Am. J. Hum. Genet. 57:1199–1210, 1995

Toward Fully Automated Genotyping: Genotyping Microsatellite Markers by Deconvolution

Mark W. Perlin,[1] Giuseppe Lancia,[2] and See-Kiong Ng[1]

[1] Computer Science Department and [2] Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh

Summary

Dense genetic linkage maps have been constructed for the human and mouse genomes, with average densities of 2.9 cM and 0.35 cM, respectively. These genetic maps are crucial for mapping both Mendelian and complex traits and are useful in clinical genetic diagnosis. Current maps are largely comprised of abundant, easily assayed, and highly polymorphic PCR-based microsatellite markers, primarily dinucleotide $(CA)_n$ repeats. One key limitation of these length polymorphisms is the PCR stutter (or slippage) artifact that introduces additional stutter bands. With two (or more) closely spaced alleles, the stutter bands overlap, and it is difficult to accurately determine the correct alleles; this stutter phenomenon has all but precluded full automation, since a human must visually inspect the allele data. We describe here novel deconvolution methods for accurate genotyping that mathematically remove PCR stutter artifact from microsatellite markers. These methods overcome the manual interpretation bottleneck and thereby enable full automation of genetic map construction and use. New functionalities, including the pooling of DNAs and the pooling of markers, are described that may greatly reduce the associated experimentation requirements.

Introduction

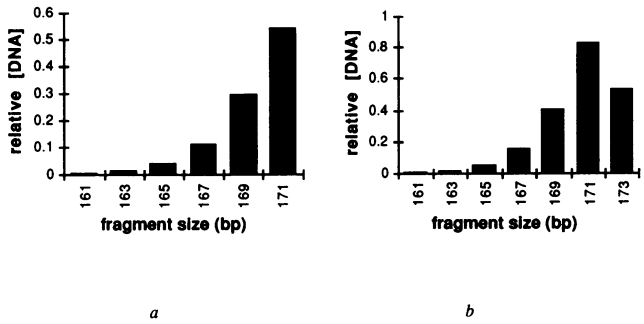
Genetic linkage maps are used to map Mendelian or complex (Ott 1991) traits by first genotyping related individuals with markers that adequately sample the genome of interest and then searching for shared chromosomal regions that are significant for the hypothesis that these regions contain causative gene(s). A variety of statistical methods (Lander and Schork 1994) and computer programs (Lathrop and Lalouel 1988) are used to carry out these genetic localizations. Genetic maps are also used in conjunction with physical maps for the positional cloning of genes (Kerem et al. 1989; Riordan et al. 1989), for diagnosing genetic disease (Schwartz et al. 1992) and assessing tumor progression, and for forensic applications (Jeffreys et al. 1985).

Modern genetic linkage maps (Lander and Botstein 1989) comprising dense informative markers were enabled by the development of recombinant DNA technology. The early RFLP markers demonstrated the power of these maps for genetically localizing Mendelian disorders but entailed Southern hybridization assays requiring substantial laboratory effort. With the advent of PCR (Mullis et al. 1986), short tandem repeat (STR or "microsatellite") marker polymorphisms (Weber and May 1989) replaced RFLPs as the marker of choice. Microsatellite markers are extremely abundant (>100,000 CA-repeat loci), readily identified, highly polymorphic (hence informative), easily shared (as PCR sequence information, rather than as laboratory reagents), and straightforward to assay via PCR amplification and subsequent size (not sequence) determination with gel electrophoresis. Microsatellite-based genetic maps have been constructed for the human and mouse genomes, with average densities of 2.9 cM and 0.35 cM, respectively (Ott 1991; Dietrich et al. 1994; Gyapay et al. 1994; Matise et al. 1994).

The original assays for microsatellite genotyping incorporated radiolabeled dNTPs into DNA sequences during PCR amplification and determined fragment lengths by using standard denaturating sequencing gels (Weber and May 1989). More recently, fluorescent end-labeling of one PCR primer (Clemens et al. 1991; Schwartz et al. 1992) with electrophoresis on automated DNA sequencers has been used to type larger numbers of markers simultaneously and to generate quantitative machine-readable gel files (Schwengel et al. 1994). Some machine-specific software has been applied to these files to assist human operators in determining genotypes (GeneScan/Genotyper for the ABI/373A [Ziegle et al. 1992], ALP for the Pharmacia ALF [Mansfield et al. 1994], and the DuPont Genesis 2000 sequencer [Perlin et al. 1994]). Newer technologies for DNA size separation are being developed that are applicable to microsatellite genotyping, including ultrathin gel slabs (Kostichka et al. 1992), capillary arrays (Mathies and Huang 1992), and mass spectrometry (Wu et al. 1993).

Received May 10, 1995; accepted for publication August 4, 1995.
Address for correspondence and reprints: Dr. Mark W. Perlin, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: perlin@cs.cmu.edu

0002-9297/95/5705-0028$02.00



a *b*

Figure 1 PCR stutter bands arising at microsatellite STR-45 in the DMD region on chromosome X. *a*, from a single allele of size 171 bp (homozygote); *b*, from two closely spaced alleles of sizes 171 and 173 bp (heterozygote) (Perlin et al. 1994). The *X* axis shows the allele size (in bp), and the *Y* axis shows the relative DNA concentrations produced by the alleles' stutter bands.

PCR amplification of an STR allele produces a stutter (or "slippage") artifact, which generates additional (generally shorter) DNA fragments. This may be due to slipped strand mispairing (Hauge and Litt 1993) or polymerase molecule slippage during replication within the repeat region of the DNA sequence. Thus, for example, a CA-repeat allele of total length 150 bp with a $(CA)_{20}$ internal repeat sequence would generate fragments of size 150 bp, 148 bp, 146 bp, . . . , corresponding to replicated $(CA)_n$ repeat units of n = 20, 19, 18, . . . , respectively. The relative concentration of each stutter fragment generally (though not invariably) decreases with fragment size (fig. 1*a*). When two alleles are close in size, their stutter bands on the gel overlap and it becomes more difficult to determine the correct alleles (fig. 1*b*). Mild stutter artifact shows a sharp stutter pattern, with most of the PCR product concentrated in the main allele band, whereas severe artifact shows a flat stutter pattern, with PCR products distributed across multiple bands. In general, a larger number of repetitive units (larger *n* for $(CA)_n$ markers) leads to an increase in both the PIC (i.e., the utility) of the marker and the severity of the stutter artifact.

