Deconvolution of nucleic-acid length distributions: a gel electrophoresis analysis tool and applications

Riccardo Ziraldo[1], Massa J. Shoura[2,3,*], Andrew Z. Fire[2,3] and Stephen D. Levene [1,4,5,*]

[1]Department of Bioengineering, University of Texas at Dallas, 800 W Campbell Road, Richardson, TX 75080, USA,
[2]Department of Pathology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA,
[3]Department of Genetics, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305, USA,
[4]Department of Biological Sciences, University of Texas at Dallas, 800 W Campbell Road, Richardson, TX 75080, USA and [5]Department of Physics, University of Texas at Dallas, 800 W Campbell Road, Richardson, TX 75080, USA

ABSTRACT

Next-generation DNA-sequencing (NGS) technologies, which are designed to streamline the acquisition of massive amounts of sequencing data, are nonetheless dependent on various preparative steps to generate DNA fragments of required concentration, purity and average size (molecular weight). Current automated electrophoresis systems for DNA- and RNA-sample quality control, such as Agilent's Bioanalyzer® and TapeStation® products, are costly to acquire and use; they also provide limited information for samples having broad size distributions. Here, we describe a software tool that helps determine the size distribution of DNA fragments in an NGS library, or other DNA sample, based on gel-electrophoretic line profiles. The software, developed as an ImageJ plug-in, allows for straightforward processing of gel images, including lane selection and fitting of univariate functions to intensity distributions. The user selects the option of fitting either discrete profiles in cases where discrete gel bands are visible or continuous profiles, having multiple bands buried under a single broad peak. The method requires only modest imaging capabilities and is a cost-effective, rigorous alternative characterization method to augment existing techniques for library quality control.

INTRODUCTION

Next-generation DNA sequencing (NGS) has rapidly become an indispensible tool in virtually every life-science discipline. NGS workflows entail various DNA-sample manipulations including PCR and enzymatic reactions to prepare DNA fragments of specific concentration, purity and size in ways that are compatible with a particular sequencing platform ([1]). The quality of the NGS library has substantial influence on the success of a sequencing run, affecting both sequence validity and the number of reads ([2]). Many current quality-control (QC) protocols require costly instruments and consumables to assess the library average size, specifically capillary-gel electrophoresis tools such as Agilent's Bioanalyzer® and TapeStation® products. With the exponential growth of genome-wide analyses and large datasets that hinge on NGS, it is crucial to optimize the accuracy and cost-effectiveness of each workflow step.

Agarose-gel electrophoresis is arguably the most widely used method for separating biopolymer components in a sample based on physical, chemical and topological properties ([3–9]). As an important example, a commonly used workflow relies on gel electrophoresis to separate DNA fragments generated by Illumina's Nextera® Tn5 transposition-based tagging/fragmentation protocol. The Nextera® kit produces sequencing libraries through enzymatic shearing of input DNA. Gel electrophoresis uses the molecular-weight dependence of linear DNA's mobility in polymer gels to separate the resulting fragments along the direction of migration ([7,8]). In applications, such as typical analyses of restriction digests, the products consist of a small number of individual fragments, producing a characteristic pattern of discrete bands. In contrast, a variety of random-shearing-based protocols (including Nextera® and other transposition-based protocols) can be used to produce broad, ideally uniformly distributed, fragment libraries ([10]). Thus, agarose-gel separations of tagmentation products, because of the large number of possible fragment sizes, generally yield a quasi-continuous distribution of DNA molecules along the path of migration.

QC on the fragment library can greatly improve the sequencing output and reduce both PCR bias and systematic errors ([11,12]). Commercial and open-source tools, such as GelAnalyzer, PyElph, GelJ or ImageQuant, offer a number of features for gel image analysis. They allow the user to

---

easily set up analyses and detect lanes and discrete bands in 'banded' profiles; however, these tools lack the capability to analyze quasi-continuous fragment-size distributions. Our image-analysis tool is designed to efficiently analyze both discrete and continuous gel-electrophoresis profiles. The technique fits a univariate composite Gaussian function to approximate the profile. Amounts of each species in the sample are then estimated using numerical integration. The goal is to obtain information about the size distribution of fragmented DNA products, from which the average molecular size and bulk concentration of DNA ends can be determined. In the particular case of amplicon-based sequencing such as Illumina's platforms, accurate knowledge of such fragment-size distributions is essential to optimize cluster density in sequencing protocols.

MATERIALS AND METHODS

Library preparation and gel-electrophoresis-based size selection

Human, *Caenorhabditis elegans* or bacteriophage-lambda genomic DNA (λ DNA) were incubated with Nextera XT Tagmentase® (Illumina), an engineered Tn5 transposase, for at least 15 min at 37°C. DNA fragments underwent PCR-based amplification using Illumina's Nextera® primers and the resulting libraries were subjected to electrophoresis for the indicated time in 1% agarose gels (Lonza) at 3.0 V cm^{-1} in TBE buffer (50 mM Tris-borate, 1 mM Na$_2$EDTA; pH 8.4). Gels and electrophoresis buffer contained 0.5 μg ml^{-1} ethidium bromide unless otherwise noted. NGS libraries consisted of tagmented DNA reisolated from excised regions of gel lanes using a Zymoclean Gel DNA Recovery® kit (Zymo Laboratories) [13].

A reference mixture of λ-DNA fragments was prepared by digesting λ DNA to completion with 2 units of *Aci*I (New England Biolabs) per μg of λ DNA (New England Biolabs) for 2 h at 37°C. A virtual digest performed using SnapGene® showed that the λ-*Aci*I digest consists of 517 DNA fragments and 206 discrete fragment sizes distributed between 2 and 1086 bp. In addition to the λ-*Aci*I reference mixture, DNA ladders were used as markers: HiLo ladder (Bionexus, Inc) and an NEB 100-base-pair (bp) ladder (New England Biolabs).

