Software for Automated Analysis of DNA Fingerprinting Gels

Daniel R. Fuhrmann,[1] Martin I. Krzywinski,[2] Readman Chiu,[2] Parvaneh Saeedi,[2] Jacqueline E. Schein,[2] Ian E. Bosdet,[2] Asif Chinwalla,[3] LaDeana W. Hillier,[3] Robert H. Waterston,[3] John D. McPherson,[3] Steven J.M. Jones,[2] and Marco A. Marra[2,4]

[1]Department of Electrical Engineering, Washington University, St. Louis, Missouri 63130, USA; [2]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada; [3]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA

Here we describe software tools for the automated detection of DNA restriction fragments resolved on agarose fingerprinting gels. We present a mathematical model for the location and shape of the restriction fragments as a function of fragment size, with model parameters determined empirically from "marker" lanes containing molecular size standards. Automated identification of restriction fragments involves several steps, including: image preprocessing, to put the data in a form consistent with a linear model; marker lane analysis, for determination of the model parameters; and data lane analysis, a procedure for detecting restriction fragment multiplets while simultaneously determining the amplitude curve that describes restriction fragment amplitude as a function of mobility. In validation experiments conducted on fingerprinted and sequenced Bacterial Artificial Chromosome (BAC) clones, sensitivity and specificity of restriction fragment identification exceeded 96% on restriction fragments ranging in size from 600 base pairs (bp) to 30,000 bp. The integrated suite of software tools, written in MATLAB and collectively called BandLeader, is in use at the BC Cancer Agency Genome Sciences Centre (GSC) and the Washington University Genome Sequencing Center, and has been provided to the Wellcome Trust Sanger Institute and the Whitehead Institute. Employed in a production mode at the GSC, BandLeader has been used to perform automated restriction fragment identification for more than 850,000 BAC clones for mouse, rat, bovine, and poplar fingerprint mapping projects.

Maps constructed from fingerprinted large-insert bacterial clones (Marra et al. 1997) have been constructed to support whole-genome and localized DNA sequencing activities, as well as gene cloning studies, in plants (Marra et al. 1999; Mozo et al. 1999; Tao et al. 2001; Chen et al. 2002), animals (McPherson et al. 2001; Gregory et al. 2002), the nematodes *Caenorhabditis elegans* (Coulson et al. 1995; The C. elegans Genome Sequencing Consortium 1998) and *Caenorhabditis briggsae* (J. Schein and M. Marra, unpubl.), insects (Hoskins et al. 2000), fungi (Olson et al. 1986; Schein et al. 2002), and bacteria (Wechter et al. 2002; J. Schein, I. Bosdet, and M. Marra, unpubl.). Starting with a sufficiently redundant large-insert library of genomic DNA, the fingerprinting process (Schein et al. 2003) involves purification of DNA from clones, treatment of the DNA with restriction enzymes to produce restriction fragments, resolution of the restriction fragments on agarose gels, identification of the restriction fragments to produce a fingerprint, comparison of the fingerprints to each other to generate contigs (clusters of overlapping clones representing the genomic regions from which the clones were derived), and finally verification of clone ordering within every contig. Even using clones containing very large inserts (i.e., bacterial artificial chromosome [BAC] clones; Shizuya et al. 1992), hundreds of thousands of fingerprints may be required to accurately represent a large (i.e., mammalian-sized) genome in large contigs. The scale of such efforts and the need to produce data in a rapid, efficient, and cost-effective fashion have provided impetus for the automation of various steps in the fingerprinting procedure. Here we describe automation of the step involving identification of restriction fragments ("bandcalling"), which is performed on digital images of agarose fingerprinting gels.

Sulston et al. (1988, 1989) were among the first to develop methods for automating bandcalling. They developed software for lane tracking and band detection that was the precursor to the IMAGE package, which is available from the Sanger Institute (http://www.sanger.ac/Software/Image). Although IMAGE is user-friendly and has impressive functionality in terms of image manipulation and display, in our experience the bandcalls it produces require significant manual verification. Presumably one reason for this is that IMAGE was designed for analysis of fingerprints generated on acrylamide gels from end-labeled DNA fragments (Coulson et al. 1986; Gregory et al. 1997). These end-labeled fragments represent typically only a portion of the DNA contained within the clone. This is in contrast to the agarose method employed currently by ourselves (Marra et al. 1997; Schein et al. 2003) and others, in which all of the restriction fragments derived from a clone are visualized by postelectrophoretic staining of agarose gels with SYBR green (Molecular Probes). Indeed, this methodological difference has made possible our bandcalling

[4]**Corresponding author.**
E-MAIL mmarra@bcgsc.ca; FAX (604) 877-6085.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.904303.

13:940–953 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

approach, which aims to identify all of the restriction fragments and, in the case of comigrating restriction fragments, their copy number (or "multiplicity").

Our software tools, collectively called BandLeader, consist of a set of MATLAB routines that are capable of automatically identifying and locating marker lanes and data lanes and the restriction fragments contained therein. Gel images, collected during the fingerprinting procedure, are subjected first to "lane-tracking", a semi-automated procedure that identifies the location of the lanes on the digital gel image. This step is performed using the excellent image manipulation tools that are part of the IMAGE package. IMAGE-extracted gel lanes are then passed automatically to BandLeader for band-calling. Currently the BandLeader software is completely dependent upon the gel and data format presently in use at the British Columbia Cancer Agency (BCCA) Genome Sciences Centre. Detailed protocols for the production of fingerprints suitable for analysis by BandLeader have been described (Schein et al. 2003). Briefly, each gel consists of 25 marker lanes and 96 "data lanes" containing large-insert clone fingerprints. DNA quantities and electrophoresis conditions are strictly controlled to ensure gel-to-gel uniformity of the data. Each marker lane contains 37 fragments, with data lanes (BACs digested with *Hin*dIII or another suitable restriction enzyme) containing 50 or more bands. Hence, each gel contains more than 6000 fragments that must be identified. The BandLeader suite accomplishes this task for each gel in approximately 10 min on a computer with 1 gigabyte of RAM and an Intel-based processor running at 1 GHz. Using heuristic data checking, the BandLeader routines flag potential artifact lanes and exclude them from the data set. These can be viewed subsequently if desired.

BandLeader has been under development at both Washington University (St. Louis) and the Genome Sciences Centre (Vancouver) for more than four years, and during this time several versions were produced and used. The earliest versions, developed during the height of activity on the Human Genome Project, were based on full two-dimensional image processing techniques and were too slow to be of practical use. The basic methodology as presented here was in place by June 2000 (Version 2.0). At that time, the software could be used as a first step, but manual postprocessing was required to correct for certain artifacts, most notably overcalls, which resulted from a mismatch between a narrowly defined data model and the actual image data. Most of the development effort of the past two years was carried out while the mouse mapping effort at the Genome Sciences Centre was underway, and concentrated on making the software more reliable and the results more robust with respect to variations in the data away from the nominal model. A full set of exception-handling routines, to flag data that the software tools were unable to interpret, were also implemented. The most recent version (Version 2.3.3) is described here.

RESULTS

An electronic image of a typical agarose fingerprinting gel is shown in Figure 1A. Figure 1B shows a marker lane, with each enumerated DNA fragment annotated with the corresponding size of the fragment. Gel (TIFF) images are collected on a Molecular Dynamics Fluorimager. Although the Fluorimager is capable of different settings, BandLeader has been tuned to use 200-micron-square pixels. Each gel image is 1000 × 1200 pixels. The gel image is partitioned into 121 single-lane images, each 1000 pixels long × 9 pixels wide, as a result of the lane tracking process in IMAGE, but is otherwise unprocessed prior to our analysis. The lanes on the gel image in Figure 1A are typical of the input provided to BandLeader, and illustrate the problem BandLeader has been designed to address; namely, the automated identification of all DNA fragments in both the marker lanes and the data lanes. Below we describe the considerations and approaches that we devised to automatically identify fragments on gels of this type, and quantify BandLeader's performance on fingerprints corresponding to a test set of fully sequenced and finished BAC clones.

Forward Synthesis Model

The methodology we adopt for the analysis of fingerprinting gels is consistent in many respects with the general model-based image analysis paradigm of O'Sullivan, Blahut, and Snyder (1998). It is based on a forward synthesis model that captures as much relevant information about the desired quantities (the molecular fragment sizes) as possible, while maintaining a level of simplicity that allows for computational efficiency. While the model is based in part on the underlying physics of the fingerprinting process, the quantitative aspects are derived or refined from the data themselves.

In brief, our model for the production of electrophoretic gel data is as follows. Under the influence of an electric field, molecular fragments of a certain size f_k migrate to a particular location on a fingerprinting gel and form a diffuse *band*. The relationship between the distance traveled, or *mobility*, and the fragment size is given by a curve which is known approximately but which must be refined empirically. A typical size/mobility curve is shown in Figure 2.