Large-scale PCR-based microsatellite genotyping has been successfully used to map complex genetic traits (Davies et al. 1994). However, this technology has only been "semi-automated" (Davies et al. 1994; Reed et al. 1994). The key remaining bottleneck is the allele calling of microsatellite data: because of PCR stutter artifact, considerable uncertainty exists when calling either closely spaced alleles of heterozygote individuals or the alleles of pooled individuals from a population. Thus, nearly all laboratories require a human technician to visually inspect the microsatellite data, with associated increases in error, cost, time, and tedium.

Several strategies have been applied to overcome PCR stutter artifact for genotyping applications. These include the following: (1) Using microsatellite markers with fewer repeating units in the alleles (i.e., using $(CA)_n$ markers with *n* small). This approach reduces stutter artifact by sharpening the stutter but also reduces the polymorphism and utility of the marker. (2) Modifying the PCR conditions. This approach works to a point but generally does not remove the artifact, since the stutter is intrinsic to the PCR amplification of a repetitive unit. (3) Shifting from dinucleotide repeat markers to tri- and tetranucleotide repeat markers. Increasing the repeat unit size does reduce stutter but requires the development of more complex, sparser, and less informative microsatellite markers. Further, the larger repeat sizes consume larger size windows (relative to their polymorphism) on the gel. (4) Calling the alleles on the basis of only the highest peaks (Ziegle et al. 1992; Mansfield et al. 1994) and ignoring the others. This approach succeeds when the alleles are widely separated. However, with closely spaced alleles, flat stutter patterns, or uncertain signal measurements, there is too much ambiguity for accurate allele calling. Moreover, this approach cannot work when more than two alleles are present, e.g., when using template DNA pooled from multiple individuals.

We have recently developed a novel approach to eliminating PCR stutter artifact. Rather than trying to suppress or ignore the artifactual bands, we *exploit* these stutter bands to mathematically eliminate PCR stutter artifact and thus determine the correct alleles (Perlin et al. 1994). Our approach enables the correct and fully automated recovery of alleles, both for individual and pooled DNAs, and can work with any DNA size–based separation technology. In the present article, we present our convolution model and deconvolution methods. Computational results are given for the conventional single-genotype situation, for more than two alleles (e.g., pooled individual experiments), and for a novel "stutter-based encoding" approach to pooling microsatellite markers that may significantly reduce experimental requirements. We conclude that the allele-calling bottleneck can be overcome using genotyping microsatellites by deconvolution (GMBD) methods; GMBD may enable fully automated computer-based application of genetic linkage maps.

Convolution Model of PCR Stutter

Convolution Model

For a given microsatellite marker assayed under fixed PCR conditions (including enzyme, cycle times, number of cycles, template and primer concentrations, and buffers), PCR amplification generates reproducible stutter patterns for each allele, even when different template

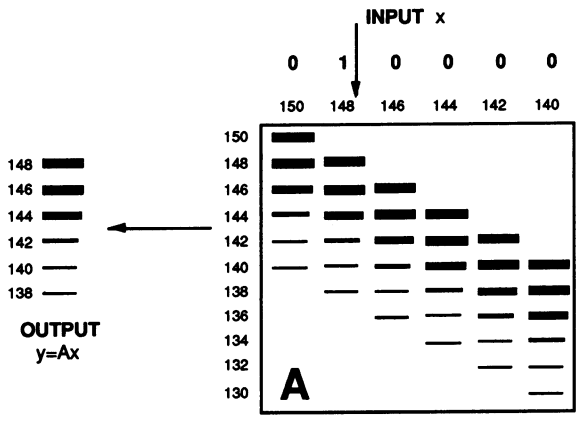


Figure 2 Action of the allele stutter pattern matrix A against one allele. The allele is encoded as a 1 in the genotype vector x, and the convolving matrix A acts to predict the observed PCR amplifier pattern at that allele by selecting the appropriate column. The response is a data vector y, which would be observed on gel electrophoresis as a series of bands. The numbers indicate illustrative allele sizes (in bp).

DNA samples are used (Perlin et al. 1994). Thus, PCR may literally be thought of as an *amplifier*. With perfect fidelity, the DNA fragment corresponding to an allele would be perfectly reproduced as a *single* band on a gel. However, when the PCR amplifier is imperfect (as with microsatellite markers), a distortion response is introduced, and an allele generates the *multiple* bands observed on a gel that correspond to the marker allele's (reproducible) stutter pattern. It is well known from electronic signal processing (Papoulis 1977) that the reproducible responses of an amplifier can be accurately modeled as a *convolution.*

The marker's PCR amplifier responses (i.e., stutter patterns) of the alleles can be written down in a matrix A, where each column corresponds to one allele, and the row entries in each column give the response (i.e., stutter pattern) of the PCR amplifier to that allele (fig. 2). A genotype x can be represented by a vector whose components correspond to allele sizes. If the actual genotype x of a haploid DNA sample has one allele, this allele can then be written as a vector having a 1 entered in the allele's size component, and having 0 entered in all other size components. When the PCR amplifier (modeled by the matrix A) acts on this allele (modeled by the vector x), and the DNA products are size separated on an electrophoretic gel, the response is a complex stutter pattern modeled by the matrix-vector product

$$y = Ax .$$

This vector y predicts the relative DNA concentrations that are present in the PCR product, and observed on the gel. (Note that this convolution model is not linear shift-invariant, since different alleles may contribute different stutter patterns; that is, the columns of A may differ.)

When the actual genotype x of a diploid DNA sample is the two alleles x_1 and x_2, the PCR amplifier matrix A acts on these alleles' vectors to generate the two response vectors (stutter patterns) $A \cdot x_1$ and $A \cdot x_2$ (fig. 3). When these DNA products are size separated by gel electrophoresis, the allele stutter patterns $A \cdot x_1$ and $A \cdot x_2$ are combined, and the observed DNA concentrations at each band is the sum of the contributions from each of $A \cdot x_1$ and $A \cdot x_2$. That is, the observed response pattern is

$$\begin{aligned} y &= (A \cdot x_1) + (A \cdot x_2) \\ &= A \cdot (x_1 + x_2) \\ &= A \cdot x , \end{aligned}$$

where x is the vector sum of the allele vectors $x_1 + x_2$.

