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

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 .

Keywords: microsatellites, capillary electrophoresis, 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 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 complexify their use, and which may come with a high price.

This is the case for GeneMapper by ThermoFisher Scientific, a commercial application running one Windows and which is, 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, 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 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 from non-human species. 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.

Description of the application

Technical characteristics

STRyper runs under macOS (version 10.13 or higher). This requirement reflects the fact that its author uses the Mac operating system at work and at home, has been a population geneticist and only develops as a hobby, and does not have the competences nor the resources to develop for other platforms. STRyper is written in the objective-C language – a superset of C – and has been conceived with the Xcode integrated development environment. STRyper relies on Application Programming Interfaces (APIs) and libraries provided by macOS (https://developer.apple.com/documentation/technologies), namely “AppKit” for the user interface (UI), “Core Data” for database management, “Accelerate” for certain accelerated vector functions, and “Core Animation” to animate changes in the UI and to support hardware-accelerated compositing. 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.

Main interface and data management

Being entirely reliant on macOS APIs, STRyper behaves as expected from a modern Mac application. For instance, text fields provide the customary contextual menus for search and spell checking, they accept any Unicode character, scrolling has the rubber-band effect, trackpad gestures are recognized, table columns can be reordered by dragging, the UI adapts to the dark theme of macOS and to high pixel density displays, and so on.

The application has a 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 since 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.

As opposed to GeneMapper, 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 by dropping chromatogram files (or using a more conventional import panel) into folders, and can be moved or copied between folders at any time. A folder and all its content, including edited genotypes, applied 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.

Samples, folders, markers, genotypes, etc. are saved in an SQLite database that permit fast queries. The database is saved automatically and has an unlimited level of undo.

Sample viewing

Selecting a folder of the database shows all its samples and associated genotypes, if any. Sample can be 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).

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.

The traces of selected samples are displayed on the right pane, regardless of the application of a size standard. Traces can be zoomed in/out via trackpad gestures such as pinch and double tap, via the scroll wheel, or by dragging the mouse over the rulers to define a size range or a fluorescence level. Zooming is animated, which helps users keep track of the viewing context. On computers equipped with an Apple chipset (M1 or more recent), the drawing of traces is hardware-accelerated, so that users can zoom in and out dozens of curves on a high-resolution display at 60 frames per second with modest resource usage. Viewing options include automatic vertical scaling to the highest visible peaks, synchronizing of vertical scales and horizontal positions, showing/hiding bins and off-scale regions, and stacking curves from several samples or channels in the same view.

Fluorescence data analysis

STRyper does not smooth fluorescent curves and does not compute trend lines. Baseline fluorescence level subtraction, which helps curve and peak interpretation, entirely relies on peak delineation (supplementary text). While this approach has no effect on signals that are too faint to contain meaningful peaks, it has the benefit of offering two baseline subtraction modes to the user: one that preserves absolute peak height, and one that maintains relative peak elevation compared to the baseline (supplementary text and Figure S1).

STRyper determines whether a peak results from crosstalk by comparing the signal between channels, accounting for saturation of the sequencer camera (supplementary text). Any region of saturation is shown behind traces as a rectangle whose color reflects the channel that caused saturation. A peak that is determined to result from crosstalk will not be automatically assigned to an allele or molecular ladder fragment but can still be manually assigned.

STRyper does not put other qualifiers on artefactual peaks (stutter, adenylation, etc.), nor does it alter fluorescence data to correct for pull-up due crosstalk (Hoffman and Riley, 2021). I do not view these functions as critical to researchers in our field. Since STRyper is not meant for human identification, it does not perform mathematical analyses to detect issues that require special attention.

Sample sizing using size standards

To define 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. Detection of molecular ladder fragments and their assignment to sizes of the size standard is based on relative peak positions and accounts for non-linear relationship between fragment size and migration speed (supplementary text). Based on size assignments, STRyper fits a polynomial of the first, second, or third degree (depending on the user choice), which is used to estimate fragment sizes. Fitting is achieved via the Cholesky decomposition implemented in the Linear Algebra Package (https://netlib.org/lapack/).

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.

Markers and bins

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, name, and the size range or their alleles. The latter two attributes can be changed after a marker is created. Markers can be copied between panels.

STRyper allows to define so called “bins”, which specify expected size ranges for individual alleles, and are used for allele calling. A set of 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. Alternatively, bins can be added/removed/modified individually, also 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).