In our model, the shape of the band is a rectangle subjected to Gaussian diffusion, and the diffusion parameters are size-dependent. The brightness, or *amplitude* of each band is also size-dependent. Qualitatively speaking, bands due to smaller fragments travel further, are dimmer, and are more diffuse, as is evident in Figure 1. In an idealized linear model, one data lane contains the superposition of bands consistent with this model at locations determined by the fragment sizes. This is described by the model equation

$$I(x,y) = \sum_{k}^{K} A_k \, B(x, y - m_k, f_k) \tag{1}$$

where x and y are the horizontal and vertical spatial coordinates, respectively, I is the acquired image intensity, A_k is the amplitude of the k^{th} band, m_k is the mobility of the k^{th} band, and B is a band shape function for the kth band. This band shape function is separable, that is,

$$B(x,y,f_k) = B_h(x,f_k) \, B_v(y,f_k) \tag{2}$$

where B_h and B_v are the horizontal and vertical factors.

Our model also accounts for the various deleterious effects that cause the image data to depart from the idealized linear model. These include: (1) a pointwise nonlinearity of fluorescent signal intensity, deliberately introduced by the Molecular Dynamics Fluorimager to compress the visual dynamic range, (2) an additive background function which appears data-dependent in an unknown way but which is smoothly-varying, (3) impulsive noise due to dust specks and other gel impurities, and (4) saturation in regions of high signal intensity. Background instrumentation noise is *not* included in the model, as the signal-to-noise ratio (SNR) is high

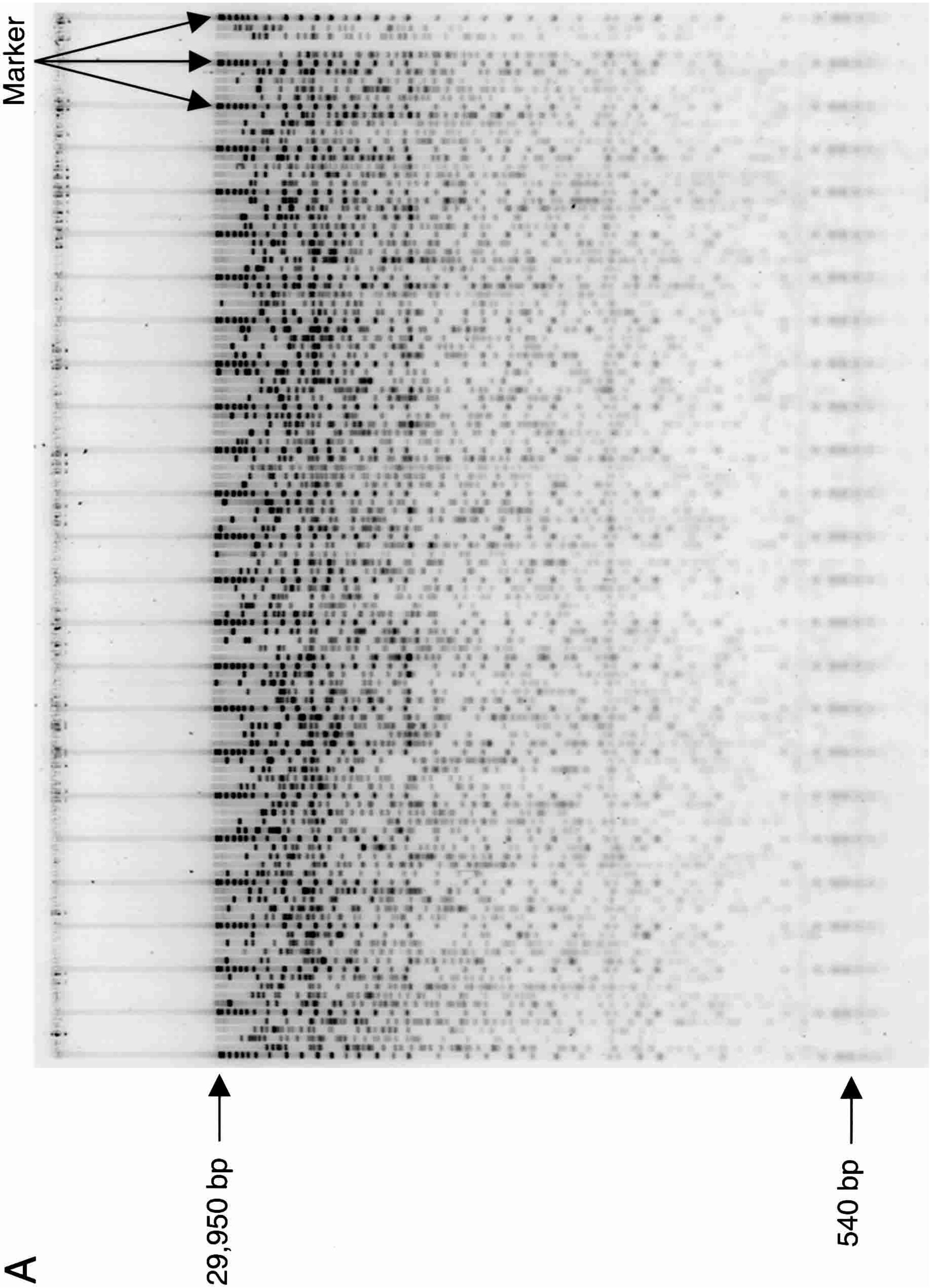

Figure 1 (Continued on facing page)

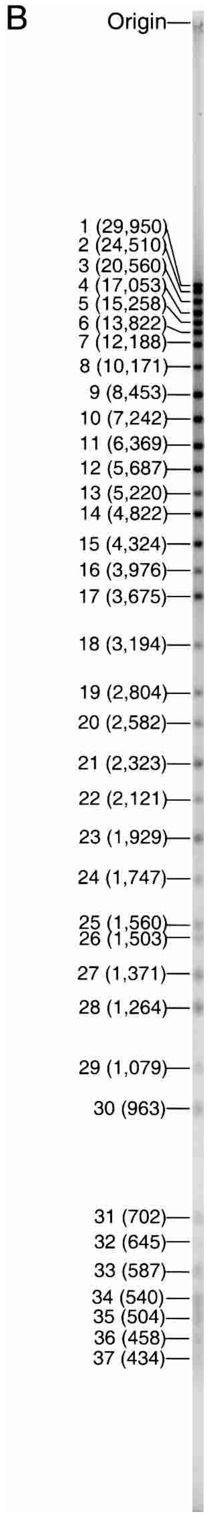

Figure 1 A typical agarose fingerprinting gel. (*A*) Shown is gel number rn0211c, produced at the BCCA GSC in July 2001. Resolvable fragments range in size from about 30,000 bp to 600 bp. The gel contains 96 data lanes and 25 marker lanes, with marker lanes occurring every fifth lane. The data lanes are never more than one lane removed from a known DNA standard. (*B*) Close-up of a marker lane, indicating the size in bp of the 37 marker fragments.

and we see little to be gained with a Poisson or Gaussian "statistical inverse problem" approach.

Based on the model described, we have developed a processing strategy which comprises the following elements: (1) an image preprocessing step for compensating for the deleterious effects in the image and putting it in a form consistent with a one-dimensional linear model, (2) marker lane analysis, for determining the quantitative aspects of the model, particularly the size/mobility curve and band shape parameters, and (3) data lane analysis, for the detection and sizing of bands in data lanes.

Image Preprocessing

The purpose of preprocessing is to mitigate the deleterious effects that are present in the data and are not directly related to the linear model and unknown fragments; in effect, it is a "data cleaning" step. The preprocessing also includes the integration of the image in the across-lane direction to obtain a one-dimensional trace $T(y)$ which is used in the subsequent model-fitting using analysis-by-synthesis.

Correction for Optical Nonlinearity

The first step in the preprocessing is to correct for the optical nonlinearity deliberately introduced to compress the visual dynamic range. This is accomplished by the trivial operation of squaring every pixel value. As a result of this step, the data are represented using real or floating-point values rather than 16-bit integers.

Background Subtraction

Background subtraction has as its goal the removal of a slowly-varying positive function, which may depend on the distribution of fragment sizes in some unknown way, but which is uninformative and does not enter into the linear model. Various algorithms for doing this can be found in several application areas in image processing. We adopt a procedure known as the MinMax Filter (J. Mullikin, pers. comm.), modified slightly for application to 2-D image data. In the modified MinMax Filter, a background image which is slowly varying in the vertical direction and constant in the horizontal direction is first determined and then subtracted from the original image. A second 1-D MinMax filter is also applied to the trace $T(y)$ which results from the preprocessing, as described below.

Impulsive Noise Filter

Dust specks and other particulate contaminating material that are found frequently in the gels appear as isolated bright pixels, or spots about 2–3 pixels in size. They are usually easy to recognize because, according to our linear model, the image is smooth in the across-lane direction. Our approach to dealing with impulsive noise is to identify outlier pixels, then delete and spline-fit through them.

Outlier pixels are identified on a row-by-row basis. In each row, every pixel is tested to see whether or not it exceeds twice the median value for that row, and if so it is considered an outlier. Pixels so identified are deleted and replaced by values obtained by cubic spline interpolation from the remaining pixels in the row. In the event that the outlier pixel lies on the edge of the lane, it can happen that the result of cubic spline extrapolation can be negative. Any negative values obtained are set to 0.