Once the PCR amplification responses (under fixed PCR conditions) of every allele for a microsatellite marker are determined, these stutter patterns can be recorded as the column vectors of the matrix A. Then, given any collection of DNA allele concentrations, the components can be represented by a vector x. (For example, the two allele entries of a heterozygote genotype would be 1, and all other entries would be 0.) To predict the observed PCR stutter pattern of DNA concentrations for this marker A with the heterozygote alleles x, one can apply our convolution model by computing the matrix-vector product

$$y = Ax ,$$

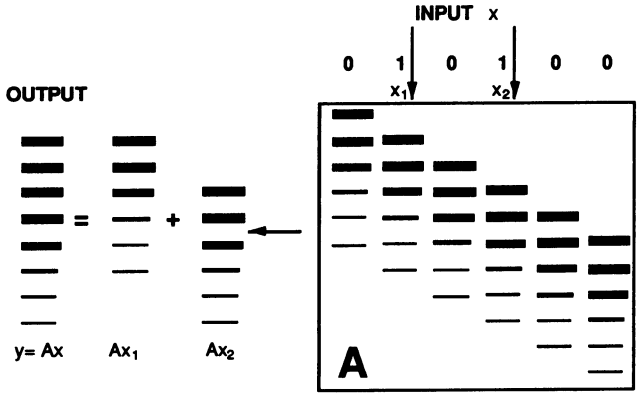


Figure 3 Action of the allele stutter pattern matrix A against two alleles. These alleles are encoded as 1s in the genotype vector x, and the convolving matrix A acts to additively superimpose the selected PCR amplifier patterns of those two alleles. The response predicts the observed gel data vector y.

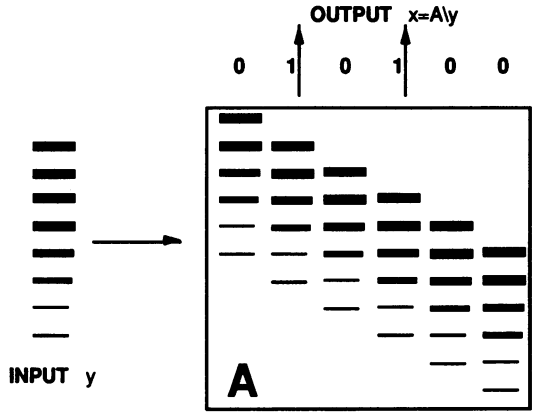


Figure 4 One deconvolution procedure. The stutter pattern matrix A is divided by the observed (input) data vector y to compute a "best" (output) genotype vector $\hat{x}$ that fits the data.

where the vector y contains the sum of the different allele contributions, thereby providing an estimate of the relative concentrations of DNA that would appear in the bands on the gel.

GMBD

To mathematically remove PCR stutter artifact from a genotyping experiment, one wants to compute the actual alleles x, given the observed gel data y (i.e., DNA sizes and concentrations) and the predetermined PCR amplification response matrix A of the marker (fig. 4). This is done using our convolution model by *deconvolving*, i.e., solving the inverse matrix-vector division problem:

$$x = A\backslash y\ .$$

This matrix division can be computed by any number of common numerical procedures, the most general and robust being a least-square's fit, such as a matrix singular value decomposition (SVD) (Press et al. 1988). These numerical procedures can be implemented by entering the unanalyzed stutter data y and the predetermined stutter responses A into computer files and then using numerical analysis tools (MatLab, The Mathworks) or programs written in a general purpose language (Common LISP, C) to perform the deconvolution. Less general deconvolution procedures for resolving closely spaced alleles in related individuals have been described by us (Perlin et al. 1994).

Deconvolution Methods

Deconvolution algorithms for three types of genotyping problems are described below.

The Single-Genotype Problem

Most current microsatellite genotyping centers on the determination of a single genotype, i.e., the (up to) two alleles present in one individual's DNA at one marker. The first deconvolution algorithm described (Perlin et al. 1994) for this problem was a statistical moment-based approach tailored for molecular diagnostics on the X chromosome rather than a general genotyping algorithm for use on any chromosome. We developed and evaluated six algorithms specifically designed for general genotyping, in the following two categories:

i) Three (linear shift-invariant) algorithms that use a single PCR stutter pattern vector a *that is independent of allele size*

- POLY.—Polynomial divides the stutter vector a by the data vector y in order to estimate the genotype vector x.
- FFT.—Fast Fourier transformation deconvolves the data vector y with the stutter vector a to recover the genotype vector x. This is done by dividing the FFT of y by the FFT of a and then recovering the deconvolved vector x by an inverse FFT.
- WIENER.—FFT method with additional Wiener filtering (Press et al. 1988) filters out possible noise from the observed data. A data-derived noise filter is used, assuming that noise arises from low-power inteference and does not exceed 15% of the observed data. ($|\Phi(f)| = \min(s_p, .15 \cdot S_p)$, where s_p and S_p are, respectively, the minimum and maximum values of the data's power spectrum.)

ii) Three algorithms that use a marker's allele size–dependent PCR stutter patterns, recorded in a matrix A

- SVD.—SVD inverts the stutter matrix A and applies this matrix inverse to the data vector y, thereby recovering the genotype vector x.
- GAUSS.—A Gaussian elimination procedure starting from the rightmost peak (largest allele size) successively subtracts off each allele's stutter pattern. This procedure provides a robust mechanism for inverting the allele stutter matrix A and applying it to the data vector y.
- ENUM.—Direct enumeration (exhaustive search) of all feasible genotypes x looks for the least error between the observed data vector y and the estimated vector Ax.

The three algorithms SVD, GAUSS, and ENUM are specifically designed to accommodate stutter patterns that vary with allele size and would be expected to perform best on actual data. The algorithms POLY, FFT, and WIENER are more conventional signal-processing algorithms that assume (usually incorrectly) that the stutter pattern does not vary with allele size. The algo-

rithms currently used in commercial genotyping software make no constructive use of stutter information.

Pooled DNAs for Population Studies

For some applications, it can be useful to pool together individual DNAs for PCR and/or gel readout. For example, the allele frequencies for each marker of the population under study are often valuable in linkage analysis (Kruglyak et al. 1995). As another example, when a founder effect is present in an ethnically homogeneous population, individuals expressing a genetic trait may be presumed (with high confidence) to share by descent a common chromosomal region containing the causative gene. Since meiotic events tend to retain flanking chromosomal regions in direct relation to the proximity of the gene (Feingold et al. 1993; Kobayashi et al. 1995), a loss of allelic heterozygosity (or linkage disequilibrium) in the gene region can localize the trait on a genetic map. Sufficiently dense chromosomewide or genomewide genotyping with microsatellites on pools of affected individuals can gather data for performing this localization. It is significant that the required number of laboratory experiments can be reduced in direct proportion to the size of the pools (e.g., 100-fold, with pools of 100 individuals). However, PCR stutter artifact has thus far precluded such quantitative pooled population PCR-based genotyping studies.

