairs (the Y axis). The relationship used to estimate fragment sizes is established by fitting a polynomial (here, of the third degree) to the points shown on the plot. This polynomial is represented by the orange curve.

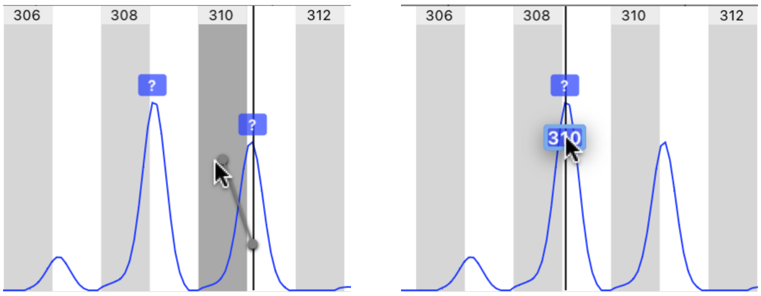


Figure 3. Genotype editing by drag and drop in STRyper. Vertical grey rectangles represent bins that define expected ranges of microsatellite alleles. Each bin has a name showing on

top. Allele names are represented by colored labels above peaks. Left screen capture: the user is dragging the mouse from a peak to a bin. This will assign the peak an allele named after the bin and replace the question mark used for alleles that are out of bins. During the operation, a handle connects the mouse location to the center of the peak at the horizontal location of the peak tip. Right screen capture: the user has decided that only the peak on the left should represent an allele and is dragging an allele label from the right-hand peak to the other. These gestures are assisted by magnetism to lock the handle or allele label to the closest suitable destination, which triggers haptic feedback on the trackpad.

Supplementary text

Peak delineation

To delineate peaks in the fluorescence data, STRyper uses a simple method that enumerates fluorescence levels from the first to the last recorded scan. A scan is a data point that is denoted by an integer index varying from 0 to the total number of data points.

The method records the lowest fluorescence level (l) and the highest level (h), and their respective scan numbers (s_l, s_h), observed up to the current scan number (s_f) whose fluorescence level is denoted as f. A peak is delineated if $h > t$, $l/h \leq r$ and $f/h \leq r$, t being the minimal fluorescence level to consider a peak and r being a parameter denoting the minimum peak elevation above the background. Horizontally, the peak starts at scan s_l, and its tip is at scan s_h. Its right boundary will correspond to the left boundary of the next peak. This method thus generates contiguous peaks.

For best results, it was found that three rounds of peak detection should be applied to the data, each round being followed by one pass of baseline fluorescence level subtraction (see next section). The first two rounds use a value of 0.7 for r, a modest peak elevation that allows the detection of relatively faint peaks. The last iteration uses a value of 0.5, which means that a peak must be at least twice higher than the background level, considering that baseline fluorescence level subtraction makes peak stand-out more.

After these three rounds, the left and right boundaries of each peak are delineated by the closest scan from each side of the peak's tip that has a fluorescence level of 0, using fluorescence levels with baseline level subtracted. This produces non-contiguous peaks.

Baseline fluorescence level subtraction

STRyper subtracts the baseline fluorescence level of a trace after peaks are delineated (see previous section) as follows. A virtual line segment is drawn from the start to the end of each peak (Figure S1). For a given scan number s_f, the height of the segment is denoted as y and is considered the "baseline fluorescence". The recorded fluorescence level for the scan is denoted as f and the fluorescence level at the peak tip is denoted as h.

For each value of s_f within the peak, a value v to subtract to the fluorescence level depends on the user preference. If they want the absolute height of peaks to be preserved, $v = y(h -$

$f)/(h - y)$. Otherwise, $v = y$. The new value for the fluorescence level is $f - v$. If the result is negative, it is set to 0. After this operation, each peak starts and ends at a fluorescence level of 0.