STRyper has unique features that address variations in electrophoretic conditions that may affect the estimated size of the same allele between different runs. A graphical interface (supplementary text and Figure S2) allows shifting the size of alleles of a specific marker for target samples, so that they match the sizes computed in the runs that were used to define bins. This functionality is meant to reduce the number of out-of-bin alleles.

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.

Allele calling

While users can identify and manually assign alleles to peak within marker ranges, STRyper can call alleles automatically. In doing so, the application identifies peak clusters that arise from the amplification of single microsatellite alleles, a phase that is subject to adenylation and indels causing “stuttering”. The most intense peak in a cluster is considered as that representing the allele. The method used is relatively robust to peak clipping due to saturation, in that the width of the saturated region is accounted for when peak height/area may not reflect the quantity of DNA material.

Reliable genotyping requires visual review and manual editing, and STRyper is optimized for these tasks. All genotypes from displayed samples are listed in a table that can be sorted by several columns. 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 file. Users can export all genotypes from a folder, or only selected ones, or even copy and paste data from selected genotypes from STRyper to a text editor or spreadsheet application.

Since STRyper is not suitable for human identification, it does not assume that genotyping errors affect human lives, hence that allele calls are reviewed by other 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.

In addition, the number of peaks assigned to putative alleles never exceeds the marker's ploidy. Hence, users rarely need to remove spurious alleles. However, they will have to visually check the presence of additional peaks possibly indicating sample contamination.

For the same reason, genotype quality is not computed by the application. Barring some trained artificial intelligence, which this application does not implement, I believe that an algorithm cannot yet replace the visual review of genotypes.

Conclusion

To conclude, STRyper constitutes a useful tool for microsatellite markers genotyping in the context of population genetic studies. Its strengths lie in its simple and responsive interface that allows the quick review of genotypes with a limited number of clicks. I stress that STRyper is not designed for the analysis of problematic samples that may contain very low-input material or a mixture of DNA from several individuals, or for any type of forensic analysis.

The current restriction of STRyper to macOS is partially balanced by its free nature (the cost of an entry-level Mac is lower than that of paid applications used for microsatellite genotyping), its responsiveness, and its "native" feel, which is rare among scientific applications.