When our matrix convolution model is applied, the problem is modeled as

$$y = \sum_i A \cdot x_i ,$$

where each individual's genotype vector contributes a partial stutter vector $A \cdot x_i$. When rewritten as

$$y = A \cdot \left(\sum_i x_i \right) ,$$

one can combine the data vector y with the stutter matrix A to recover the pooled allele frequency vector $\Sigma_i x_i$.

We have developed six deconvolution algorithms that can solve this genotyping problem. The first five algorithms (SVD, GAUSS, POLY, FFT, and WIENER) were described above and exploit the linearity property of convolution models that allows integer combinations of alleles. When no use is made of PCR stutter information, genotyping of dinucleotide repeat markers is not feasible on pooled data by either visual inspection or computer analysis. Direct enumeration (ENUM) is too computationally prohibitive for practical use with large pools. A sixth algorithm is SEARCH. An initial solution is computed using the allele size–dependent SVD or GAUSS algorithms. A local hill-climbing search procedure (Rich and Knight 1991) is then applied to find a better solution using a statistical model, such as least-squared deviation between predicted and observed allele distributions.

Pooled Markers Using Stutter-Based Encoding

To increase laboratory throughput, PCR products from different microsatellite markers can be pooled together prior to the rate-limiting gel readout step (Reed et al. 1994). With current genotyping analysis methods, at most one marker may appear in any detectable allele size range. This is because a polymorphic STR marker can take on a wide range of possible allele size values, the artifactual PCR stutter bands further extend this range, and the marker pooling organization must ensure that the bands of one marker do not overlap greatly with the bands of another. This requirement for disjoint size ranges imposes several interacting constraints on size-based microsatellite pooling strategies (Reed et al. 1994):

Limited pooling.—To ensure nonoverlapping allele sizes, only a limited number of markers can be pooled in any lane on the gel. The pooling is reduced even further when tri- or tetranucleotide repeat markers are used.

Reduced informativeness.—To allow as many markers as possible in one lane, the size range allocated to each marker should be as small as possible. But since PIC is directly related to the number of possible alleles, selecting microsatellites with small allele size ranges is (by definition) reducing the informativeness of each marker.

Reduced modularity.—A given microsatellite generally does not easily replace another marker in a preexisting pooled set, since its allele size characteristics are fairly unique. Thus, it is not practical to design modular sets that allow diverse markers, particularly microsatellites customized to particular applications.

It is interesting that our deconvolution approach can exploit PCR stutter patterns to *eliminate* these constraints entirely, overcome the current limitations on microsatellite pooling, and thereby increase throughput.

A microsatellite's PCR stutter pattern is generally viewed as an artifact that needs to be eliminated or suppressed. The alternative view taken here is that stutter provides a useful *encoding* of the marker. The idea is that if the stutter patterns from two or more microsatellite markers are superimposed, then they can be separated into their component markers on the basis of their unique stutter pattern signatures. In the decoding process, the alleles are determined. The effect is to enable the pooling of markers *whose allele size ranges overlap,* and thus eliminate the usual constraints on nonoverlapping allele size ranges.

With our matrix convolution model, each microsatellite marker j contributes a stutter matrix A_j. The cumulative effect of each marker's genotype vector x_j is the data vector

$$y = \sum_j A_j \cdot x_j \,. \qquad (1)$$

By combining the observed data vector y together with the predetermined stutter matrices A_j, one can deconvolve to recover the marker allele vectors x_j.

Enumerating all combinations of candidate allele solutions $\{x_j\}$ and calculating each candidate's deviation (e.g., least squared) from the measured data vector y determines the correct alleles for multiple markers. This is computationally tractable. For a microsatellite with n candidate alleles, the number of candidate diploid solutions is n^2. Since n is generally <20 (even for extremely informative CA-repeat markers), n^2 is <400. With k-fold pooling of size-overlapping markers, the total number of integer candidate vectors to explore is n^{2k}. For example, with $n = 20$ and $k = 3$, this set has size 64,000,000. Such sets are amenable to direct enumerative search.

For practical k-fold pooling of size-overlapping markers when $k \geq 3$, we developed dynamic programming methods using branch-and-bound techniques (Papadimitriou and Steiglitz 1983) that considerably reduce the required search effort. The key idea is that virtually all the "feasible," but incorrect, solutions have (least squared) error deviation values that eliminate them as candidates in a branch-and-bound search. Thus, rapid pruning of the search space is possible, and a set of best candidate solutions can be maintained.

Specifically, each node in the branch-and-bound search either fixes an allele or excludes some alleles, and makes these decisions for multiple markers simultaneously. Starting from the largest allele size observed in the pooled marker data, the first component of vector equation (1) is solved to determine which markers could contribute combined alleles that account for the data, within some error tolerance. This monodimensional subproblem is computationally hard but is solved effectively with a dynamic programming approach that exploits the finite width of stutter patterns. Successive monodimensional subproblems are solved for decreasing allele sizes, each subproblem constrained by the results of the preceding subproblems. These combinatorial algorithms were implemented in the C++ programming language on a UNIX workstation.

Additional Material and Methods

Data Sources

Fluorescently labeled microsatellite marker data that can accurately quantitate DNA fragment sizes and concentrations were collected from automated DNA sequencers for testing the deconvolution algorithms. Gel data on dinucleotide-repeat markers was provided by Pharmacia Biotech as detected on their ALF system (Alastair Brown, personal communication) in electropherogram file format and as bands quantitated by their Fragment Manager software for DNA size and concentration. Millennium Pharmaceuticals provided ABI/373A gel image collection files (Jeffrey Thomas, personal communication) that we analyzed for DNA fragment size and concentration. These data were used in the testing, evaluation, and refinement of the algorithms.