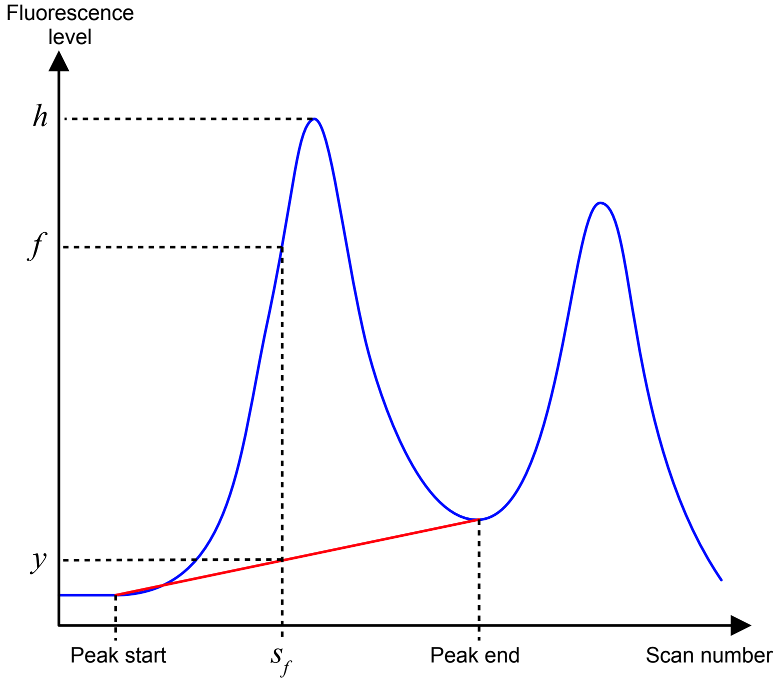


Figure S1. Subtraction of baseline fluorescence level. Symbols are defined in the supplementary text.

Determination of crosstalk

STRyper determines whether a peak in fluorescence results from interference between channels, i.e., crosstalk. This inference relies on the presence of saturation, or of higher peak of similar shapes, in other channels.

A chromatogram file lists each scan number for which the signal saturated the sequencer camera but does not specify which channel caused the saturation. STRyper determines this channel by first delineating regions composed of consecutive scan numbers where saturation occurred.

For each region, the channel that is considered to have caused saturation is the one whose fluorescence level is the highest at the first scan of the region. This criterion does not compare maximum/average fluorescence levels over the region between channels, because the peak at the channel that caused saturation if often clipped and may be shorter than peaks of other channels in the region. However, this peak has the highest fluorescence level at the point where saturation began.

A peak is considered to results from crosstalk if the following conditions are met: (i) its tip lies within a region where saturation is caused by another channel, and (ii) the fluorescence level at the peak tip is at least twice those recorded at the scan preceding the start and the scan after the end of the region. Criterion (ii) accounts for the fact that several DNA fragment may have migrated at the same speed, such that legit peaks appear at the same locations. However, the fluorescence level at a peak resulting from crosstalk should not be high before the saturation from another channel is recorded.

Alternatively, crosstalk may cause a “crater” in other channels, that is, sharp peaks at the edges of the saturated region. If a short peak lies near such edge and sharply decreases within the saturated region, the peak is considered to result from crosstalk.

If a focus peak is not considered are resulting from crosstalk based on the above checks, the program inspects other channels to find the one with highest fluorescence level at the peak tip, and for which the fluorescence level is at least 1.6 times that at the peak tip. If it finds one, it then evaluates how much peaks of both channels overlap, using two criteria. The program first scales down the taller peak such that is elevation corresponds to the shorter. It then measures the peak areas by summing fluorescence levels. The first criterion is considered passed if the area representing the intersection between peaks is at least 30% of the area representing the union of the peaks. The second criterion precisely evaluates how much the peak horizontal positions are offset. For that, the difference in fluorescence level (curve height) between channels is computed at each scan along the range encompassing both peaks. The sign of the difference is reversed if the scan is greater than the scan of a given peak’s tip. For each peak, these differences are summed across all scans of the range. The second criterion is considered passed if the absolute value of each sum is less than 30% the combined areas of the peaks. If both criteria are met, the program checks if other peaks in the channel that may have induced crosstalk also induced crosstalk in the focus channel. This inspection relies on the expected ratio of peak heights between the two channels, which should be rather constant in the case of crosstalk and in the absence of saturation. If another peak does not appear to have induced crosstalk, then the peak under consideration is not considered to result from crosstalk.

Size assignment of molecular ladder fragments

The algorithm conceived to assign sizes to molecular ladder fragments inspects peak in the appropriate channel, ignoring those resulting from crosstalk (see previous section). In the following, the “scan number” of a peak refers to the scan at its tip.

Peaks are first enumerated by decreasing scan numbers, and the average peak height is computed at each step. Any peaks whose height is at least twice the current average and whose scan number is less than 1/3 total number of scans in the trace is discarded. This eliminates high-intensity peaks of short size (in base pairs) resulting from degradation of the molecular ladder.

The algorithm then discards weak peaks amounting to “noise”, which sometimes affect the data. To do so, remaining peaks are enumerated by decreasing height. The enumeration stops when the number of enumerated peaks corresponds to the number of sizes specified in the size standard, or when a peak is at least three times shorter than the previous one. Any peak that is twice as short as the least enumerated peak, or shorter, is discarded.

To assign remaining peaks to sizes defined in the size standard, peaks are ordered by increasing scan number. The method assigns the lowest size to the first peak, and the largest size to the last peak. To understand the process, picture a straight line of equation $y = a + bx$ passing through these two peaks on a plot where the x axis represents scan numbers, and the y axis sizes in base pairs.