Symmetry and Monotonicity Constraints

The band shape model we use is a rectangle diffused via convolution with a Gaussian kernel. In the horizontal, or across-lane, direction, this shape is symmetric about the center pixel and monotonically decreasing from the center pixel to the edges. To force our image data to conform to this model, several simple steps are taken. The "center of mass" is computed for each row, and these values are fit to a straight line running in the vertical direction. Each row is then interpolated onto a grid, which moves this line to the center of the lane. Following this, the lane is averaged with its mirror image about the center to enforce the symmetry constraint, and finally, to enforce monotonicity, each pixel is thresholded so that its value does not exceed the value of its neighbor toward the center.

Conversion to One–Dimensional Trace

Because of the separability of our band shape model, it is possible to convert the 2-D image to a 1-D trace prior to analysis via model fitting. All of the desired fragment-size information remains present in the 1-D data; this greatly simplifies the analysis in terms of computational and memory requirements. At every vertical position y, we eliminate the across-row factor by *matched filtering*:

$$T(y) = \frac{\int I\,(x,y)\; B_h(x,y)dx}{\int B_h^2(x,y)dx} \qquad (3)$$

In this expression, $B_h(x,y)$ is the across-row band shape identified with vertical position y. If the 2-D image which results from the first preprocessing steps conforms to the noiseless linear model, then

$$T(y) = \sum_{k=1}^{K} A_k\, B_v(y{-}m_k,f_k) \qquad (4)$$

This matched filtering procedure would be optimal under an additive Gaussian noise model, even though we have not postulated such a model. There are other ways to obtain $T(y)$ in the no-noise case, such as integration across the lane or even simply sampling the center pixel. However, matched filtering does provide some robustness to noise and is more amenable to modifications to handle model inaccuracies, such as saturation and adjacent-lane effects.

All of the computations implied in Equation 3 are carried out on a discrete grid, with the integrations being replaced by a summation over nine pixels, corresponding to the width of the lane extracted by IMAGE from the fingerprinting gel. All of the shape functions $B_h(x,y)$, or *templates* are stored in a look-up table which is computed prior to analysis.

Gain Correction

In the linear model there is a large dynamic range in the amplitudes of the bands, due to the compounding of the effects of fragment size and band diffusion. To counteract this effect, both for purposes of visualization and also to ensure that all sections of the lane are treated as equally important in the model fitting, we have introduced a normalization which applies increasing gain to pixels at increasing mobility. The gain applied to the row at vertical position y is the inverse of the integrated intensity of an unnormalized band at position y. The nominal result of this normalization is that all bands have equal area.

Examples of the results of the preprocessing steps described above are illustrated graphically in Figure 3. In this figure, the various stages of the preprocessing are shown as a sequence of panels. The top panel shows the raw image data displayed in false color using the MATLAB "jet" colormap. The second panel shows this image after background subtraction and correction for optical nonlinearity. In the third panel, we see the image after the output of the dust-speck filter. This image also has a line running down the middle, indicating the estimated center of the lane. The fourth panel shows the result of additional processing to center the bands in the lane and enforce constraints of monotonicity away from the center. The result of gain correction to normalize the bands is shown in the fifth panel. Also in the fifth panel the image data have been shifted to the left to account for a bulk mobility shift (relative to a fixed standard) that is determined during marker lane analysis. Finally, in the sixth panel is shown the 1-D trace obtained after the application of the across-lane matched filter.

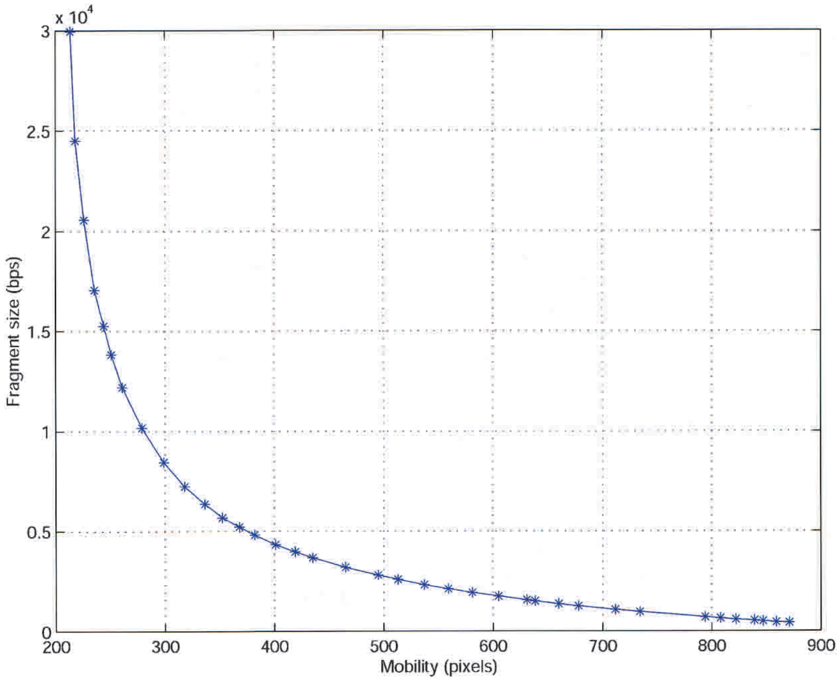

Figure 2 Curve illustrating the typical relationship between restriction fragment size and restriction fragment mobility. Size vs. mobility data are for the 37 marker fragments, taken from one lane of one of our standard gels. Individual data points are indicated by asterisks.

Marker Lane Analysis

The purpose of the marker lane analysis is to determine the exact mobilities of 37 different fragments (Fig. 1B) that have exactly known sizes. The nominal mobilities of these fragments are known fairly accurately, for a given experimental protocol,

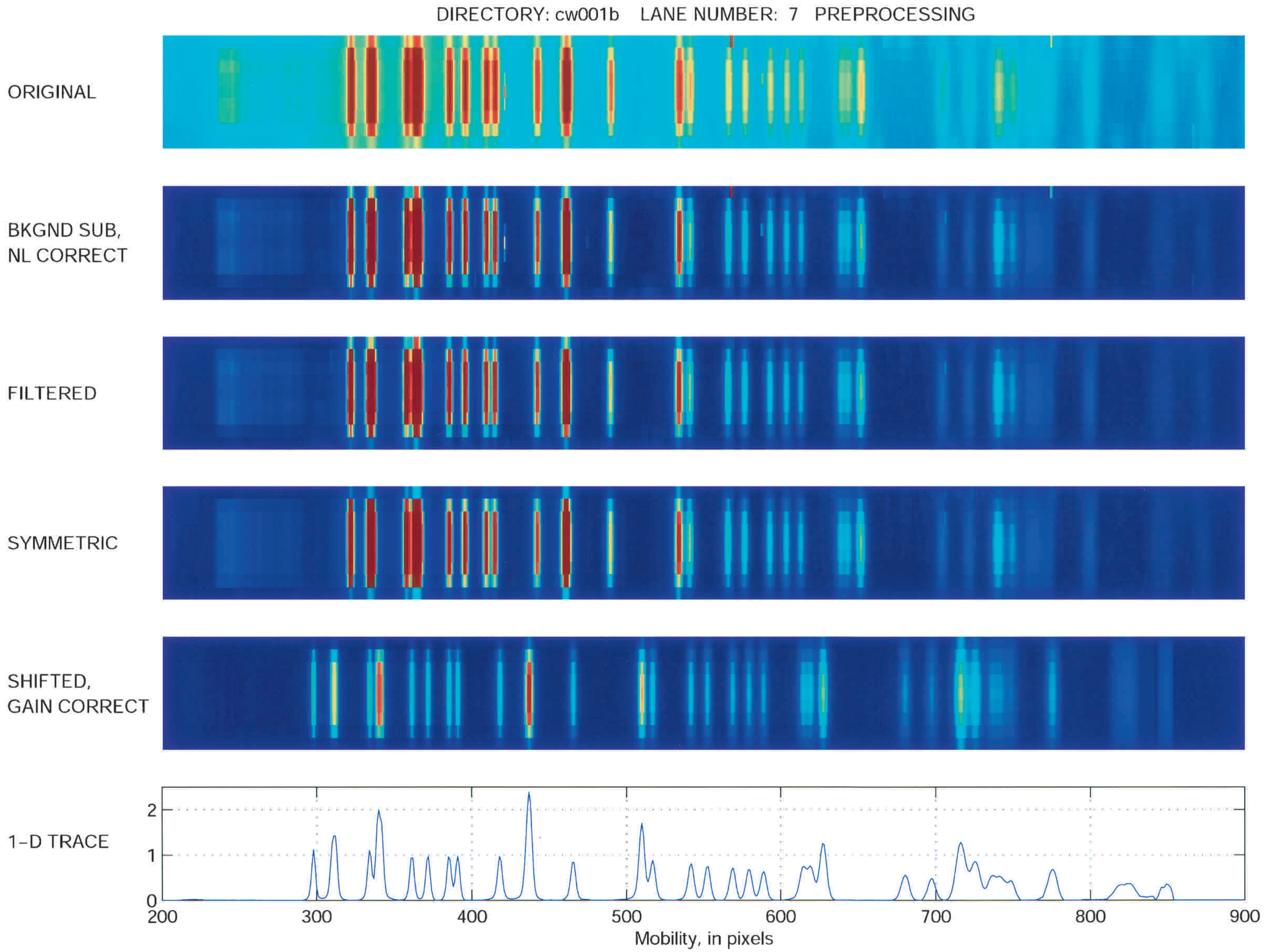

Figure 3 Data lane preprocessing steps. *Panel 1:* Raw image data (after lane tracking) using MATLAB "jet" colormap. *Panel 2:* Result of background subtraction and correction for pointwise nonlinearity. *Panel 3:* Result of impulsive noise filtering. *Panel 4:* Result of enforced symmetry and monotonicity constraints. *Panel 5:* Result of gain correction and mobility shift to standard location. *Panel 6:* Extracted one-dimensional trace.