Stutter Library Construction

To apply the deconvolution methods to an actual microsatellite marker of interest, the stutter pattern over a range of allele sizes must first be determined. This determination is redone whenever the marker's PCR conditions (hence, stutter patterns) are changed. The allele-size dependent PCR stutter patterns correspond to the columns of matrix A; the task is to determine this matrix A. Since $y = Ax$, from a known set of (column) reference genotype vectors X used to probe A, a corresponding set of experimentally observed data (column) vectors Y can be generated. Note that each set of column vectors (i.e., X and Y) is a matrix. This extends the stutter pattern matrix relation to

$$Y = AX \,,$$

where Y, A, and X are matrices. By matrix division (i.e., numerical solution using least-square minimization) of the under- or overdetermined linear system, the relation

$$A = Y/X$$

allows the determination of the stutter pattern matrix A. Each probing column vector in X can be constructed from one individual (i.e., a known allele pair), or, more efficiently, from a pool of previously genotyped DNAs. The matrix division can be performed in MatLab.

Data Simulator

We constructed a software program in Common LISP for generating simulated microsatellite markers and data. Each simulated marker was a dinucleotide repeat with fragment sizes ranging from 100 bp to 200 bp, having from 10 to 25 normally distributed alleles and an associated simulated stutter matrix A. The number of stutter bands in each allele's pattern was a variable that could be preset for any marker and was typically varied from 3 to 12 bands. Each column of A corresponded to an allele's real-valued stutter pattern vector, with an approximate exponential decay rate inversely

proportional to allele size; columns were normalized to sum to unity. The simulator was used to generate a library of 150 simulated microsatellite markers, including each marker's stutter matrix and allele distribution.

For a given simulated marker, random observed data vectors were constructed from the marker's stutter matrix A and allele frequencies. This was done by randomly generating a genotype vector x with alleles drawn from a marker's allele frequencies, and then setting the random observed data vector y equal to the convolution product Ax plus an additional noise component (see Noise Models, below).

Noise Models

Our model of noise consisted of two components: (1) N_b for random background noise, and (2) N_m for scaled normally distributed measurement error. Given a clean measured value x and a preset noise level of $k\%$, the total simulated noise reading was the sum $N_b + N_m(x)$.

The first component N_b was modeled as a uniformly distributed random variable with support on the interval $[-x_{min}, +x_{min}]$, where x_{min} was the minimum measured value in the simulation. The second component N_m was modeled as a function of the clean measured value x, where $N_m(x)$ was normally distributed, with zero mean and the variance scaled relative to x such that

$$\text{Prob}(-k \leq N_m(x)/x \leq +k) > 0.99 ,$$

where k was the given percentage noise level. This scaling provided a measurement error that was normally distributed and that was within the preset noise level with high probability.

Gel Image Analysis

To quantitatively analyze ABI gel image files, we developed software in MatLab that accurately determines DNA sizes and concentrations of the bands on the gel. Closely spaced molecular weight (MW) markers (20 bp ladder, Bioventures) fluorescently labeled with TAMRA were loaded in each lane together with the PCR products of the microsatellite markers. The fully automated image analysis of the resulting ABI gel image file began by using these MW size markers to construct a mapping between the expected (lane,bp) coordinates and the observed gel image coordinates. The MW marker data was also used to model the peak shape of bands on the gel. For every lane found by the coordinate mapping function, a one-dimensional electropherogram trace was constructed. At every expected base pair location in the lane, the DNA concentration of a detectable band was determined by applying the model peak shape to the electropherogram data. These quantitated (lane,bp) events were recorded for subsequent deconvolution analysis.

Results

Genotyping by Deconvolution: Pharmacia Sequencer Data

PCR fragment products of microsatellite marker D11S527 for 39 individuals from 10 families were size separated and detected in separate lanes of a Pharmacia A.L.F. DNA sequencer. From this data, 23 lanes were identified for which the Fragment Manager quantitation was moderately reliable, showing at least one stutter band per allele and no detector saturation (or other) artifacts. Of these 23 moderately reliable lanes, a subset of 13 highly reliable lanes was identified, showing adequate signal heights, with little or no baseline error. Lane, fragment size (bp), and peak area (DNA concentration) information from the data table produced by Fragment Manager analysis was used, with size (bp) adjusted to evenly spaced integer values. The true genotypes (evident by human visual analysis on this data set) were recorded.

Restricting attention to the 13 highly reliable lanes, we constructed the two calibration tables X_0 and Y_0, where X_0 was the 0-1 matrix of true genotype column vectors and Y_0 was the matrix of observed DNA concentration column vectors, with each column of Y_0 renormalized to sum to 2 (i.e., two alleles present). We then computed D11S527's stutter matrix A using X_0 and Y_0 (see Stutter Library Construction above, under Additional Material and Methods). Missing columns of A were inferred by linear interpolation from neighboring columns.

For the entire set of 23 moderately reliable lanes, we constructed the 0-1 matrix of true genotype column vectors X (a superset of X_0), and the matrix of observed DNA concentration column vectors Y (a superset of Y_0). We performed the MatLab matrix left division operation "$A\backslash Y$" on the previously computed stutter matrix A and the observed data matrix Y to obtain an estimated genotype matrix $\tilde{X}$. The columns of $\tilde{X}$ corresponded to the columns of Y but had the PCR stutter artifact removed by deconvolution. The automated allele calling was 100% accurate on the data set analyzed. Without deconvolution, much of the relative distribution of the bands (28%) is spread out over incorrect alleles; however, with deconvolution, virtually all the distribution (>95%) is located in the bands of the two correct alleles (fig. 5, table 1). This recentering effect relative to the true genotypes X was shown to hold for all the observed data Y and the deconvolved genotypes $\tilde{X}$ (table 1) and can greatly facilitate the unambiguous determination (by human or machine) of the alleles (fig. 5).

Single-Genotype Deconvolution: Simulated Data

We generated 300 random single genotype (i.e., two allele) vectors for markers randomly selected from the set

Table 1

Twenty-three Lanes of Pharmacia Data Genotyped by Deconvolution Using MatLab Matrix Division

Lane	Without Deconvolution[a]	With Deconvolution[b]	Allele 1	Allele 2
3	.699	.950	148	158
4	.696	1.000	154	162
6	.671	1.000	148	162
13	.744	.967	152	156
14	.731	.961	152	156
15	.764	.960	156	158
16	.632	.815	156	156
17	.761	.968	162	164
18	.669	.876	156	164
19	.665	.885	156	164
20	.687	1.000	152	162
21	.719	.905	156	160
22	.747	.992	156	162
23	.751	.786	156	162
25	.786	.890	152	158
27	.771	.960	148	152
29	.792	.980	148	152
30	.798	.981	148	152
31	.668	.928	152	162
32	.660	1.000	152	164
34	.610	.890	152	164
35	.767	1.000	144	156
36	.784	.976	156	158

NOTE.—For every lane, the fraction of band distribution that is centered on the correct two alleles (X) is shown both with ($\tilde{X}$) and without (Y) deconvolution analysis.