Image acquisition

Several cameras were used to record digital images of the agarose gels. Both high bit-depth (16-bit) grayscale images and RGB (24-bit) color images were acquired. Table 1 lists the different cameras used for the comparison and the specifications of the acquired images.

Sequencing protocol and bioinformatic analysis

Gel-purified DNA libraries were sequenced using the MiSeq® platform (Illumina) according to the vendor's recommended protocol. Paired-end reads were obtained by sequencing DNA libraries using 150v3 MiSeq® kits (Illumina) for 76 paired cycles. Insert sizes were determined bioinformatically from the distance between paired reads

Table 1. Cameras used for image acquisition

Camera 1 (C1)	Flea2 1.4 MP Mono FireWire 1394b 1392x1032, 12 bit, uncompressed
Camera 2 (C2)	SONY, DSC-WX350, 125ms exp 1440x1080, sRGB 24 bit, compressed
Camera 3 (C3)	One Plus, Android phone, 3480,2610x4640i,RGB 24 bit, compressed
Camera 4 (C4)	Apple, iPhone SE, 67 ms exp 3024x4032, sRGB 24 bit, compressed

(R1 and R2) after alignment. The size histogram of sequenced reads was obtained as described by Shoura *et al.* [13].

Plug-in operation

The workflow for analysis of the gel images is implemented using an ImageJ (https://imagej.net/ImageJ2) plug-in (Gel Lanes Fit). The ImageJ platform was chosen because of its wide distribution as an open-source platform for application and development of image analysis tools. The JFreeChart (http://www.jfree.org/jfreechart) library was used for output plots and the Apache Commons Math (http://commons.apache.org/proper/commons-math/) library was used to implement the fitting algorithm.

The Gel Lanes Fit plug-in includes a number of features that streamline the processing of gel images. Each gel lane is associated with a rectangular 'region of interest' (ROI). The dimensions of each ROI can be selected interactively through the plug-in's interface and, as the ROI is resized, the line-profile plot of the corresponding gel lane is dynamically updated. The interface provides access to parameters for non-linear least-squares fitting of the line profile discussed below.

Function describing the line profile

The image-intensity line profile generated for the corresponding ROI is modeled using a composite univariate function of a coordinate y corresponding to the apparent electrophoretic migration distance of DNA fragments in the gel. This function is the sum of a background signal and a set of discrete Gaussian peaks associated with each DNA species in the mixture. In cases, where there are relatively few species for which the centers of mass of each component are well separated relative to peak width (defined in terms of full peak width at half-maximum (FWHM), for example), the gel pattern is made up of discrete bands and the line profile consists of discrete peaks. Continuous distributions, in which the separation of adjacent peak maxima is comparable to their respective FWHM values, are expected to have the appearance of a 'streak' or 'smear.'

Some background contribution to the signal is expected due to scattering from the gel and fluorescence from unbound ethidium dye. The background function maximum is assumed to be less than the smallest minimum corresponding to peaks in the discrete case. A constraint on the maximum degree of the background polynomial and its maximum slope (first derivative) can be set by the user using the plug-in's interface.

Equation (1) shows the target function $F(y)$. $B_q(y)$ is a polynomial of degree q. Both q and a limit on the first derivative of B_q can be selected by the user. G_i are Gaussian functions with given amplitude, a_i, mean, m_i and standard deviation σ_i.

$$F(y) = B_q(y) + \sum_i G_i(y, a_i, m_i, \sigma_i)$$

$$B_q(y) = b_0 + b_1 y + b_2 y^2 + \ldots + b_q y^q$$

$$G_i(y) = a_i \exp\left\{-\frac{(y - m_i)^2}{2\sigma_i^2}\right\} \tag{1}$$

The function is fitted using the Levenberg–Marquardt least-squares algorithm (14). For discrete patterns, each peak is assumed to be Gaussian with the mean at the location of a local maximum and having variance proportional to the band spread. Thus, initial guesses for m_i and σ_i are made using local maxima and peak widths. The peak tolerance parameter sets a threshold for the peak height in the initial guess and thereby filters any peaks contributed by noise. This parameter is defined as a percentage of the full line-profile intensity range.

Discrete and known continuous patterns

Well-resolved image-intensity line profiles show peaks that correspond to various species present in the sample. However, as discussed above, well-separated peaks cannot always be obtained. The function given in Equation (1) is assumed to be valid for both discrete and continuous profiles, but stronger constraints are needed to fit profiles composed of under-resolved peaks. If a small peak overlaps a larger peak, it cannot be detected automatically. However, an initial guess can be included using the custom peak feature, which specifies the potential location and width of the undetected peak.

When the line profile approaches a continuum, there are no discernible peaks and a different fitting strategy has to be used. Each fragment in a known distribution is used to define constraints on the location and the width of the Gaussian peaks in Equation (1). As a model for a quasi-continuous distribution of DNA-fragment sizes, we used the mixture of products generated by complete digestion of λ DNA by the restriction enzyme *Aci*I. This product mixture contains 517 fragments of known size (see 'Library preparation and gel-electrophoresis-based size selection' section) and constitutes a basis set for fitting unknown, quasi-continuous fragment distributions. The vast majority of fragments in this mixture lie in a size range comparable to that selected during library preparation.