but vary slightly even within one gel due to subtle variation in the gel conditions and nonuniformities in the electric field. Once the marker band locations have been determined, the fragment size/mobility relation can be determined for every data lane by interpolation across the gel and down each lane. After the locations of the bands have been identified, the shapes of the bands are also analyzed to develop the templates needed for a complete linear model for the data lanes in a gel.

Marker Band Detection

The first step in the marker lane analysis is the image preprocessing described previously. The templates used for the across-lane matched filtering are taken from a "standard" generated by the analysis of a typical gel produced under a given experimental protocol.

In a marker lane there are 37 bands (Fig. 1B), most of which are distinct and easily identified, except for the pair numbered 18–19, and the group of seven at high mobility, numbers 31–37. All of the marker lanes appear similar, differing only in some translation and distortion of the mobility axis, and in the overall lane amplitude. Thus, the primary task in marker lane analysis is to fit a distorted version of a standard template to the marker trace. In this respect, the analysis has much in common with algorithms in pattern matching or pattern recognition using deformable templates (Grenander and Miller 1994; Jain et al. 1996; Zalubas et al. 1997).

In the marker template, the first 17 bands form a distinctive and easily recognized pattern. This pattern is approximated by a translated and dilated version of a standard template, with all the band peak amplitudes equal. The top section of the trace is matched to a set of 4000 versions of the template (100 translations times 40 dilations) until a best fit is found.

As the distortion of the mobility axis may be something other than a simple translation and dilation, each marker band must be individually isolated. This is accomplished by sequentially finding each band using a prediction based on previously identified bands. This sequential procedure is carried out beginning at marker band 9 and operates in both directions, up and down the trace, from this point. Quadratic peak-finding is used to identify peak locations to subpixel accuracy for known singlets, whereas a slightly different version of the previously described pattern-matching procedure is used for bands that do not have clearly identifiable peaks.

Marker Band Verification

The accuracy of the bandcalling depends critically on the correctness of the marker lane analysis; in short, there is little room for error at this step. Accordingly, measures must be taken to ensure that the marker lane analysis was successful. For verification, we generate a synthetic marker trace using the called band locations and the nominal band shapes. The correlation between the synthetic and the true (preprocessed) trace is then computed. This correlation must exceed 0.95; otherwise, the marker lane is discarded.

Additional steps are taken to verify the marker lane analysis, once all the individual lanes have been called. The 25 marker lanes across the gel are examined for any discrepancies. For each marker band *k*, the 25 called mobilities across the gel, $m_k(I), I = 1 \cdots 25$, should form a smooth curve. Each function $m_k(I), I = 1 \cdots 25$, is fit to a low-order polynomial. Any called locations that deviate significantly from this curve are replaced by an estimated mobility found by polynomial interpolation. The same interpolation procedure can be used to replace data from "bad" marker lanes discarded by the correlation analysis.

Figure 4 shows an image depicting the raw data from all 25 marker lanes in a typical gel, with results of the marker lane analysis superimposed. In this figure we have shown only the high-molecular-weight bands at low mobility (roughly the top half of the gel) to better illustrate the performance. Note the subtle but significant variation in the marker band location from lane to lane, the very reason that accurate marker lane analysis is critical.

Band Shape Analysis

The marker bands, once identified, can be used to develop a complete band shape model for a fingerprinting gel. This is done by an empirical analysis of the second moments of the bands, and fitting these to a sequence of second moments consistent with the model. Because the model band shape is separable, we can analyze the horizontal and vertical moments separately. A horizontal band is found by summing pixels in the vertical direction, and vice versa. Furthermore, for shape analysis, the horizontal and vertical bands are easily normalized to unit area.

The analysis of the band shapes proceeds by computing the horizontal and vertical second moments of the 28 well-resolved singlets in the marker lanes. According to our model, the second moment of each band can be attributed to three sources: (1) a fixed rectangular pulse, (2) a fixed Gaussian pulse with different horizontal and vertical widths, and (3) a variable-width Gaussian pulse with circular symmetry and width increasing with mobility. We have found that a useful model describing the diffusion is that the standard deviation (square root of the second moment) of the variable Gaussian pulse grows quadratically with mobility. A complete description for the band shapes is found by fitting the two sequences $\sigma_h^2(1) \cdots \sigma_h^2(28)$ and $\sigma_v^2(1) \cdots \sigma_v^2(28)$ to a model that incorporates all the features described above.

Because of the computational impracticality of building a separate model for each lane in the gel, the results of the analysis for all 25 marker lanes are combined to give an "average" band shape model for the gel. From the band shapes and the known fragment sizes, a nominal model for the amplitude curve can be generated as well. All of this information is combined to generate a set of templates and other data structures used in the data lane analysis, which we call the complete gel model.

Data Lane Analysis

After the marker lanes have been analyzed, and a full parametric model has been developed for the gel, the analysis of the data lanes with the unknown fragments can be carried out. The approach used is one of *analysis-by-synthesis*, wherein synthetic data are generated and matched to the true data.

The basic data model, after the preprocessing described previously, is given by

$$\tilde{T}(y) = \sum_{k=1}^{K} \tilde{A}_k B_v(y - m_k, F_k) \tag{5}$$

The band shapes are assumed to be exactly known, and the amplitudes $\tilde{A}_k$ are nominally all equal to a constant. The amplitudes will be subject to slight corrections as the analysis progresses. The objective of the data lane analysis is to determine a set of fragment sizes f_k which when used to generate a