[a] Mean = .7205; and SD = .0551.

[b] Mean = .9511; and SD = .0509.

of 150 markers in the simulated marker library. The number of bands produced from the allele and its stutter artifact was set to 5 or 10. From each known genotype x, a data vector y was developed by adding the stutter convolution Ax to noise vectors at 0% and 10% noise levels (see Data Simulator above, under Additional Material and Methods), since measurement errors for fluorescent data peaks from single genotypes are generally much less than 10%. Comparisons were then made between our single-genotype deconvolution algorithms, which estimated x from these generated data sets. For each individual, we compared the number of mismatches between the estimated and known genotypes, and then computed the average number of mismatches per deconvolution for each algorithm. With moderate stuttering (5 bands per allele) or severe stuttering (10 bands per allele), all the deconvolution algorithms showed no mismatches, and all were effective in removing the stutter artifact and correctly calling alleles, at all noise levels studied.

All the algorithms were effective in recentering most of the allele distribution back onto the two correct alleles (table 2). The allele-dependent deconvolution algorithms were more effective than allele-independent signal processing algorithms, particularly in the presence of severe stutter. The least effective algorithm was Wiener filtering, most likely due to inexact modeling of observed noise. The most effective algorithm was complete enumeration, which works by recentering the stutter bands to the best alleles.

Single-Genotype Deconvolution: ABI Sequencer Data

Separate PCR amplifications were performed for eight microsatellite markers on each of 32 individuals. For each individual, the amplified products of these markers (labeled in the FAM, TET, and HEX fluorescent dyes) were added to MW markers labeled in the TAMRA fluorescent dye (20 bp ladder, Bioventures), with size separation and detection then performed on an ABI/373 sequencer. The starting point for our analysis was the ABI gel image file (Jeffrey Thomas, personal communication).

The five markers (d10s186, d11s1347, d16s499, d17s802, and d22s281) that were free of ABI bleed-through artifacts and multiple polymorphisms were selected for deconvolution analysis. The bands on the gel were quantitated for DNA size (in bp) and relative DNA concentration using our fully automated image analysis software (see Additional Material and Methods). One

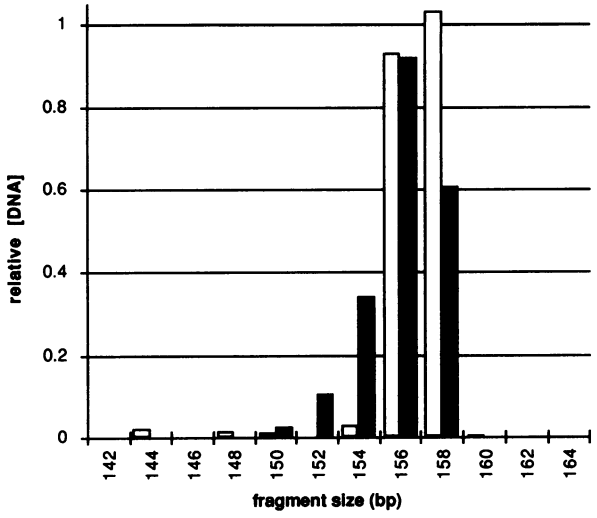


Figure 5 Effect of deconvolution in removing PCR stutter artifact for alleles differing by 2 bp, corresponding to lane 15 in the Pharmacia data for microsatellite marker D11S527. Shown are the allele distribution of the uncorrected data y without deconvolution (*blackened bars*), and the distribution of the corrected data $\tilde{X}$ with deconvolution (*unblackened bars*). The allele distributions are normalized to sum to 2, i.e., the number of alleles present. Note that the distribution corrected by deconvolution is largely (96%) centered on the correct two alleles x.

Table 2

The Fraction of Allele Distribution Centered on the Correct Genotype, Determined for Six Single-Genotype Deconvolution Algorithms

	Noise			
	0%		10%	
	Mean	SD	Mean	SD
Moderate stutter[a]:				
INPUT[b]	.635	.084	.635	.086
POLY	.991	.012	.950	.035
FFT	.990	.013	.948	.036
WIENER	.959	.037	.920	.061
SVD	1.000	.000	.955	.033
GAUSS	1.000	.000	.981	.023
ENUM	1.000	.000	1.000	.000
Severe stutter[c]:				
INPUT[b]	.468	.093	.469	.093
POLY	.977	.023	.921	.042
FFT	.977	.023	.919	.043
WIENER	.896	.102	.852	.110
SVD	1.000	.000	.938	.034
GAUSS	1.000	.000	.979	.026
ENUM	1.000	.000	1.000	.000

NOTE.—Simulation studies were conducted with 300 genotypes of closely spaced alleles that were separated by 0–3 (equally weighted cases) dinucleotide repeat units.

[a] Five bands.

[b] Input data, uncorrected by deconvolution.

[c] Ten bands.

quarter of the data (lanes 25–32) were used to automatically construct the stutter library for each marker (Additional Material and Methods).

Deconvolution analysis for fully automated allele calling was performed on each of the five microsatellite markers for all 32 individuals using the marker's computed stutter library. Our two allele-dependent deconvolution algorithms SVD and ENUM were applied, since our simulation studies had indicated that these would be the most effective methods. With both algorithms, 100% of the 320 alleles were correctly called, relative to manual scoring. SVD was highly effective in recentering the allele distribution, while ENUM was constrained to find the two (correct) alleles.

Pooled DNA Genotyping by Deconvolution

We generated 300 pooled genotypes for markers randomly selected from the simulation library. Each pooled genotype comprised 100 alleles (50 individuals, each with 2 alleles) drawn from the marker's allele frequency distribution. Three hundred noiseless data vectors were constructed by matrix convolution (i.e., $y = Ax$) of a marker's stutter matrix A with its pooled genotype vector x. The number of bands produced from the allele and its stutter artifact ranged from 3 to 12 bands. Noise was then added to these data vectors at 0%, 5%, 10%, and 15% levels (see Additional Material and Methods). For each of the 300 pooled genotypes, we estimated the allele distribution vector x by using our pooled DNA deconvolution algorithms. The average mean squared errors between the estimated and known allele distribution vectors were then determined for our pooled DNA deconvolution algorithms (table 3).