An initial guess for the position (mean) of a fragment's peak is determined from the locations of known species in a discrete, well-resolved ladder of standard fragments of known molecular weight. The position of the unresolved fragments along the continuum is interpolated from the discrete-ladder pattern using the spline function in Equation (2), with spline nodes located along positions that correspond to the maxima of ladder bands. We assume a piecewise linear relationship between the logarithm of the molecular weight, $\log_{10}(M)$, and displacement in the direction of

migration, y (9). The molecular weight, M, of a particular double-stranded fragment is calculated from the molecule size in base pairs, l, and the linear relation $M(l) = 607.4l + 157.9$. Here, 607.4 is the molecular weight of the average nucleotide pair and 157.9 is the combined molecular weight of two 3'- or 5'-monophosphate groups, which are the termini generally left behind by the restriction enzyme (15). The displacement in the direction of migration, y, observed in the gel is estimated using Equation (2), based on a piecewise-linear relationship between $\log_{10}(M)$ and y (16). Thus, the linear segments are delimited by the components of the standard ladder of molecular weights, $M_{\text{ladder}} = \{M_1, \ldots, M_k\}$, and the respective observed displacements, $Y_{\text{ladder}} = \{Y_1, \ldots, Y_k\}$. Each segment of y is defined over the intervals $M_{i-1} < M \le M_i$ for the ladder species $i = 2, \ldots, k$.

$$\begin{cases} y(\log_{10}(M)) = h_i \log_{10}(M) + c_i \\ h_i = \dfrac{Y_{i-1} - Y_i}{\log_{10}(M_{i-1}) - \log_{10}(M_i)} \\ c_i = \log_{10}(M_{i-1}) - Y_{i-1} \end{cases} \tag{2}$$

In the case of a 'discrete' pattern the standard deviation of the Gaussian peak associated with the band, σ_i, is estimated using the FWHM of the peak intensity profile. For 'continuous' patterns, a value of σ_i for each fragment in the distribution is determined by interpolation of a piecewise least-squares linear fit to the σ_i values corresponding to bands in a discrete-ladder standard. Additional constraints are applied to the final profile fit by restricting deviations allowed relative to the original σ_i estimates. The observed band intensity, or peak amplitude, a_i, is assumed to be proportional to the product of mass fraction and DNA molecular weight. The initial guess for the peak amplitude is thus estimated as $a_i = f_i w_i d$, where f_i is the fragment's relative abundance, w is the molecular weight normalized by the maximum value and d is the full-scale range of the line profile.

Therefore, for a generic continuous profile with 'unknown' fragments, the initial guess is represented by a set of peaks associated with a 'known' distribution of n_f fragments and assumed similarities between known and unknown distributions of fragments. The final fit is constrained so that the location and FWHM of each Gaussian component are similar to that of the initial guess with the estimate for the initial amplitude of each component based on the abundance and size of fragments in the initial distribution. Adherence of the amplitude for each peak to that expected from the known distribution is constrained by an area-drift parameter described below. The fitted peaks' mean (y displacement, proportional to the molecular weight, $w_{i,\text{fit}}$) and area, $p_{i,\text{fit}}$ (proportional to the fragment's molecular weight and abundance) are used to calculate an average molecular weight for the distribution of fragments, using Equation (3).

$$\bar{w} = \frac{\sum_i w_{i,\text{fit}} p_{i,\text{fit}}}{n_f \sum_i p_{i,\text{fit}}} \tag{3}$$

The bulk DNA concentration, C as determined by Qbit fluorescence assays for example, can then be converted to number-averaged fragment molarity $\bar{M}$ through the relation $\bar{M} = C/\bar{w}$.

Area-drift constraint for continuous line profiles

Initial peak areas based on the parameters corresponding to the known distribution are taken as elements of an array, P in Equation (4), calculated for peak i as $a_i\sigma_i\sqrt{2\pi}$, from the peak's amplitude, a_i, and standard deviation, σ_i. At each fitting iteration, the set of parameters is updated based on the algorithm's gradient requirements. The new list of parameters is first checked to exclude non-physical parameter values, for a_i, m_i and σ_i. After the exclusion of non-physical parameter values, the algorithm recalculates an array of areas, Q, in Equation (5). Particular constraints are enforced so that the peak-area values from the initial guess retain their proportionality, based on the known DNA mass distribution in the λ-*Aci*I digest.

$$P = \{p_0, \ldots, p_n\} = \{p_i\} \tag{4}$$

$$Q = \{p_0 + \delta_0, \ldots, p_n + \delta_n\} = \{q_i\} \tag{5}$$

$$\Xi = \left\{\frac{p_i}{q_i}\right\} \approx \Gamma = \{\gamma, \ldots, \gamma\} \tag{6}$$

The array Ξ contains element-by-element ratios of P and Q arrays. The applied area constraint ensures that Q stays proportional to P. Thus, the elements in the array of ratios, Ξ, are bounded, and would ideally have the same values as those in the array Γ, in Equation (6). This constraint is enforced using the condition in Equation (7). If the standard deviation of the elements in Ξ, $s(\{\xi_i\})$, exceeds a value ϵ, chosen by the user, then the array of areas Q is adjusted to compensate, using Q'.

$$\text{If } s\left(\{\xi_i\}\right) > \epsilon, \quad \text{then } Q' = \left\{\left[\bar{\bar{\xi}}_i a_{i,0} + \mathcal{U}(-\epsilon, \epsilon)\right]\sigma_i\sqrt{2\pi}\right\} \tag{7}$$

Since the FWHM is subjected to a stricter constraint (see previous subsection), the area constraint is mostly enforced using the peaks' amplitudes, $N' = \{a_1, \ldots, a_k\}$, which are obtained by scaling the array of initial amplitude guesses, $N = \{a_{1,0}, \ldots, a_{k,0}\}$, by $\bar{\xi}_i$, the sum of mean values of the components of $\{p_i/q_i\}$ plus a uniformly distributed random increment, $\mathcal{U}$.