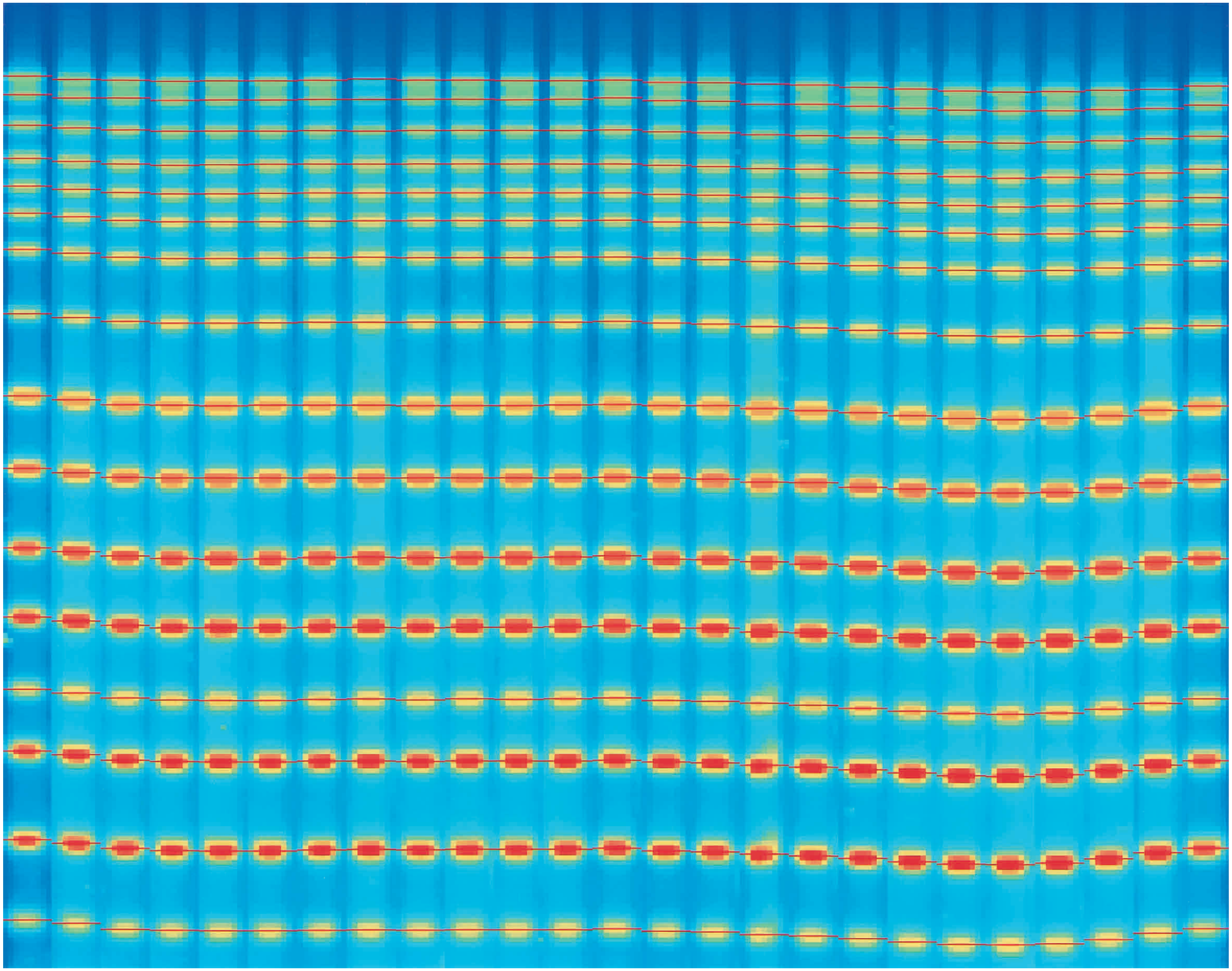

Figure 4 Results of marker lane analysis, showing low-mobility, high-molecular-weight marker bands 1–16. Shown in false color are 25 marker lanes, isolated from the full 121-lane gel and juxtaposed. Superimposed are red horizontal bars indicating the marker band locations as determined by BandLeader's marker lane analysis.

synthetic trace according to the model of Equation 5, provide the best least-squares fit to the preprocessed data.

We adopt a discrete implementation of the model, in which the possible mobilities m_k are quantized onto a grid of 1500 possible values, logarithmically spaced between a minimum and maximum mobility determined by the modeling step. Typically this leads to step sizes on the "mobility grid", as it is called, of approximately 0.2 pixels at low mobilities and 1 pixel at high mobility. This corresponds roughly to the resolution available from the band shapes, which decreases with increasing mobility. The typical quantization error in mobility leads to errors on the order of 0.25% in fragment size, ignoring other bandcalling errors.

The reason for the discretization of the mobilities onto the mobility grid is that it simplifies the search procedure. We use a search algorithm that shares properties of both a gradient algorithm and exhaustive search. As the band shape model is stored in a look-up table, it is not possible to compute gradients analytically; a numerical approach is required.

Cluster Analysis

One of the characteristics of the trace $T(y)$ is that the bands tend to occur in isolated groups containing typically anywhere from 1–10 or 12 bands. We call these groups *clusters*. In the space between the clusters, the signal value is near 0, and this fact can be used to isolate clusters. In effect, by searching for signal-absent regions the trace is broken down into a sequence of contiguous signal-present and signal-absent regions. In this way the global model-fitting problem is reduced to a number of much smaller local model-fitting problems.

Grid Search

Following the partitioning of the trace and the mobility grid into isolated clusters, each cluster is analyzed for the best model fit. Suppose that a cluster occupies pixels $N_1 \cdots N_2$ and that these same pixels correspond to mobilities $M_1 \cdots M_2$ on the mobility grid. Define $N = N_2 - N_1 + 1$ (number of pixels) and $M = M_2 - M_1 + 1$ (number of mobilities to test). Define the test vector as $\mathbf{s} = \boldsymbol{T}[N_1 : N_2]$ in MATLAB notation. We seek a model of the form

$$\mathbf{s} = \mathbf{A}\mathbf{x} \qquad (6)$$

where $\mathbf{A}$ is an $N \times M$ matrix whose columns contain the individual band model. $\mathbf{x}$ is a vector of M integers, describing the finite combination of bands to include in the model fit. Most of the entries of $\mathbf{x}$ will be either 0 or 1, but our model does allow for multiple copies of bands at the same mobility.

The knowledge of the amplitudes of the bands, or equivalently the fact that the entries of the solution vector $\mathbf{x}$ are integers, eliminates the model-order problem which often plagues model-fitting procedures. There is no risk of overfitting the data with too many bands. Increasing the number of bands over that which gives the optimal fit will simply increase the error between the data and linear combination; thus the fitting procedure is in a sense self-limiting.

We have crafted a hybrid numerical gradient search to solve the model-fitting problem for one cluster. We adopt a cost function $h(\mathbf{s},\mathbf{Ax})$, and seek the value of the vector $\mathbf{s}$ which minimizes this cost function. For simplicity, the details of the search algorithm are omitted here. The cost function is a modified least-squares function, where the modifications address the uncertainty in the amplitude curve. The modified cost function places more emphasis on the *shape* of the target function, and less on its amplitude.

Amplitude Curve Estimation

The determination of the amplitude curve a_k, $k = 1 \cdots 1500$ is critical to the success of the algorithm described above. Nominally, the amplitude curve is known to within a single scale factor prior to the data analysis. However, the amplitude curve varies from lane to lane, and the model based on integrated intensities is not sufficiently predictive to be used without modification. Accordingly, the full data lane analysis requires three passes through the data, with refinements of the amplitude curve at each pass.

Pass 1. The amplitude curve is found by scaling the normalized amplitude curve by a factor α, where α is chosen so that 15% of the values in the trace vector $\mathbf{y}$ are above the curve, and the remaining 85% below. We have found that this normally causes the adjusted curve to "hug" most of the single peaks, and that it allows the multiplet peaks to exceed the curve. Using this scaled nominal amplitude curve, the algorithm described above is run; however, only clusters with singlets and resolved doublets are retained.

Pass 2. The normalized amplitude curve is again scaled by a factor α, this time chosen to achieve a least-squares fit between the trace vector $\mathbf{t}$ and the retained clusters. This new amplitude curve is again used in the gradient search procedure, and this time all the clusters are retained.

Pass 3. The amplitude curve is multiplied pointwise by a cubic polynomial. The coefficients of this polynomial are chosen to minimize the squared error between a synthetic trace and the data.

The results of the analysis of a typical data lane are summarized graphically in Figure 5. The top panel shows the image of the data lane in false color, after preprocessing. The second panel shows the one-dimensional trace and the nominal amplitude curve based on the 15% rule plus the Pass 1 bandcalls indicated as small black circles. The fourth panel shows the same trace with the Pass 3 amplitude curve and the final bandcalls, indicated with small red circles. The bottom panel contains a synthetic trace, generated according to our forward synthesis model using all the results of the data lane analysis. The agreement between the model and the preprocessed data is evident here; the correlation between the actual trace and the synthetic trace is 0.98 in this example.

Exception Handling

Several heuristic safeguards have been built into BandLeader to detect data lanes that are defective in some sense, and also to recognize when there has been an error in processing and thus the results cannot be used with confidence. Specifically, there are eight conditions that will generate errors and four conditions that will generate warnings.

An error will cause the lane data and any bandcalling results to be discarded, while a warning is recorded in a log file for further manual inspection if desired. Most conditions are tested on the preprocessed one-dimensional trace signal. Different tests occur at different points in the processing.

The conditions that generate errors are as follows:

1. *Empty lane*. The total signal level is below a threshold determined from the signal level in the marker lanes.
2. *Nonrecombinant lane*. Thirty percent of the total signal is found within a single 10-pixel window. Nonrecombinant clones are those which contain the vector DNA without any insert DNA, causing there to be just one or two medium-sized bands, depending on how many enzyme motifs are contained in the vector.