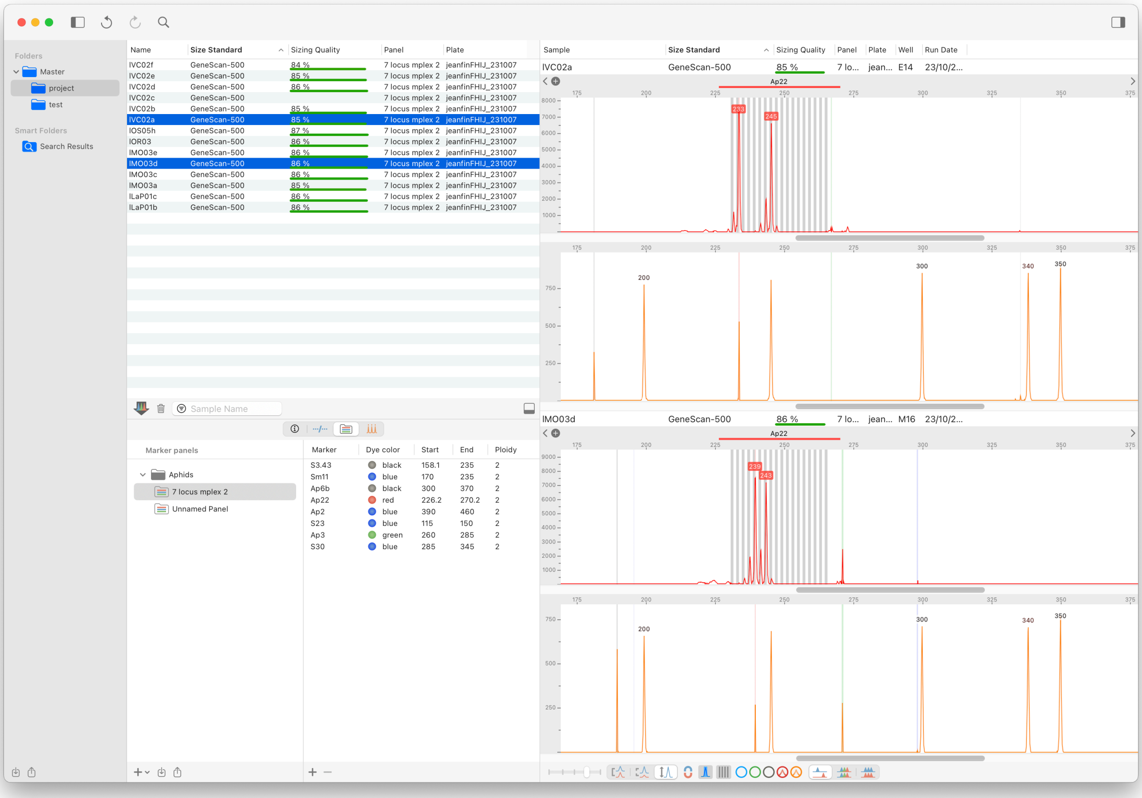
Acknowledgements

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. Part of this work was funded by intramural funds from the CNRS and the University of Poitiers, and by Agence Nationale de la Recherche Grants ANR-20-CE02-0004 (SymChroSex).

References

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

Figures



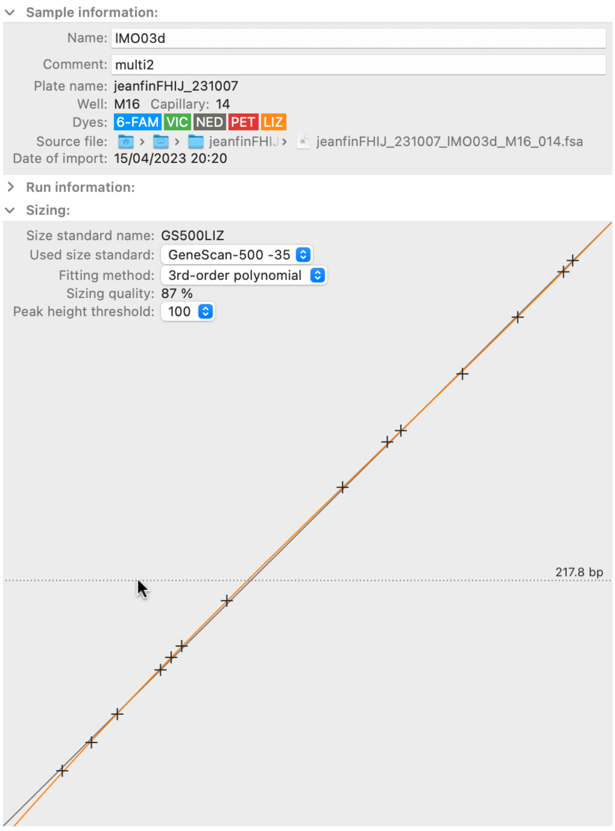
Name	Size Standard	Sizing Quality	Panel	Plate
IVC02f	GeneScan-500	84 %	7 locus mplex 2	jeanfinFHIJ_231007
IVC02e	GeneScan-500	85 %	7 locus mplex 2	jeanfinFHIJ_231007
IVC02d	GeneScan-500	86 %	7 locus mplex 2	jeanfinFHIJ_231007
IVC02c	GeneScan-500		7 locus mplex 2	jeanfinFHIJ_231007
IVC02b	GeneScan-500	85 %	7 locus mplex 2	jeanfinFHIJ_231007
IVC02a	GeneScan-500	85 %	7 locus mplex 2	jeanfinFHIJ_231007
IOS05h	GeneScan-500	87 %	7 locus mplex 2	jeanfinFHIJ_231007
IOR03	GeneScan-500	86 %	7 locus mplex 2	jeanfinFHIJ_231007
IMO03e	GeneScan-500	86 %	7 locus mplex 2	jeanfinFHIJ_231007
IMO03d	GeneScan-500	86 %	7 locus mplex 2	jeanfinFHIJ_231007
IMO03c	GeneScan-500	86 %	7 locus mplex 2	jeanfinFHIJ_231007
IMO03a	GeneScan-500	85 %	7 locus mplex 2	jeanfinFHIJ_231007
ILaP01c	GeneScan-500	86 %	7 locus mplex 2	jeanfinFHIJ_231007
ILaP01b	GeneScan-500	86 %	7 locus mplex 2	jeanfinFHIJ_231007