Peaks are then enumerated in decreasing order, starting from the second-to-last. The size of the fragment causing a peak is estimated as $a + bx$, x being the peak scan number. The size of the size standard that is the closest to the estimated size is assigned to the peak, only if the difference between both sizes is less than 15 bp in absolute value.

The next peak is evaluated in the same fashion. If it is assigned to the same size as a previous peak, both peaks are confronted to retain the one whose predicted size is the closest. The a and b parameter are updated to correspond to the line connecting the two peaks that were assigned last. Hence, the size/scan relationship dynamically changes to account for non-linearity.

At the end of the procedure, the shortest size of the size standard may be assigned to a different peak than the one of lowest scan number. This is not the case for the longest size, which remains assigned to the peak of largest scan number, although this assignment might be erroneous (this is addressed using subsequent iterations, as described below).

A quality index is computed to evaluate the assignments. This index relies on the residuals of the linear regression between scan number and size in base pairs, using ordinary least squares. For each pair of successive points (peaks), the difference between residuals is divided by the difference between scan numbers, both in absolute value. The mean of these ratios is computed. The inverse of this mean, multiplied by the percentage of sizes that were assigned to peaks, constitutes the quality index. If this index is higher than a certain value (chosen at 100), the number of assigned sizes is recorded as a reference.

Further iterations of assignments are performed by decrementing the longest assignable size (to consider the possibility that electrophoresis failed or stopped before the last fragment was detected), then by decrementing the last assignable peak. Assignments are not recorded if the number of assigned sizes is lower than the reference, and iterations stop when the number of assignable sizes/peaks is lower than the reference.

In the end, the set of assignments that yielded the best quality index is retained.

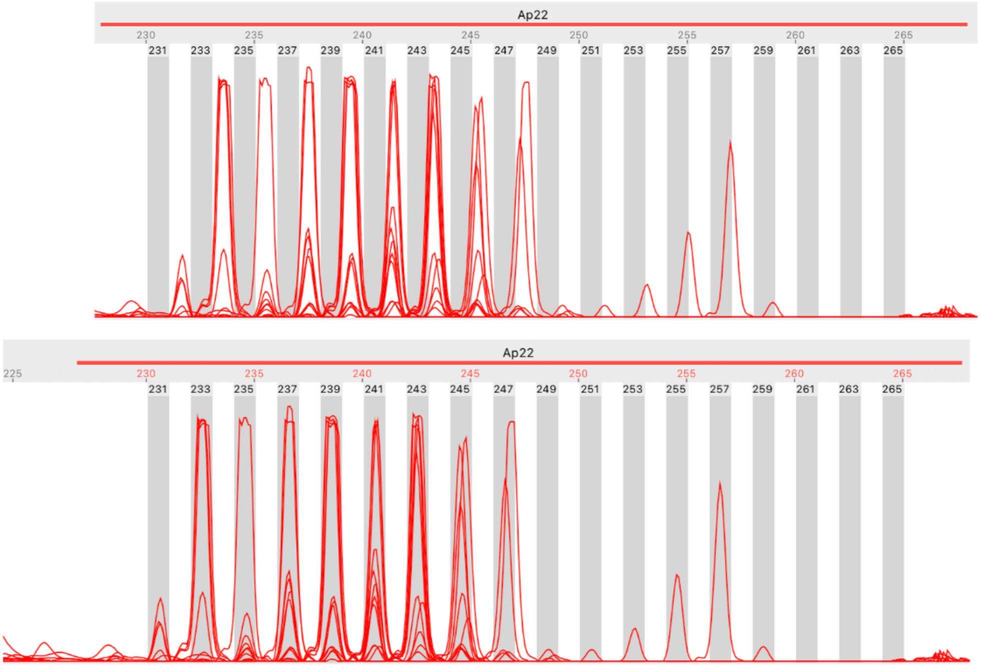


Figure S2. A case of out-of-bin alleles that is resolved. Both screen captures show the stacked traces from 12 samples of the same sequencer run. Peaks represent amplicons of

the marker, and grey rectangles the bins for the marker's alleles. Top: peaks are shifted to the right with respect to bins, and more so for shorter alleles although bins are separated by exactly two base pairs. Bottom: the user has moved and resized the bin set such that bins coincide with peaks, using a graphical editing mode that does not change bin coordinates. Indeed, the top graduations in base pairs have moved in sync with bins within the marker's range, and have turned red to denote the shift. This shift translates into a linear combination of parameters b = 1.023 and a = –6.34 (see main text). As a result, the estimated size of peaks overlapping bin 231 has changed from ~231.7 bp to ~230.7 bp.