RESULTS

Discrete profiles

The major goal of plug-in development was to provide a tool that was equally capable of quantifying discrete, banded, as well as continuous, gel profiles. An example of discrete-band quantification is shown in Figure 1, in which the plug-in is used to analyze a standard ladder composed of known molecular-weight species. The location of clearly defined bands is automatically detected using the location of local maxima in the profile. The peak position, maximum intensity above background, integrated intensity and the FWHM are reported by the plug-in for each peak in the gel lane. As an example of a discrete profile fit, Figure 1 shows the output plot with the fitted profile's distribution. The plug-in also reports the 'root-mean square' (RMS) difference between the fit and the line profile, a measure of the
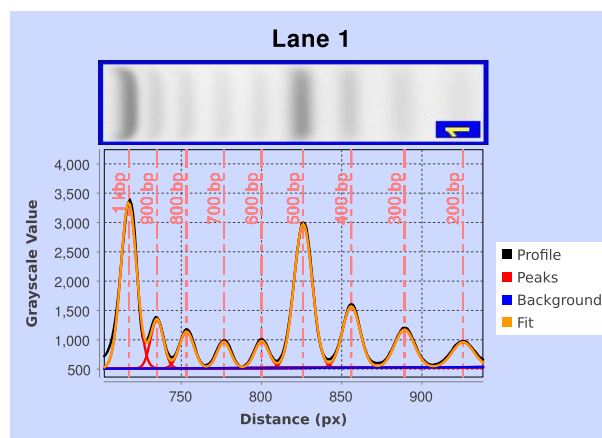

Figure 1. Fit to a discrete profile obtained for gel lane shown in the top panel (also present in lane 1 of Figure 2A). The complete non-linear least-squares fit (orange) and its component background (blue) and Gaussian peaks (red) are shown. The vertical dashed lines (light red) indicate the locations of peak maxima, corresponding to known molecular-weight values for the ladder species.

closeness of the fitted function to the original profile (not shown). This discrete-profile fitting strategy applies in general, but it is used specifically by the plug-in to analyze continuous line profiles.

Continuous profiles

The plug-in can quantify profiles with suboptimally resolved bands, such as DNA-fragmentation products obtained with the Illumina Nextera® tagmentation kit and quantitate the average size of fragments contained in the mixture. Lanes labeled 2–5 in the gel in Figure 2A and 1–4 in Figure 2C are all examples of continuous gel profiles, in which single bands are not visible.

The fragments in an unknown distribution are approximated by a known distribution of fragments having similar electrophoretic mobilities. Figure 2 shows fits to continuous profiles for an unknown fragment-size distribution (a tagmentation reaction on bacteriophage-λ DNA, lanes 3–5 in 2A and lanes 2–4 in 2C) using the λ-*Aci*I digest, a known fragment-size distribution, as a model (lane 2 in Figure 2A and lane 1 in Figure 2C). As expected, the algorithm performs best in the range of fragment sizes that are well covered by the known distribution, but produces lower-quality fits in size ranges where there is limited coverage by the λ-*Aci*I digest (i.e. > 425 bp).

*Aci*I digestion produces a fragment distribution that lies approximately in the range that is appropriate for Illumina tagmentase library preparation. Other quasi-continuous distributions can be generated using a different enzyme or a combination of enzymes. For example, digestion with *Fin*I will produce a quasi-continuous fragment distribution spanning a different molecular-weight range (∼1000 bp) than that produced by digestion with *Aci*I. The plug-in is designed to accommodate different molecular-weight standards, which are specified in an ASCII input file.
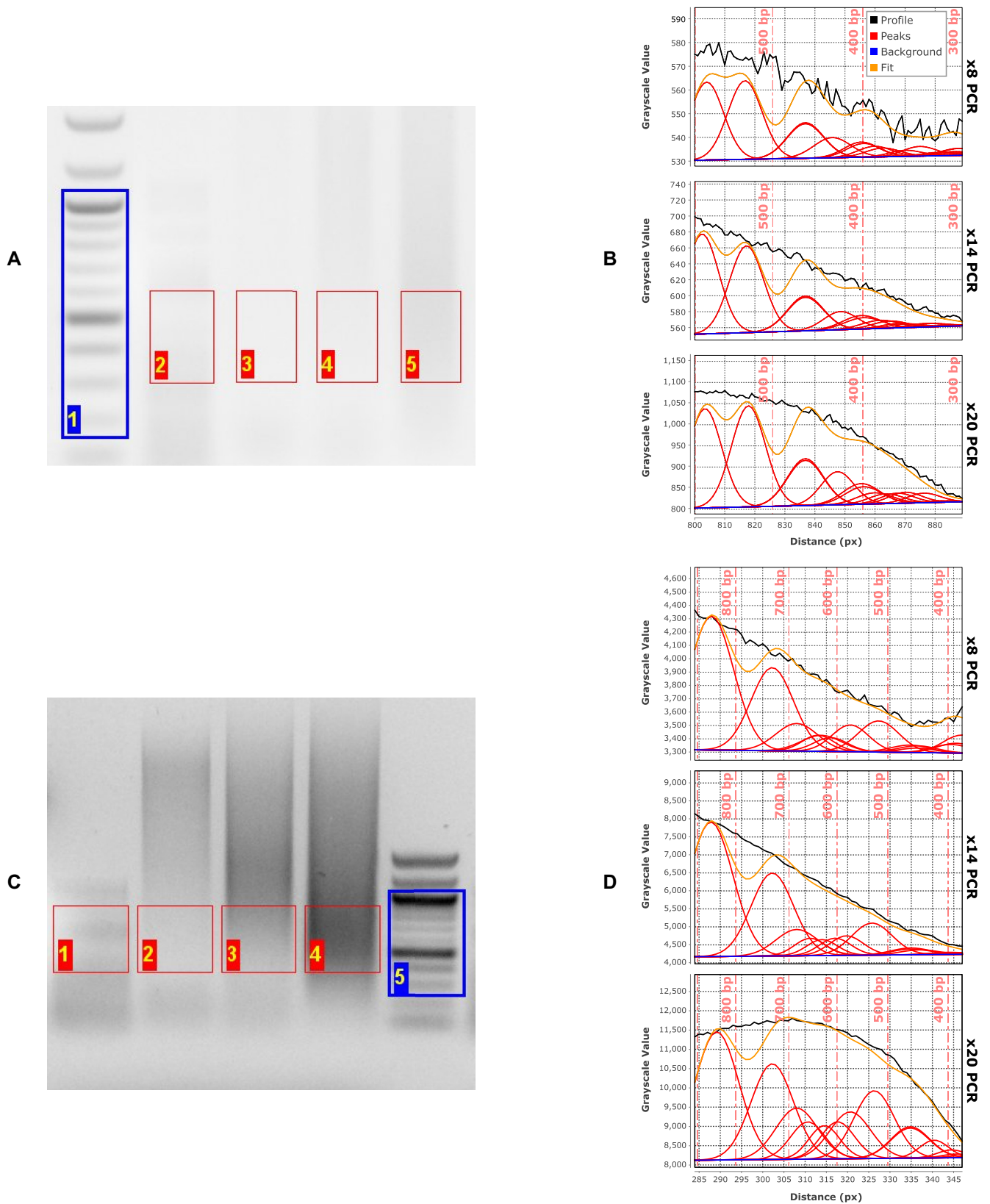
Figure 2. Agarose-gel analysis of tagmentation products of phage-λ DNA analyzed on (**A**) high-resolution and (**C**) low-resolution (i.e. mini) agarose gels. Size-distribution fits for the high-resolution gel (**B**) and mini gel (**D**) of the same tagmented λ-DNA sample subjected to increasing numbers of PCR-amplification cycles (lanes 3–5 in (A), lanes 2–4 in (C)): 8 cycles (top plots), 14 cycles (middle plots) and 20 cycles (bottom plots) in both (B) and (D). Vertical dashed lines (light red in (B), (D)) give the positions of maxima in the discrete molecular-weight ladder (blue ROI in (A), (C)).