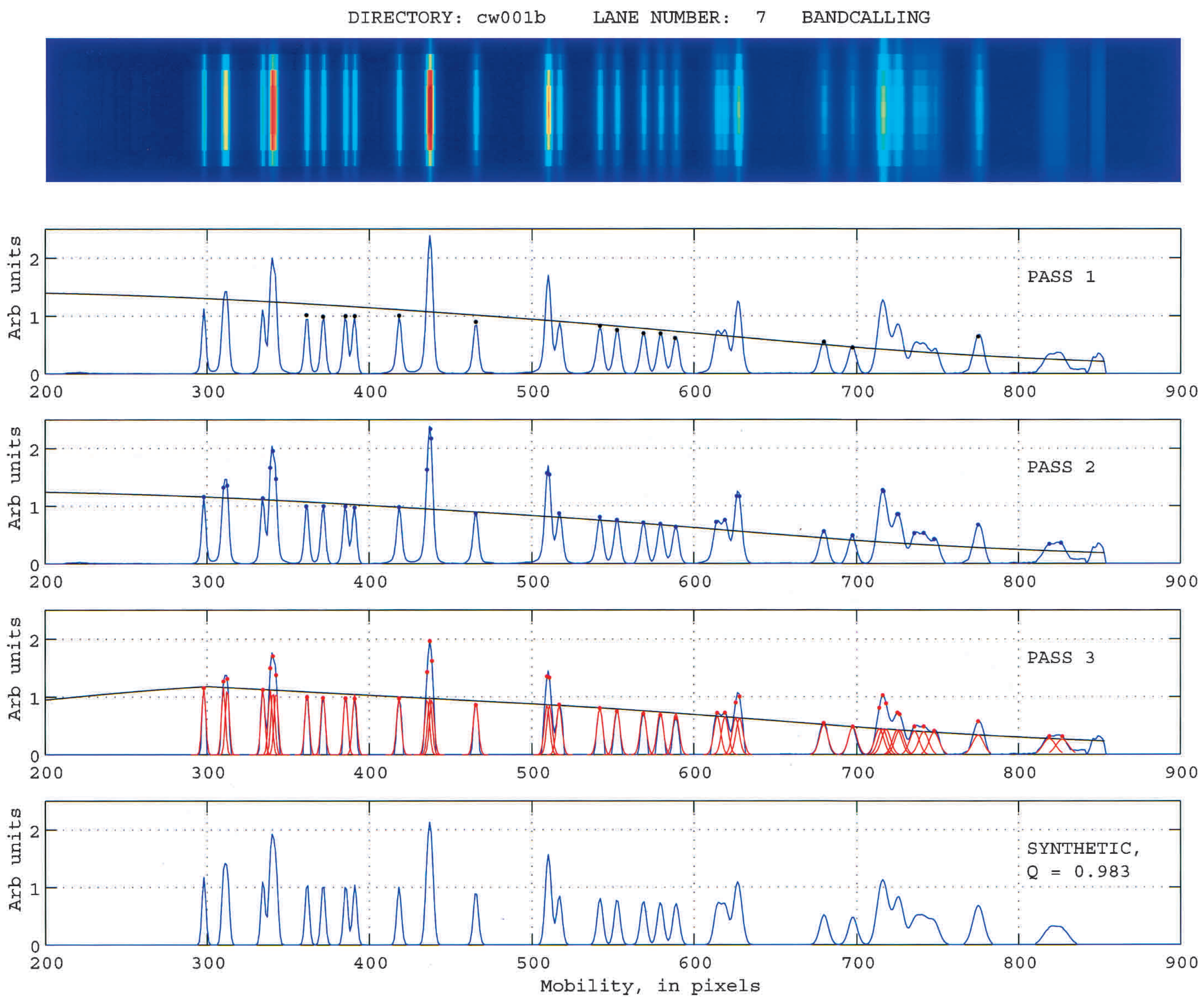

Figure 5 Data lane bandcalling steps. *Panel 1:* Image data after preprocessing. *Panel 2:* Result of bandcalling, first pass. *Panel 3:* Result of bandcalling, second pass. *Panel 4:* Result of bandcalling, third pass, with individual bands superimposed in red. *Panel 5:* Synthetic trace based on called bands and estimated model parameters.

3. *Low-mobility concentration*. The total signal in the first 100 pixels is greater than 80% of the total signal in the lane.
4. *Overcount*. The sum of all called fragment sizes exceeds a user-specified limit (e.g., 350 kbp).
5. *Undercount*. The sum of all called fragment sizes is below a user-specified limit (e.g., 50 kbp)
6. *Poor quality measure*. The correlation between the preprocessed trace and a synthetic signal generated using the bandcalls as input to our forward model is less than 0.9.
7. *No singlets found*. No singlets were identified in the first bandcalling pass, thus making amplitude curve estimation impossible.
8. *Unknown error*. A run-time software error such as divide-by-zero or subscript out-of-bounds is trapped by the MATLAB error handling routines. This prevents any remaining "bugs" in the software from halting production bandcalling, although naturally it is a cause for concern and usually leads to investigation and correction of the problem.

The conditions that generate warnings are as follows

1. *Low-mobility contamination*. Some lanes contain high-molecular-weight genomic DNA which does not belong to either the vector or the insert. There are two tests for this condition: (a) The number of pixels that are assigned to band clusters in the first 180 pixels exceeds 60, and (b) the number of bands called in the first 20 pixels exceeds three. There is some modification to the processing under these conditions; under condition (b), all bandcalls in the first 10 pixels are disregarded.
2. *Possible overcall*. There are 10 or more bandcalls in any region of four or fewer pixels.
3. *Saturation*. The number of pixels in the high-mobility region of the lane which are set equal to the largest output value of the imaging A/D converter exceeds a given threshold.
4. *High-mobility concentration*. The total signal level in the last 200 pixels exceeds the total signal in the first 400 pixels.

All of the errors and warnings are recorded in a separate log file for each gel analyzed. In our experience it is rare that all 96 data lanes are analyzed without error, with an average of nine lanes per gel generating a warning or error record in the log files.

Assessment of BandLeader Performance

To evaluate BandLeader's performance on agarose fingerprinting gels, we identified a "test set" of 140 human BAC clones and 185 mouse BAC clones. These were selected from among a set of BACs, available in GenBank (http://www.ncbi.nlm.nih.gov), that had been sequenced completely and accurately (i.e., were "finished") at Washington University Genome Sequencing Center, the Whitehead Institute for Biomedical Research Sequencing Center, or at the Sanger Institute. We generated *Hin*dIII fingerprints of the BACs using our standard laboratory conditions (Schein et al. 2003). The resulting gel images were analyzed with BandLeader, and the bandcall data compared to the "in silico" fingerprints identified by computer analysis of the sequenced BACs (see Methods). In our analyses, the sizes of "in silico" and actual restriction fragments that were within an arbitrarily chosen 2% window were classified as identical. This criterion was applied to all restriction fragments except those less than 600 base pairs (bp). These were excluded because under our standard fingerprinting gel conditions they tend to be diffuse with low levels of signal, obviating their accurate identification by any approach. The adoption of a test set of fingerprinted BACs was crucial, as it provided us with objective "ground truth" test data that made critical evaluation of BandLeader performance possible.

The results of our analyses are shown in Figure 6A. For comparison, we have included the results of a similar analysis performed using automatically generated IMAGE bandcalls (Fig. 6B). Of the 140 fingerprinted human BACs and 185 fingerprinted mouse BACs considered for this analysis, BandLeader accepted 139 and 183, respectively. Analysis of the sequences corresponding to these clones revealed that they contain 16,782 *Hin*dIII restriction fragments. Of these, BandLeader correctly identified 16,134, corresponding to a sensitivity measure of 96.13%. Of the 16,736 fragments identified by BandLeader, 16,138 correctly identified a sequence-predicted fragment, for a specificity measure of 96.42%. For comparison, automated IMAGE bandcalls are only 60.99% sensitive and 88.65% specific. Hence, although BandLeader is not perfect, it offers a remarkable improvement over IMAGE. Further, BandLeader outperforms the manual efforts of even our most experienced technical staff, at throughputs far exceeding those possible by manual analysis of the fingerprinting gels.

A major goal of the BandLeader project was to produce software that would reliably detect multiplets, which we defined as restriction fragments that comigrated within the same data lane on a fingerprinting gel. We assessed the performance of BandLeader in multiplet identification as follows. First, we identified as multiplets all fragments in BAC sequence data falling within a restriction fragment size window of +/− 2%. A similar grouping was done for the BandLeader bandcalls corresponding to these clones. A total of 322 human and mouse BAC sequences were analyzed and 3431 multiplets found in the sequence, for an average of approximately 10 multiplets per sequenced BAC. BandLeader correctly predicted band multiplicity in 96.0 % of these cases (see Fig. 7 and Methods). BandLeader's performance, even on the larger fragment clusters, is striking. For example, BandLeader correctly identifies a cluster of 11 bands. Again, although BandLeader is not perfect, it is distinctly superior to any other bandcalling option we are aware of, manual or automatic.

DISCUSSION

We have described a set of software tools for the automated analysis of agarose gel images acquired during the DNA fingerprinting process. The various steps involved are common to many image analysis and related engineering problems. First, a model was established for the process of electrophoresis and the digital image acquisition. The model included sufficient detail to allow for variations in the model parameters, but was not so complex as to lead to unwieldy image analysis tasks. Based on this model, procedures were derived for determining the model parameters (through the use of marker lanes) and for solving the inverse problem on the data lanes. The integrated suite of tools, written in MATLAB and collectively called BandLeader, has been used for the mouse fingerprint mapping project at the BCCA (Gregory et al. 2002) and is currently being used in fingerprint mapping projects targeting the rat and bovine genomes (J. Schein, I. Bosdet, C. Mathewson, N. Wye, R. Chiu, C. Fjell, H. Shin, S. Jones, and M. Marra, unpubl.).