With both moderate 5 band stutter (table 3), and severe 10 band stutter (table 3), the deconvolution algorithms that permitted allele-dependent variation in the stutter pattern (SVD, GAUSS, SEARCH) showed, on average, less error by a factor of 10 than those deconvolution algorithms that assumed a constant stutter pattern (POLY, FFT, and WIENER). With an average mean squared error of $<.10$ for even severe stutter at the 5%–15% noise levels, these allele-dependent deconvolution algorithms could prove acceptable candidates for determining allele frequencies in a population.

Pooled Marker Genotyping by Deconvolution

Simulated pooled marker data were generated for $k = 5$, 6, and 7 dinucleotide-repeat markers using 15 possible alleles per marker. For each value of k, three sets of five problems were generated. Each set corresponded to the maximum relative error allowed in the simulated data, with low ($N_m = .03$, $N_b = .2$), medium (N_m

Table 3

The Average Mean Squared Errors for Pooled DNA Deconvolution Algorithms on Simulated Data

	Noise			
	0%	5%	10%	15%
Moderate stutter[a]:				
POLY	.362	.387	.374	.437
FFT	.192	.203	.221	.276
WIENER	.192	.202	.220	.274
SVD	.000	.029	.058	.093
GAUSS	.000	.031	.065	.101
SEARCH	.000	.021	.048	.079
Severe stutter[b]:				
POLY	.650	.662	.688	.736
FFT	.492	.518	.538	.582
WIENER	.492	.515	.543	.589
SVD	.000	.033	.081	.125
GAUSS	.000	.033	.084	.138
SEARCH	.000	.025	.071	.116

NOTE.—Three hundred simulated pools of 50 genotypes (50 × 2 alleles) were constructed using markers containing from 10 to 25 alleles (normally distributed).

[a] Five bands.

[b] Ten bands.

= .05, N_b = 0.3), and high (N_m = .10, N_b = .5) errors introduced (see Noise Models above, under Additional Material and Methods). For each noise level, microsatellite stutter was generated having six bands per stutter pattern (see Data Simulator above, under Additional Material and Methods). The branch-and-bound algorithm was set to retrieve the five best solutions. In 43 of 45 cases, the correct genotype was found; in 36 cases, this solution minimized the least-squared error. In both cases where no correct genotype was found, the simulated data had a high maximum relative error. This result suggests that the pooled marker approach may prove workable in the more typical (e.g., fluorescently labeled) situation where the signal-to-noise ratio is adequate.

With $k = 5$ pooled markers, the size of the solution space was 5.7×10^{11} possible genotypes, and the average execution time of the combinatorial algorithm was 1 min. With $k = 6$ pooled markers, the solution space size contained 1.3×10^{14} genotypes, and the execution time averaged 2.5 min. With $k = 7$ pooled markers, solution space contained 3.0×10^{16} genotypes, and the run time averaged 4 min. These results highlight the practical advantages of combinatorial search procedures (e.g., branch-and-bound and dynamic programming) relative to brute force enumeration when computing genotypes with pooled markers.

Discussion

The ability to accurately determine the alleles of microsatellite markers would overcome the key bottleneck currently precluding fully automated genotyping. The presence of PCR stutter artifact in dinucleotide repeat data has led to alternative approaches such as tri- and tetranucleotide markers and the use of pedigree information for consistency checking. However, the unbiased use of the very abundant, highly polymorphic, extensively mapped, and easily constructed dinucleotide repeat markers remains a highly desirable goal for effective localization of genetic traits.

Building on our previous work (Perlin et al. 1994), in this article we developed a convolution model for PCR stutter artifact and an associated set of deconvolution methods that can mathematically eliminate this artifact. The reproducible PCR stutter patterns of each microsatellite marker can be measured and then applied as calibration data to remove stutter from new data. Our deconvolution methods were extensively compared on realistic simulation data, and initial testing was entirely successful on data collected from automated DNA sequencers. The most effective algorithms used stutter patterns that depended on each allele of a marker, rather than assuming an invariant pattern for all alleles. Further improvements to the models, methods, and results may be possible by applying more refined statistical (e.g., mixture) models (Devlin et al. 1991).

We have used our allele-calling methods to eliminate "+A" artifact from microsatellite data. This artifact adds to the usual PCR stutter pattern a companion pattern of variable height that is shifted by 1 bp. Intense +A artifact can confound simple allele-calling methods based on peak height, since it produces two peaks of maximum height spaced 1 bp apart. Our approach was to use robust gel image quantitation software to determine highly accurate DNA concentrations at 1 bp intervals. The artifactual +A bands were then mathematically excised from the electropherogram trace. Our usual deconvolution analysis could then be performed on this adjusted data to correctly call alleles.

Our deconvolution methods were applied in three situations. (1) *Single-genotype* analysis is the conventional approach to high-throughput genotyping (Reed et al. 1994), where each unique size region of a gel contains at most two alleles. Considerable effort is currently expended in calling closely spaced alleles, even when using automated fluorescence-based DNA sequencers. We showed on Pharmacia, ABI, and simulated data how our deconvolution algorithms would be effective in reducing this effort. (2) *Pooled DNA* genotype analysis would be highly useful for population studies, including determining allele frequency distributions, and mapping methods based on allelic variation. We showed on simulated data how our deconvolution algorithms could be used to determine such frequency distributions. (3) *Pooling markers* having overlapping size windows would be highly desirable for increased throughput but is not possible with current analysis methods. However, by exploiting PCR stutter artifact, a set of markers can be selected so that each marker's stutter pattern serves as a unique identifying signature, even when the size windows overlap. We showed on simulated data how a demultiplexing analysis could accurately infer genotypes from such pooled marker data, and we assessed highly optimized algorithms for this approach.