Sensitivity to gel resolution

We carried out a pilot assessment of utility of the plug-in by analyzing apparent fragment-size distributions for samples subject to distinct PCR amplification protocols and electrophoretic-separation conditions. In particular, we were interested in assessing whether profiling of gel-electrophoresis patterns using our tool could anticipate differences in size distributions observed in high-throughput Illumina DNA-sequencing experiments for different samples and conditions. Working with a set of samples amplified through different numbers of PCR cycles, we compared plug-in and sequencing output from (i) long-format, high-resolution agarose gels (Figure 2A and B) and (ii) short-format, lower-resolution (minigel) electrophoretic separations (Figure 2C and D). Identical samples were run at the same field strength ($3.0\,\mathrm{V\,cm^{-1}}$) on both gels, but for different durations: 16 h in the case of the long-format gel and 2 h in the case of the minigel. Standards consisted of the Hi-Lo ladder, 100-bp ladder and and the λ-*Aci*I digest; experimental lanes contained products of a tagmentation reaction with phage λ genomic DNA that were subjected to differing numbers of PCR-amplification cycles.

The effect of suboptimal fragment separation on minigels is apparent when we compare electrophoretic fragment-size distributions obtained after 8, 14 and 20 PCR cycles (Figure 2C and D). In particular, between 14 and 20 cycles there is a substantial increase in the subpopulation of fragments >500 bp in size. The effect of size-distribution distortion is especially prominent in the minigel results and, because of the relatively small region being excised, more difficult to mitigate through selective excision of an appropriate region of the gel. These results both confirm the utility of the plug-in analysis package and highlight the extent to which gel resolution has a significant effect on NGS-library characterization.

Further evaluation using the plug-in examined differences in size distributions obtained both from analysis of the lane profiles and from MiSeq post-sequencing fragment-size distributions for the completed library (see 'Sequencing protocol and bioinformatic analysis' section). Although input- and output-size profiles for the MiSeq will show some distortion due to the known preference for smaller fragments during bridge amplification, measurement of input material with the plug-in allows prediction of conditions most likely to alter that pattern of eventual fragment densities (11,17–20). Figure 3 quantitatively shows the effect of increasing numbers of amplification cycles on the fragment-size distributions obtained from MiSeq output. As with the data from the plug-in analysis, there is a clear increase in low-molecular-weight bias between 14 and 20 amplification cycles. Although additional aspects of library-preparation and sequencing workflow may contribute to biases in the apparent molecular-weight distributions, it is clear that libraries prepared using limited PCR amplification and careful size selection on high-resolution agarose gels are likely to provide higher fidelity sequencing output.

Sensitivity to image-acquisition hardware

The images for the gel in Figure 2A were acquired using two different cameras (C1 and C2 in Table 1), whereas images
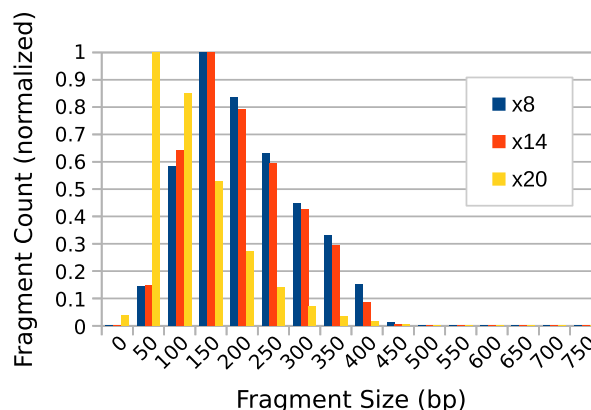

Figure 3. Fragment-library size distributions obtained from bioinformatic analysis of MiSeq sequencing output. A tagmented sample of phage λ genomic DNA was subjected to 8, 14 and 20 cycles of PCR amplification, respectively. The fragment counts in each bin are normalized with respect to the maximum value in each distribution.
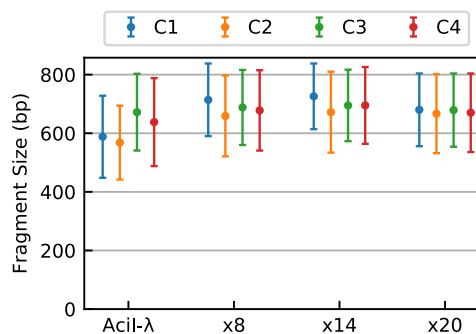

