Computer Automation of STR Scoring for Forensic Databases

Mark W. Perlin*
Cybergenetics, Pittsburgh, PA

Abstract

Forensic databases are becoming an increasingly valuable law enforcement tool for convicting repeat offenders and exonerating the innocent. However, constructing such databases is quite laborious. After generating STR profiles in the lab, people expend even greater effort visually reviewing the data before it enters the database. All artifacts must be detected, and no error can be tolerated. With millions of samples to analyze every year, this has become a formidable task.

We have developed software analysis methods that can automate this data review and potentially eliminate 90% of the work. Our fully automated TrueAllele™ system inputs raw fluorescent DNA sequencer gel files, processes the gel image (separating colors, tracking and sizing lanes), and analyzes the STR experiments (quantitating and sizing peaks, comparing with ladder peaks, calling alleles). For each allele call, TrueAllele assigns a quality score and applies artifact detection rules. These quality checks enable a user to focus on just the 5%-10% of suspect data, thereby eliminating most of the review effort.

We are currently developing more powerful extensions to TrueAllele for advanced forensic processing. TrueAllele can already read data from any DNA sequencer, and process it on any computer. Our immediate goal is to have TrueAllele replicate much of the reasoning of the forensic database analyst, and present its focused conclusions visually, rapidly and intuitively. Longer-term, we expect TrueAllele to develop into an intelligent casework assistant.

Introduction

Assuring the quality of short tandem repeat (STR) (1-4) forensic databases (5, 6) is a demanding, labor-intensive task. For each deoxyribonucleic acid (DNA) sample, an analyst must visually review the STR data, identify artifacts, provide correct designations, and decide whether or not to record it in the database. This repetitive task requires time, patience, and extensive knowledge of the allelic behavior at each STR locus. However, much of this decision making may be automatable by a knowledge-based computer program. Coupling a human data reviewer with such a software assistant would greatly increase both data throughput and database quality.

We have developed TrueAllele™, an expert system computer program for STR data analysis (7). TrueAllele automates most of the data review process, transforming raw DNA sequencer files into quality-checked allele designations. By focusing an analyst's attention on just the 5%-10% of problematic designations, TrueAllele can eliminate most of the human data review.

This overview paper is organized as follows:
- System. How TrueAllele transforms data from sequencer files to quality-checked designations.
- Knowledge. Some STR behavior that TrueAllele uses in designating alleles.
- Example. A typical session of TrueAllele processing and user review.
- Software. The platform-independent TrueAllele program, user support, and processing results.
- Conclusion. The implications of TrueAllele for forensic databasing, and future research directions.

System

TrueAllele is a flexible automated genotyping system that accurately sizes and quantitates DNA fragment data (Figure 1). The TrueAllele software can read data formats from both capillary and gel electrophoresis automated fluorescent DNA sequencers (8). The program runs on Macintosh, Windows, and Unix computers.

* Address for correspondence and reprints: Dr. Mark W. Perlin, Cybergenetics, 160 N Craig St, Ste 210, Pittsburgh, PA 15213, USA, http://www.cybgen.com, Email: perlin@cybgen.com

TrueAllele processing begins with an initial image and/or signal analysis that builds a set of sized electropherogram traces. On capillary data, TrueAllele conducts this analysis separately for each capillary.

- Acquire data. TrueAllele reads in the raw data from sequencer files.
- Process signal. The program performs useful signal or image processing, such as removing the baseline, filtering out noise, or eliminating specific signal artifacts.
- Separate colors. TrueAllele automatically resolves the observed fluorescent data into their component dyes. By determining the color separation matrix directly from the run data (e.g., for each capillary), TrueAllele eliminates the need for prior dye calibration runs.
- Remove primers. The software strips the primer region from the allele data.
- Track sizes/lanes. TrueAllele matches the observed size standard peaks (e.g., in the red dye signal) to their expected sizes. With gel data, TrueAllele simultaneously tracks the lanes.
- Extract profiles. Using the size/lane tracking results, TrueAllele extracts a set of one dimensional (1D) electropherogram profiles, indexed by lane/capillary and color.

The extracted 1D profiles are mathematically transformed into a uniform size coordinate system. Since the new domain of each profile is calibrated in base pairs (bps), and not in pixels (e.g., scan lines), TrueAllele can more accurately quantitate the peaks, check for artifacts, and designate alleles.

TrueAllele designates alleles using floating windows by comparing quantitated peaks against allelic ladders.