The use of automated image analysis for high-

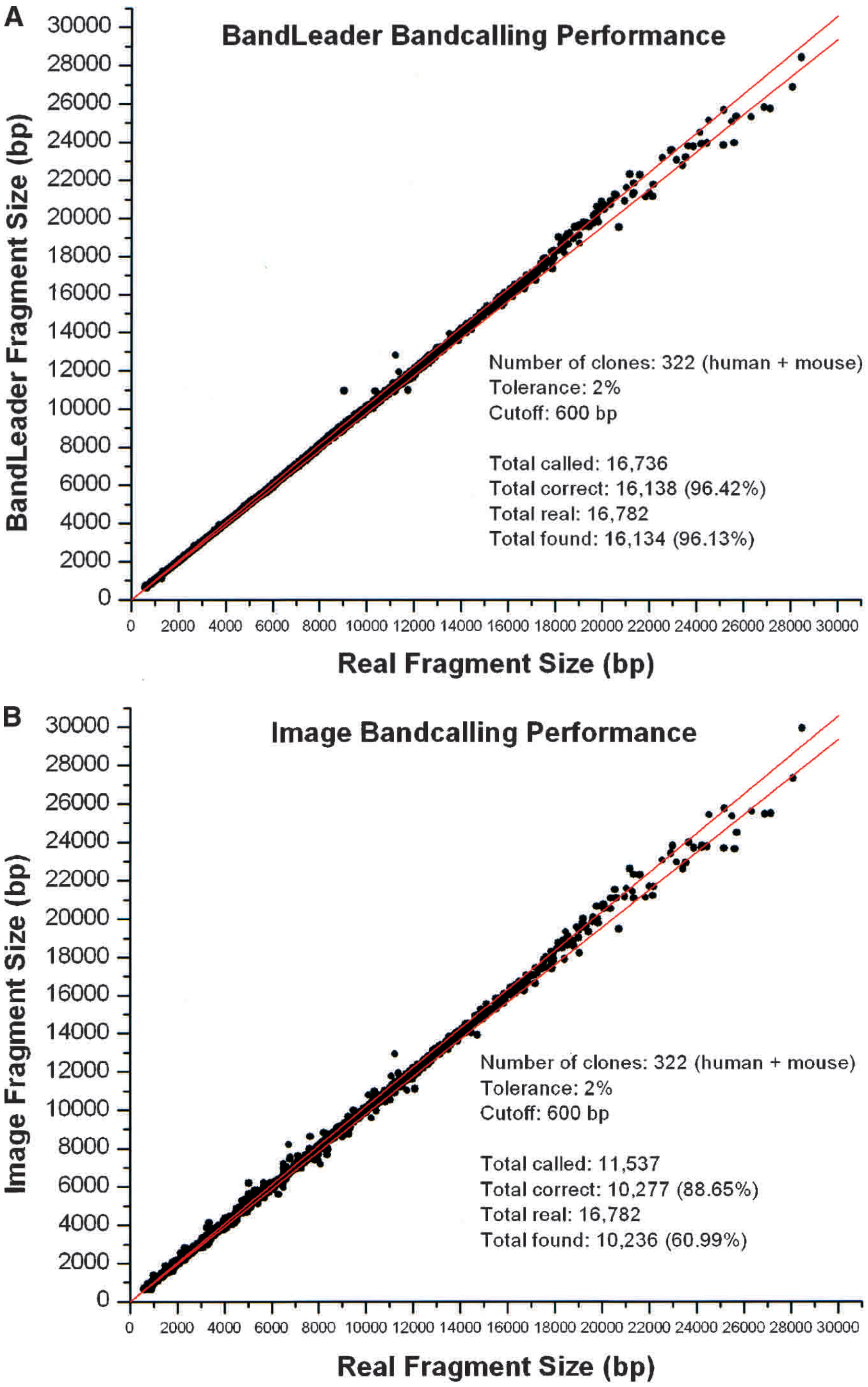

Figure 6 Comparison of BandLeader (*A*) and IMAGE (*B*) performance on test gels. "Real fragment size", plotted on the *x*-axes, refers to restriction fragment size as determined by computer analysis of "finished" human and mouse BAC sequence data. Fragment sizes determined by BandLeader or IMAGE are plotted on the *y*-axes. Fragments less than 600 bp are not considered. Each data point represents the comparison of a restriction fragment predicted by sequence analysis to a restriction fragment identified by either IMAGE or BandLeader. Red lines on the plot indicate a 2% size window. Points falling within the 2% window are considered to represent identical restriction fragments. "Total called" refers to the number of restriction fragments identified by IMAGE or BandLeader. "Total correct" refers to the number of these fragments that match the corresponding sequence-predicted fragments. This is a measure of specificity. "Total real" refers to the number of sequence-predicted fragments, and "total found" refers to the number of these detected by the software. This is a measure of sensitivity.

throughput fingerprint mapping projects has important advantages. Chief among these are the increases in both the rate and accuracy of data analysis and the opportunity to reanalyze the very large fingerprint data sets if more suitable parameters are found. Further, the opportunity exists to repeatedly analyze the gel images to collect statistics. For example, our entire set of mouse (C57BL/6) fingerprints (3500 gels containing more than 330,000 fingerprints) can be reanalyzed in 600 CPU hours. Since each gel analysis is independent, the process is amenable to parallelization, such that only about 24 processors would be needed to reanalyze the 3500-gel mouse set in one day.

The BandLeader software has already proven of enormous value in completing mapping projects that would otherwise be unfeasible given time and budgetary constraints. We have used versions of BandLeader to analyze 13,629 fingerprinting gels, generated in fingerprint mapping efforts aimed at bacterial, fungal, plant, and animal genomes. Although BandLeader's performance is excellent, there is room for incremental improvement. With continued careful modeling and algorithm improvement, we see the potential for increased bandcalling performance, with gains in sensitivity, specificity, and sizing accuracy. One of the more challenging aspects of the automated trace analysis has been the estimation of the amplitude curve, which facilitates detection of multiple bands. The human eye adapts quite easily to model variations and aberrations, and in all but the most pathological cases it is a simple matter for our technical staff to identify singlets visually and hypothesize a smooth curve connecting the peaks. In our automated analysis, it has proven difficult to sort out the singlets, multiplets, and clusters, and derive a reliable amplitude curve. The current version of the software is doing a satisfactory job with this particular task, but on rare occasions, errors may cause a lane to be failed.

There are several safeguards built into the software to recognize bad data and to abort the processing for a given lane. Causes of unusable data include: (1) an empty

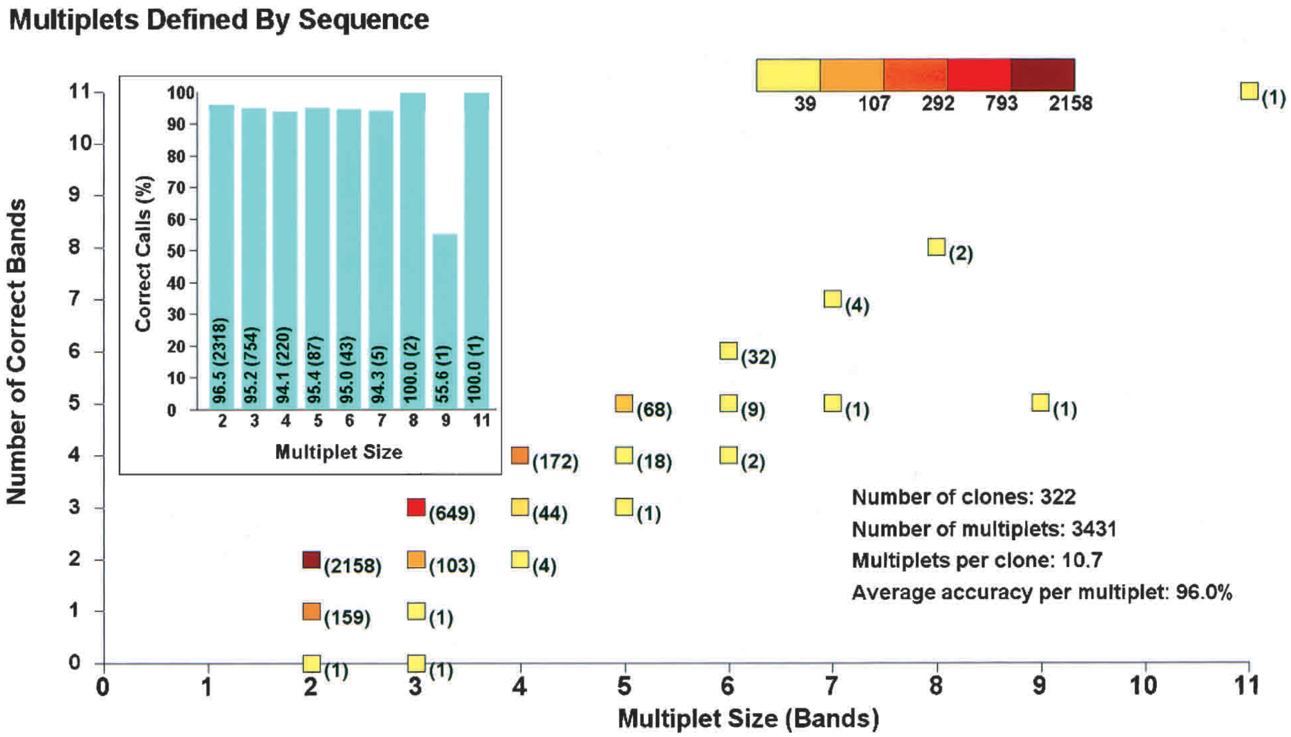

Figure 7 Performance of BandLeader in multiplet identification: 3431 multiplets were extracted from 322 "finished" human and mouse BAC sequences. Plotted on the *x*-axis is the number of fragments per multiplet detected by sequence analysis. Plotted on the *y*-axis is the number of correct bands per multiplet identified by BandLeader. Numbers under the individual data points indicate the number of multiplets detected in the sequences for a given number of bands in the multiplet. For example, there was a single 11-plet identified in the sequences, and BandLeader identified correctly all 11 bands in the 11-plet. In another example, there were two 8-plets identified in the sequences, and all of the bands in all these multiplets were correctly identified by BandLeader. The *inset* summarizes BandLeader performance for all multiplet sizes.

lane, (2) contamination, due to traces of a second clone or other genomic material in the sample, and (3) nonrecombinant BACs, which yield only one or two large bands in the lane. On very rare occasions, there are good lanes for which BandLeader analysis fails, presumably because the fit of the data to the model is poor. One way to recognize this is to compute the estimated clone size by summing all of the detected fragment sizes; if this sum is outside of acceptable limits, the lane can be rejected. In addition, any unexpected software errors (such as a divide by zero) are "trapped" and allow for the analysis of subsequent lanes to continue without the entire process halting.