Figure 4. Dependence of plug-in output on camera hardware. Estimates of average fragment size obtained for images of the low-resolution gel (Figure 2C collected using four different camera systems, C1–C4 (Table 1). The plug-in measurements of average fragment size were compared for the known λ-*Aci*I digest and the same λ-genome tagmented samples analyzed in Figure 2. Error bars indicate the standard deviation (i.e. $\sqrt{\mathrm{variance}}$) of the fitted distribution. Numerical data for this experiment are also provided in Supplementary Table S2.

for the gel in Figure 2C were acquired using all four camera systems. Figure 4 and Supplementary Figure S1 compare the values of average fragment size estimated using the plug-in for both the mini- and high-resolution gel data (numerical data for Figure 4 can be found in Supplementary Tables S2 and S1). Each dataset reports the apparent average fragment size for the λ-*Aci*I digest and for different numbers of PCR-amplification cycles of a tagmentation reaction analyzed on a single gel, but imaged using different hardware. The error bars show the standard deviation above and below the mean, $\pm 1\sigma_d$, of the fitted distribution of fragments, as a measure of the distribution's spread. In general, the differences are modest with the largest discrepancies occurring between high-resolution scientific cameras having appropriate bandpass optical filters and simpler camera systems (such as cell phone cameras), which generate uncalibrated RGB images. We speculate that the absence of a bandpass filter alters the spectral content of the captured signal, which in the absence of calibration, may report inaccurate signal intensities upon conversion from RGB to grayscale.
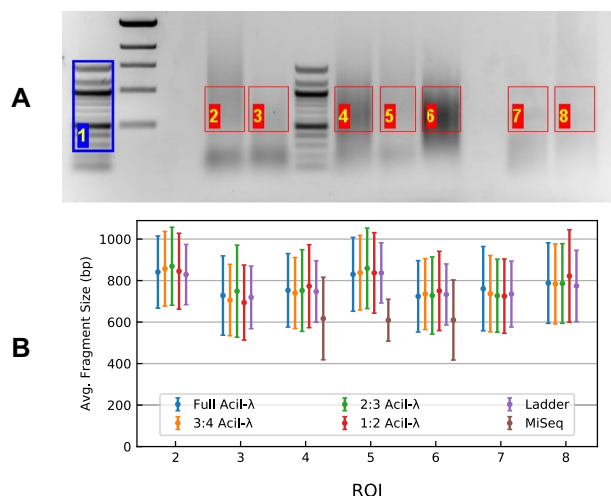
Figure 5. Data used to examine the effect of fragment-density test. (**A**) The gel was analyzed considering the portions delimited by the ROIs, using lane 1 (blue) as the reference ladder. ROIs 2–6 are tagmentation products of human genomic DNA; 7 and 8 are λ-*Aci*I digest samples (**B**). Summary plot of the fit for the gel in panel (**A**). Fragment-size mean and standard-deviation values for model distributions containing varying numbers of reference fragments (peaks). Symbols are colored according to the proportion of the original λ-*Aci*I basis fragments retained in the reference distribution; thus, 1:2 means that every other peak was eliminated from the fit, 2:3 every third peak etc. Data are also reported in Supplementary Table S4.

To further test the robustness of our approach, we processed a gel image of tagmentation products taken at different exposure levels. The ROIs in Figure 6A and B show the excised regions from which libraries were prepared for this analysis. As shown in Figure 6C and D, the apparent average fragment sizes are independent of the imaging exposure time. Detailed results are also reported in Supplementary Table S3.

Effect of fragment-size densities in reference distributions

The plug-in provides better profile fits when the reference distribution predominantly consists of DNA fragments in the size range of interest. This is because we generally wish to restrict band widths (FWHM) for both discrete and continuous profiles to values similar to those of co-localized peaks present in the discrete reference ladder. To provide improved profile fits in regions where the reference distribution has fewer peaks, the constraint that the FWHM of an unknown fragment's peak would be similar to that of a comparably sized fragment in the reference ladder (see 'Discrete and known continuous patterns' section for details) has to be relaxed, with the assumption that a single Gaussian peak in the basis set covers a range of fragment sizes that are present in the analyzed distribution.

To examine the effect of limiting reference-peak coverage, we created hypothetical reference distributions by regularly removing some basis-set peaks present in the λ-*Aci*I digest. Figure 5A shows the gel used for this example and Figure 5B (Supplementary Table S4) reports the fitted distribution's mean fragment size and standard deviation. Supplementary Table S5 reports the absolute RMS as a measure of the closeness of fit for the respective cases.

Comparison with TapeStation® output

Tagmented libraries prepared from *C. elegans* genomic DNA were analyzed by agarose-gel electrophoresis as shown in Figure 6 and also using Agilent's TapeStation® (TS) system. We excised two different sections of the gel from each lane to obtain distinct fragment distributions having different average fragment sizes, as indicated by the respective ROIs. The libraries were reisolated from the gel as described (see 'Library preparation and gel-electrophoresis-based size selection section') and reanalyzed on an Agilent 2200 TapeStation. Here we compare the fragment-size distributions obtained by using the plug-in with those provided by TS output. Generally, fragments of lower molecular size (300–500 bp) are relevant for NGS; however, we investigated distributions of larger fragment sizes to determine the effectiveness of our plug-in for other applications. Comparison of plug-in and TS results are shown in Figure 6E and F (Supplementary Table S6).

DISCUSSION

We have developed an ImageJ plug-in for the quantitation of gel-electrophoresis images that can analyze both discrete and continuous gel patterns. This work was motivated by the lack of rigorous and inexpensive DNA size-analysis tools for NGS library preparation. Knowing the distribution of fragment sizes and, hence, the average concentration of DNA termini in a sequencing library is an important quality-control (QC) step, particularly in amplicon-based sequencing.