Marker	Dye color	Start	End	Ploidy
S3.43	● black	158.1	235	2
Sm11	● blue	170	235	2
Ap6b	● black	300	370	2
Ap22	● red	226.2	270.2	2
Ap2	● blue	390	460	2
S23	● blue	115	150	2
Ap3	● green	260	285	2
S30	● blue	285	345	2

Sample	Size Standard	Sizing Quality	Panel	Plate	Well	Run Date
IVC02a	GeneScan-500	85 %	7 lo...	jean...	E14	23/10/2...
IMO03d	GeneScan-500	86 %	7 lo...	jean...	M16	23/10/2...

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 red channel currently shows traces for a diploid DNA marker ("Ap22"). This marker specifies 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: › › jeanfinFHIJ › 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.

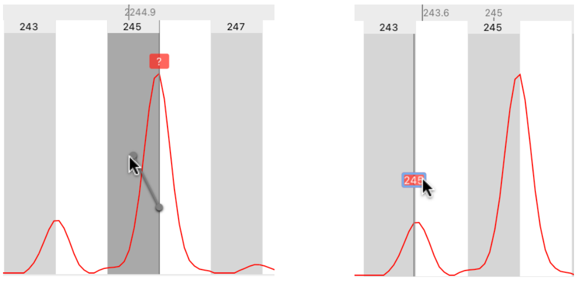


Figure 3. Genotype editing by drag and drop in STRyper. Vertical grey rectangles represent bins that define expected ranges of microsatellite alleles. Each bin has a name showing on top. Allele names are represented by colored labels above peaks. Left screen capture: the user is dragging the mouse from 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 the peak on the left, rather than the peak on the right, should represent an allele, and is dragging an allele label from the right-hand peak to the other. This will assign an allele to the destination peak. The allele will be named after the bin encompassing the peak tip, if any.

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). It works 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, $v = y$. If v is negative, it is set to 0. The new value for the fluorescence level is $f - v$.

After this operation, each peak starts and ends at a fluorescence level of 0.

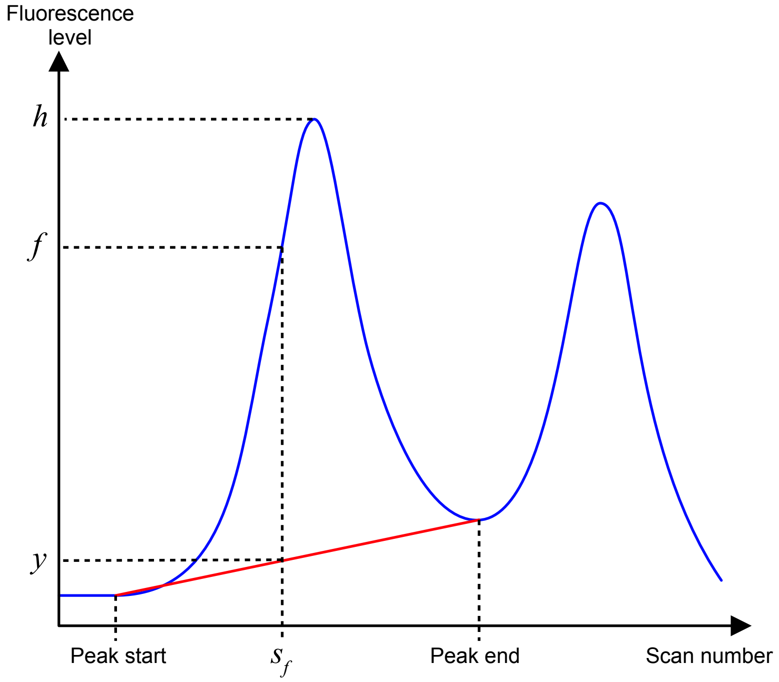


Figure S1. Subtraction of baseline fluorescence level. Symbols are defined in the supplementary text.

Determination of crosstalk

STRyper determines whether a peak (identified via the method described above) 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 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 start and end of the region. The second criterion 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, a peak resulting from crosstalk should not start before the saturation from another channel is recorded.