Currently, BandLeader relies on the data collection format used in our laboratories, and there is no flexibility in gel format or choice of marker DNA. As the fingerprinting data generation protocols are published and the marker DNA is commercially available, this inflexibility is not a major obstacle in the use of BandLeader to support fingerprinting activities in other laboratories. However, we recognize that there are several applications for a more flexible version of BandLeader, including restriction analysis of plasmid and other clones, and also genotyping. Hence, near-term future research and development will focus on methods for the generalization of our techniques to other protocols. For example, we intend to work towards the substitution of restriction digested, sequenced BAC clones in place of the commercially prepared markers. This will permit BandLeader to generate data models from marker lanes that are equivalent to the data lanes, and this in turn is anticipated to positively impact bandcalling accuracy, especially for comigrating restriction fragments. The extent to which accuracy can be improved is limited however, as BandLeader already is capable of 96.42% specificity and 96.13% sensitivity when used in a fully automated mode. This performance, and the robustness and reliability of the code, have made BandLeader the only restriction fragment identification system used in all of the large-scale, high-throughput fingerprinting activities at the GSC.

METHODS

To evaluate the performance of BandLeader, we compared the restriction fragment sizes determined by BandLeader analysis of fingerprinted clones to those generated by in-silico digests of the sequences of the same clones. Three hundred and twenty-five fully sequenced and finished BACs residing in GenBank were identified and the sequences downloaded from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov). All clone sequences were analyzed computationally to generate "in silico" fingerprints. The actual clones were recovered from our local copy of the RPCI-11 library (Osoegawa et al. 2001) and fingerprinted using our standard protocols (Schein et al. 2003). The resulting fingerprinting gels were analyzed by BandLeader version 2.3.3, and the BandLeader-identified restriction fragments were compared to the corresponding in silico restriction fragments. The comparison was performed using a modified version of the Needleman-Wunsch algorithm (Needleman and Wunsch 1970). The modification involved setting a cutoff of 2% of the restriction fragment size in bp, such that only those BandLeader and in silico fragments that were within a 2% size window were classified as identical. In addition, a cutoff of 600 bp was introduced, as our standard laboratory protocols do not yield reliable data for fragments that are smaller than this size. The in silico and BandLeader-generated datasets were each used in turn as the reference fingerprint set, and the percentage of matching bands for all of the test clones was taken and designated as the "sensitivity" and "specificity" merits, respectively.

ACKNOWLEDGMENTS

This work was supported in part by NIH grants 1-U01-HG02042, Sequencing the Human Genome, and 1-U01-HG02155, Sequencing the Mouse Genome. We gratefully acknowledge the support of the British Columbia Cancer Foundation, the British Columbia Cancer Agency (BCCA), and all members of the Mapping Group at the BCCA Genome Sciences Centre. M.M. is a Michael Smith Foundation for Health Research Scholar.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

The *C. elegans* Genome Sequencing Consortuim 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Chen, M., Presting, G., Barbazuk, W.B., Goicoechea, J.L., Blackmon, B., Fang, G., Kim, H., Frisch, D., Yu, Y., Sun, S., et al. 2002. An integrated physical and genetic map of the rice genome. *Plant Cell* **14:** 537–545.

Coulson, A., Huynh, C., Kozono, Y., and Shownkeen, R. 1995. The physical map of the *Caenorhabditis elegans* genome. *Methods Cell Biol.* **48:** 533–550.

Coulson, A.R., Sulston, J., Brenner, S., and Karn, J. 1986. Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **83:** 7821–7825.

Gregory, S.G., Howell, G.R., and Bentley, D.R. 1997. Genome mapping by fluorescent fingerprinting. *Genome Res.* **7:** 1162–1168.

Gregory, S.G., Sekhon, M., Schein, J., Zhao, S., Osoegawa, K., Scott, C.E., Evans, R.S., Burridge, P.W., Cox, T.V., Fox, C.A., et al. 2002. A physical map of the mouse genome. *Nature* **418:** 743–750.

Grenander, U. and Miller, M.I. 1994. Representations of knowledge in complex systems. *J. Royal Stat. Soc. B* **56:** 549–603.

Hoskins, R.A., Nelson, C.R., Berman, B.P., Laverty, T.R., George, R.A., Ciesiolka, L., Naeemuddin, M., Arenson, A.D., Durbin, J., David, R.G., et al. 2000. A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. *Science* **287:** 2271–2274.

Jain, A., Zhong, Y., and Lakshmanan, S. 1996. Object matching via deformable templates. *IEEE Trans. Pattern Analysis and Machine Intelligence* **18:** 267–278.

Marra, M., Kucaba, T., Sekhon, M., Hillier, L., Martienssen, R., Chinwalla, A., Crockett, J., Fedele, J., Grover, H., Gund, C., et al. 1999. A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nat. Genet.* **22:** 265–270.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7:** 1072–1084.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., et al. 2001. A physical map of the human genome. *Nature* **409:** 934–941.

Mozo, T., Dewar, K., Dunn, P., Ecker, J.R., Fischer, S., Kloska, S., Lehrach, H., Marra, M., Martienssen, R., Meier-Ewert, S., et al. 1999. A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nat. Genet.* **22:** 271–275.

Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., and Frank, T. 1986. Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci.* **83:** 7826–7830.

Osoegawa, K., Mammoser, A.G., Wu, C., Frengen, E., Zeng, C., Catanese, J.J., and de Jong, P.J. 2001. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11:** 483–496.

O'Sullivan, J.A., Blahut, R.E., and Snyder, D.L. 1998. Information-theoretic image formation. *IEEE Trans. Info. Theory* **44:** 2094–2123.

Schein, J., Tangen, K., Chiu, R., Shin, H., Lengeler, K.B., MacDonald, K., Bosdet, I., Heitman, J., Jones, S.J.M., Marra, M., et al. 2002. Physical maps for genome analysis of serotype A and D strains of the fungal pathogen *Cryptococcus neoformans*. *Genome Res.* **12:** 1445–1453.

Schein, J., Kucaba, T., Sekhon, M., Smailus, D., Waterston, R., and Marra, M. 2003. High-throughput BAC fingerprinting. In *Bacterial artificial chromosomes: methods and protocols* (eds. S. Zhao and M. Stodolsky), Humana Press Inc., Totowa, NJ (in press).

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci.* **89:** 8794–8797.

Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T., and Coulson, A. 1988. Software for genome mapping by fingerprinting techniques. *Comput. Appl. Biosci.* **4:** 125–132.

Sulston, J., Mallett, F., Durbin, R., and Horsnell, T. 1989. Image analysis of restriction enzyme fingerprint autoradiograms. *Comput. Appl. Biosci.* **5:** 101–106.

Tao, Q., Chang, Y.L., Wang, J., Chen, H., Islam-Faridi, M.N., Scheuring, C., Wang, B., Stelly, D.M., and Zhang, H.B. 2001. Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis. *Genetics* **158:** 1711–1724.

Wechter, W.P., Begum, D., Presting, G., Kim, J.J., Wing, R.A., and Kluepfel, D.A. 2002. Physical mapping, BAC-end sequence analysis, and marker tagging of the soilborne nematicidal bacterium, *Pseudomonas synxantha* BG33R. *OMICS* **6:** 11–21.

Zalubas, E.J., O'Niell, J.C., Williams, W.J., and Hero, A.O. 1997. Shift and scale invariant detection. *Proc. ICASSP (Munich, Germany)* **5:** 3637–3640.

WEB SITE REFERENCES

http://www.sanger.ac/Software/Image; IMAGE software is available at this Sanger Institute site.

http://www.ncbi.nlm.nih.gov; NCBI home page. Access to GenBank database.

Received October 11, 2002; accepted in revised form February 26, 2003.