In the case of discrete patterns, the plug-in can directly quantify relative amounts of DNA based on the intensity and width of the gel bands. In the case of continuous patterns, the plug-in estimates the fragment-size distribution based on a model of superimposed Gaussian peaks derived from a suitable known standard. For NGS library preparation, the user specifies, through a rectangular ROI, the region of the gel lane and hence the subset of fragmented products that make up the library. The plug-in provides an estimate of the actual DNA-mass distribution, detailed information that typically goes beyond that provided by specialized QC tools.

We carried out a case study for the determination of library-size distributions using the Nextera® tagmentation protocol. The distribution of fragment sizes in the input library is important in the performance of amplicon-based NGS. Fragments that are too small will be favored during the bridge amplification step, hindering the sequencing of longer fragments; fragments that are too large pose problems with dye localization. Using the plug-in, we were able to quantify a fragment-size bias caused by excessive PCR amplification. Awareness of this amplification bias can improve QC in the preparation of NGS libraries.

The plug-in performs well with both high-resolution and short-format (i.e. mini-) gels (see 'Image acquisition' section); however, analysis is most efficient when bands in the reference ladder of DNA standards are well separated and peaks in the reference lane can be detected automatically. Apparent DNA-size distributions were broadly insensitive to image characteristics and quality in our study, which
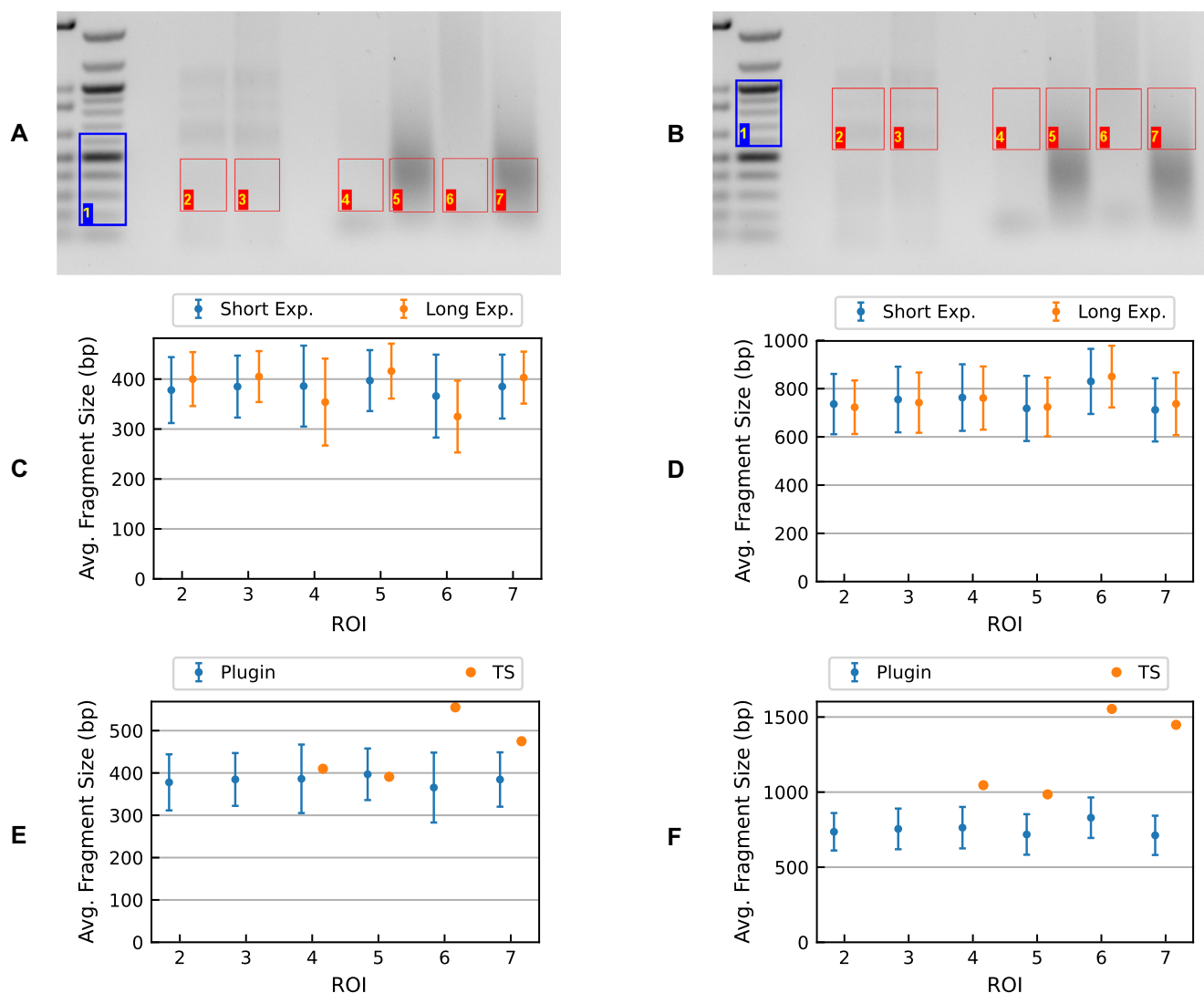
Figure 6. Tagmented libraries prepared from *C. elegans* genomic DNA. Gel images show the sections corresponding to panel (**A**) low-molecular-weight and panel (**B**) high-molecular-weight fractions. In both cases, lanes 2 and 3 are λ-*Aci*I digest samples; lanes 4–7 *C. elegans* genomic-DNA tagmentation reactions amplified by PCR under different conditions. Summary plot of average fragment size based on short- or long-exposure images, panel (**C**) low-molecular-weight and panel (**D**) high-molecular-weight fractions (Supplementary Table S3 in table form). Comparison of average fragment sizes obtained using the plug-in to those generated by TapeStation output for the same DNA fractions: panel (**E**) low molecular weight and panel (**F**) high molecular weight (Supplementary Table S6 in table form).

used four different camera systems ranging from scientific CCDs to cell phone cameras. This shows that the approach implemented by the plug-in can robustly analyze raw image data from a wide variety of sources. From our experience, the precision and accuracy afforded by the approach is more than sufficient for guiding sample preparation in a wide range of sequencing modalities and applications. Ongoing tool development and applications will investigate effects of digital image compression, conversion from RGB to grayscale or changes in pixel-value precision when using sensors with different dynamic ranges. For NGS applications, the magnitude of the variations observed in predicted fragment-size averages obtained using different gel cameras are not expected to be propagated in such a way to affect cluster density predictions.