If no saturation was detected at the peak, the application inspects other channels to find the one with highest fluorescence level at the peak tip, and for which the fluorescence level is at least four times that at the peak tip. If it finds one, it then evaluates whether this other channel shows a peak of similar shape at the location of the focus peak. It does so by comparing fluorescence level at each scan with the peak range, after standardizing by the average ratio of fluorescence levels over scans within this range. If peak shapes appear similar, the application then 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 is 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

To assign sizes to molecular ladder fragment, STRyper 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 one 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.

STRyper then tries to discard 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.

Addressing variations in electrophoretic conditions causing out-of-bin alleles

Amplicons and molecular ladder fragments often react differently to variations in electrophoretic conditions. Hence, the estimated size of the same allele may slightly vary between runs. Such variations can shift the location of alleles with respect to bins, defeating the purpose of bins.

Moving bins to address this issue would require managing multiple bin sets per marker. Also, it is the estimate sizes of the alleles, not the position of bins, that should be shifted. To this aim, STRyper allows users to offset the size of peaks found in a marker range, for specific samples. Peak sizes are computed with the formula $y = a + bx$, where x is the "original" size of a DNA fragment computed thanks to the molecular ladder, 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 offset, a = 0 and b = 1. This approach assumes that the effect of varying electrophoresis conditions can be approximated by this linear combination.

The linearity allows users to find appropriate values for a and b parameters by modifying the horizontal location and width of a rectangle representing the marker range, such that bin positions neatly match those of peaks arising from reference alleles (Figure S2). The shift in marker position defines b, while the change in width affects both a and b.

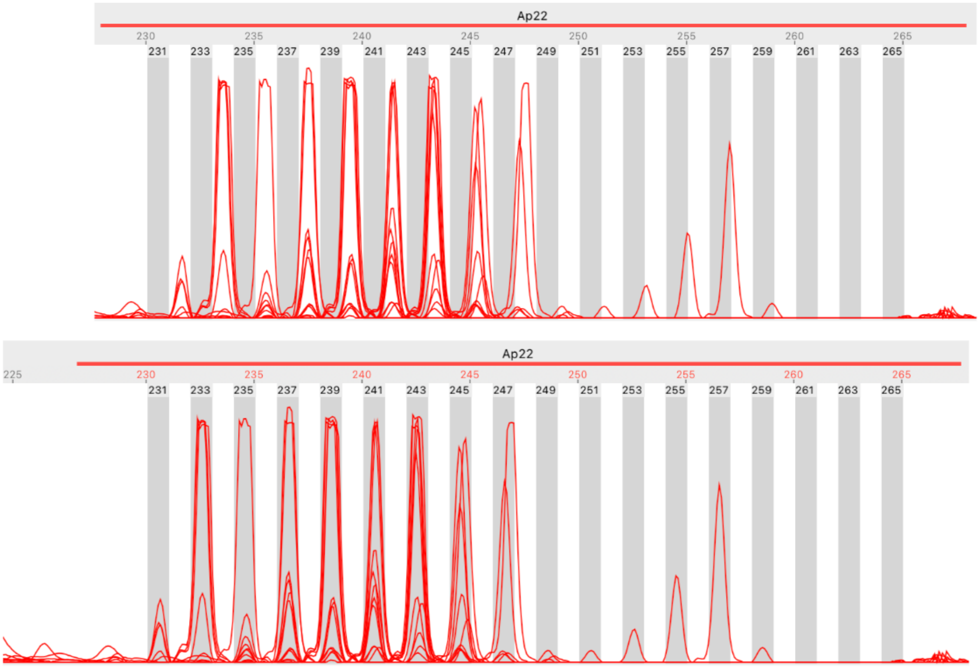


Figure S2. A case of out-of-bin alleles that is resolved. Both screen captures show the stacked traces from 12 samples of the same sequencer run. Peaks represent microsatellite alleles of a dinucleotide marker, and grey rectangles the user-defined bins for the marker. Top: alleles appear to be shifted to the right with respect to bins, and more so for shorter

alleles (although all adjacent bins are separated by exactly two base pairs). Bottom: the user has shifted the peak locations so that they coincide with the bins. This shift corresponds to linear combination of parameters b = 1.023 and a = –6.34 (see supplementary text). Hence, the estimated size of peaks overlapping bin 231 has shifted from ~231.7 bp to ~230.7 bp, which should match their estimated sizes in other runs where these peaks coincide with the bin. The fact that sizes are offset within the marker range is denoted by the red color of size graduations, and by the large gap between graduation 225 and 230.