Our deconvolution-based genotype studies entail a change in how gel electrophoresis data is viewed. For most molecular biology applications, gel fragment data is understood as discrete all-or-none results that provide qualitative information. In our genotyping convolution models, however, the data signals are necessarily viewed as a sequence of continuous, real-valued quantities. This new perspective changes the relative importance of certain experimental parameters. The measured real-valued quantities (range values of signal) correspond to DNA concentrations, which are accurately quantitated with little noise as electropherogram peak areas in fluorescent detection experiments. However, current automated

DNA sequencer genotyping protocols are not yet optimized for highly accurate determination of fragment size (domain values of signal). This determination can be accomplished by using more closely spaced MW size standards, including partial DNA sequencing ladders, genetic markers for individuals of known genotype, enzymatic cleavage of reference molecules, or chemical modification and cleavage of synthesized polymers. Our results suggest that accurate measurement of DNA sizes and concentrations as real-valued data signals will enable the effective use of our deconvolution methods.

By providing greater accuracy and throughput, GMBD may improve the application of microsatellite marker data to genetic localization. For single-genotype applications, this would result primarily in reduced error and cost. With pooled DNA genotyping, deconvolution methods would reduce by one to two orders of magnitude the number of experiments required in population studies. A qualitative change in the resolution and analysis of genetic studies could result from the novel pooled stutter-encoded marker techniques described here. For example, with 24 lanes, 4 microsatellites per lane, and 3 fluorescent data planes, a typical ABI gel can assay roughly 300 microsatellite markers per individual, i.e., 10 cM genomic resolution. By pooling markers on the basis of PCR stutter, a fivefold improvement could assay 1,500 markers, thereby obtaining 2 cM resolution in a single readout experiment.

Acknowledgments

This work was supported by grant R01 NS32084 from the NIH National Institute of Neurological Disorders and Stroke (M.W.P.). Dr. Clark Tibbetts provided formatting descriptions for the DuPont Genesis and ABI/373A DNA sequencer systems data files. Daniel Richards developed the prototype gel image analysis software used for automatically quantitating the ABI microsatellite data. The Pharmacia/ALF data were generated by Drs. Alastair Brown and Alan Wright of the Medical Research Council (MRC) Human Genetics Unit in Edinburgh and comprise one of the demonstration files (democ.alf at anonymous ftp.hgu.mrc.ac.uk:/pub/ALP) for MRC's ALP software. The ABI data were generated in collaboration with Dr. Jeffrey Thomas of Millennium Pharmaceuticals, Inc. in Cambridge, MA.

References

Clemens P, Fenwick R, Chamberlain J, Gibbs R, de Andrade M, Chakraborty R, Caskey C (1991) Carrier detection and prenatal diagnosis in Duchenne and Becker muscular dystrophy families, using dinucleotide repeat polymorphisms. Am J Hum Genet 49:951–960

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, et al (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371:130–136

Devlin B, Risch N, Roeder K (1991) Estimation of allele frequencies for VNTR loci. Am J Hum Genet 48:662–676

Dietrich WF, Miller JC, Steen RG, Merchant M, Damron D, Nahf R, Gross A, et al (1994) A genetic map of the mouse with 4,006 simple sequence length polymorphisms. Nat Genet 7:220–245

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–252

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, et al (1994) The 1993–94 Généthon human genetic linkage map. Nat Genet 7:246–339

Hauge XY, Litt M (1993) A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. Hum Mol Genet 2:411–415

Jeffreys AJ, Brookfield JFY, Semeonoff R (1985) Positive identification of an immigration test-case using human DNA fingerprints. Nature 317:818–819

Kerem B-S, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Kobayashi H, Matise TC, Perlin MW, Marks HG, Hoffman EP (1995) Towards fully automated genotyping: use of an X linked recessive spastic paraplegia family to test alternative analysis methods. Hum Genet 95:483–490

Kostichka AJ, Marchbanks ML, Brumley RL Jr, Drossman H, Smith LM (1992) High speed automated DNA sequencing in ultrathin slab gels. Bio/Technology 10:78–81

Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56:519–527

Lander E, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lathrop GM, Lalouel J-M (1988) Efficient computations in multilocus linkage analysis. Am J Hum Genet 42:498–505

Mansfield DC, Brown AF, Green DK, Carothers AD, Morris SW, Evans HJ, Wright AF (1994) Automation of genetic linkage analysis using fluorescent microsatellite markers. Genomics 24:225–233

Mathies RA, Huang XC (1992) Capillary array electrophoresis: an approach to high-speed, high-throughput DNA sequencing. Nature 359:167–169

Matise TC, Perlin MW, Chakravarti A (1994) Automated construction of genetic linkage maps using an expert system (MultiMap): application to 1268 human microsatellite markers. Nat Genet 6:384–390

Mullis KB, Faloona FA, Scharf SJ, Saiki RK, Horn GT, Erlich HA (1986) Specific enzymatic amplification of DNA *in vitro:* the polymerase chain reaction. Cold Spring Harbor Symp Quant Biol 51:263–273

Ott J (1991) Analysis of human genetic linkage, rev ed. The Johns Hopkins University Press, Baltimore

Papadimitriou CH, Steiglitz K (1983) Combinatorial optimiza-

tion: algorithms and complexity. Prentice-Hall, Englewood Cliffs

Papoulis A (1977) Signal analysis. McGraw-Hill, New York

Perlin MW, Burks MB, Hoop RC, Hoffman EP (1994) Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular dystrophy. Am J Hum Genet 55:777–787

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1988) Numerical recipes in C: the art of scientific computing. Cambridge University Press, Cambridge

Reed PW, Davies JL, Copeman JB, Bennett ST, Palmer SM, Pritchard LE, Gough SCL, et al (1994) Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. Nat Genet 7:390–395

Rich E, Knight K (1991) Artificial intelligence. McGraw-Hill, New York

Riordan JR, Rommens JM, Kerem B-S, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, et al (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066–1073

Schwartz LS, Tarleton J, Popovich B, Seltzer WK, Hoffman EP (1992) Fluorescent multiplex linkage analysis and carrier detection for Duchenne/Becker muscular dystrophy. Am J Hum Genet 51:721–729

Schwengel DA, Jedicka AE, Nanthakumara EJ, Weber JL, Levitt RC (1994) Comparison of fluorescence-based semi-automated genotyping of multiple microsatellite loci with autoradiographic techniques. Genomics 22:46–54

Weber J, May P (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Wu KJ, Stedding A, Becker CH (1993) Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix. Rapid Commun Mass Spectrom 7: 142–146

Ziegle JS, Su Y, Corcoran KP, Nie L, Mayrand PE, Hoff LB, McBride LJ, et al (1992) Application of automated DNA sizing technology for genotyping microsatellite loci. Genomics 14:1026–1031