As a final validation step, we compared the plug-in's estimates of average fragment size with that generated by Agilent's TapeStation® (TS) instrument. The TS system has a list price of ∼USD 50,000 and significant consumables costs. Moreover, the TS software reports the most-probable fragment size (*i.e.*, the mode) along with graphical output of the apparent size distribution. In most cases, the plug-in output is in good agreement with TS output. However, for some of the samples TS reports unreasonable mode values for the size distribution. This happens with very flat profiles where a peak is not detectable. Our approach, while not as 'hands-off' as the TS system, can handle lower signal-to-noise line profiles where at least some slope is detectable in the fragment-size distribution.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank early users whose feedback helped refine the plug-in.

FUNDING

REFERENCES

1. Head,S., Komori,K., LaMere,S., Whisenant,T., Van Nieuwerburgh,F., Salomon,D.R. and Ordoukhanian,P. (2014) Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, **56**, 61–77.
2. Marine,R., Polson,S.W., Ravel,J., Hatfull,G., Russell,D., Sullivan,M., Syed,F., Dumas,M. and Wommack,K.E. (2011) Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl. Environ. Microbiol.*, **77**, 8071–8079.
3. Stellwagen,N.C. (2009) Electrophoresis of DNA in agarose gels, polyacrylamide gels and in free solution. *Electrophoresis*, **30**, 188–195.
4. Shimooka,Y., Nishikawa,J.I. and Ohyama,T. (2013) Most methylation-susceptible DNA sequences in human embryonic stem cells undergo a change in conformation or flexibility upon methylation. *Biochemistry*, **52**, 1344–1353.
5. Levene,S.D. (2009) Analysis of DNA Topoisomers, Knots, and Catenanes by Agarose Gel Electrophoresis. In: Clarke,DJ (ed). *DNA Topoisomerases: Methods and Protocols*. Humana Press, Totowa, pp. 11–25.
6. Vetcher,A.A., McEwen,A.E., Abujarour,R., Hanke,A. and Levene,S.D. (2010) Gel mobilities of linking-number topoisomers and their dependence on DNA helical repeat and elasticity. *Biophys. Chem.*, **148**, 104–111.
7. Southern,E. (1979) Measurement of DNA length by gel electrophoresis. *Anal. Biochem.*, **100**, 319–323.
8. Elder,J., Amos,A., Southern,E. and Shippey,G. (1983) Measurement of DNA length by gel electrophoresis: I. Improved accuracy of mobility measurements using a digital microdensitometer and computer processing. *Anal. Biochem.*, **128**, 223–226.
9. Elder,J. and Southern,E. (1983) Measurement of DNA length by gel electrophoresis II: Comparison of methods for relating mobility to fragment length. *Anal. Biochem.*, **128**, 227–231.
10. Boone,M., DeKoker,A. and Callewaert,N. (2018) Capturing the ome: the expanding molecular toolbox for RNA and DNA library construction. *Nucleic Acids Res.*, **46**, 2701–2721.
11. Aird,D., Ross,M.G., Chen,W.S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.
12. Taub,M.A., Corrada Bravo,H. and Irizarry,R.A. (2010) Overcoming bias and systematic errors in next generation sequencing data. *Genome Med.*, **2**, 87.
13. Shoura,M.J., Gabdank,I., Hansen,L., Merker,J., Gotlib,J., Levene,S.D. and Fire,A.Z. (2017) Intricate and Cell Type-Specific Populations of Endogenous Circular DNA (eccDNA) in Caenorhabditis elegans and Homo sapiens. *G3*, **7**, 3295–3303.
14. Marquardt,D.W. (1963) An algorithm for least-squares estimation of nonlinear parameters. *J Soc. Indust. Appl. Math.*, **11**, 431–441.
15. Akhras,M.S., Pettersson,E., Diamond,L., Unemo,M., Okamoto,J., Davis,R.W. and Pourmand,N. (2013) The Sequencing Bead Array (SBA), a next-generation digital suspension array. *PLoS ONE*, **8**, 1–13.
16. Bearden,J.C. (1979) Electrophoretic mobility of high-molecular-weight, double-stranded DNA on agarose gels. *Gene*, **6**, 221–234.
17. Gohl,D.M., Magli,A., Garbe,J., Becker,A., Johnson,D., Anderson,S., Auch,B., Billstein,B., Froehling,E., McDevitt,S.L. *et al.* (2019) Measuring sequencer size bias using REcount: a novel method for highly accurate Illumina sequencing-based quantification. *Genome Biol.*, **20**, 85.
18. Schirmer,M., Ijaz,U.Z., D'Amore,R., Hall,N., Sloan,W.T. and Quince,C. (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.*, **43**, e37.
19. Bustin,S.A. and Nolan,T. (2004) Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J. Biomol. Tech.*, **15**, 155–166.
20. Devonshire,A.S., Whale,A.S., Gutteridge,A., Jones,G., Cowen,S., Foy,C.A. and Huggett,J.F. (2014) Towards standardisation of cell-free DNA measurement in plasma: Controls for extraction efficiency, fragment size bias and quantification. *Anal. Bioanal. Chem.*, **406**, 6499–6512.