- Quantitate peaks. TrueAllele models each data peak as a Normal-Cauchy function, and forms a best least-squared fit of the electropherogram data to a parameterized sum of model peaks (9). This approach accounts for band overlap, and robustly estimates each peak's relative DNA concentration from the modeled peaks (using height or area).
- Derive ladders. For each locus, the program matches the observed allelic ladder data to the expected ladder sizes. It then forms a virtual ladder to account for designations not in the ladder data.
- Designate alleles. For each polymerase chain reaction (PCR) experiment, TrueAllele determines which candidate peaks to retain. To designate the alleles, the program matches these peaks against the ladder's floating window. With gel data, each gel loading has a separate floating window.

TrueAllele checks the quality of every allele designation by (a) applying rules that detect data artifacts or potential scoring problems and (b) heuristically ranking the remaining data on a scale of 0 (worst) to 1 (best).

(a) TrueAllele's *rule architecture* is quite flexible, and programmers can rapidly modify existing rules or add new ones. Typical rules check for signal strength, dispersed peaks, sizing deviations, and crosstalk artifacts. Alleles that trigger rule firings are presented for visual review to a human analyst.

(b) TrueAllele's *quality measure* enables an analyst to focus on the problematic data. After reviewing rule firings, the human analyst then checks those designations which have low scores. High scoring designations (the vast majority) have a low probability of error, and do not require human review (10).

TrueAllele makes it processing results available in several formats.

- User Interface. The *AlleleView* navigator interface shows TrueAllele's designations in the context of the appropriate allelic ladder (Figure 2). *AlleleView* uses rule firings and quality scores to help focus the analyst on problematic data. Other visual interfaces (*PrepView*, *ImageView*, *MarkerView*, *SizeStdView*, etc.) provide useful data interactions.
- Computer Output. TrueAllele can generate tabbed text results in diverse formats useful for downstream processing. Programmers can customize these results to include many variables for each allele designation, including internal quality measurements.
- Database Entry. TrueAllele's tabbed text output can be routed to a forensic database.

Knowledge

TrueAllele applies considerable knowledge of STR systems to its processing and presentation of genotyping data. A few examples are given here.

Floating windows. DNA sequencers can size inconsistently, due to differential migration of the STR PCR products relative to the size standards. This deviation is most evident when comparing different brands or models, but even occurs with runs on the same machine. Therefore, software analysis methods based on fixed size criteria for allele binning can sometimes lead to nonoptimal allele designations (e.g., PE Biosystems Genotyper).

To correct for this error, forensic scientists have long used allelic ladders as within-run calibration data (11, 12) . When sample peaks are analyzed in the context of allelic ladders (e.g., peak center ± 0.5 bps), comparisons using *floating windows* can more accurately designate the alleles. This comparison can be done visually, or in software (e.g., Forensic Science Service ALLEATOR).

TrueAllele uses such floating windows when analyzing forensic STR data. For every locus, the program sizes each ladder and sample lane relative to its internal lane size standards. TrueAllele then matches the ladder peaks to their designations; even and odd gel loadings are processed separately. The program compares a sample's peaks with the floating allele windows of the appropriate ladder. Large deviations may reduce TrueAllele's quality score for the designation. TrueAllele's *AlleleView* user interface displays all forensic data in the context of the allelic ladder, making the floating window visually apparent.

Stutter artifact. PCR through an STR's tandem repeat region can skip a repeat unit and synthesize a differently sized fragment (13). After many cycles, the PCR then generates a stutter (or "shadow band") artifact comprised of multiple allele sizes. Calibration of this reproducible artifact for a given locus under fixed PCR conditions can mathematically eliminate the artifact (7). Stutter is generally more pronounced with shorter repeat units (e.g., dinucleotide repeats) than with the longer units (e.g., tetranucleotide repeats) used in forensic science.

TrueAllele can automatically detect, calibrate, and remove PCR stutter artifact from STR data. TrueAllele does this by (optionally) building a table of locus stutter patterns for each allele, and using this table to deconvolve the stutter from the observed signal (14). The deconvolved signal can more accurately describe the DNA components present in the original sample.

Pref amp. Preferential amplification occurs in STR systems when some fragment sizes (e.g., shorter ones) PCR more efficiently than others. The pref amp artifact may mask small allelic peaks, and cause heterozygotic genotypes to appear homozygotic. Forensic analysts use preset allele ratios to reproducibly designate pref amp alleles.

TrueAllele can automatically calibrate pref amp for each locus, and then (optionally) use this information in making allele designations. When excessive deviation in allele quantitation ratios triggers a rule, TrueAllele can report this potential designation problem to the analyst.

Rule system. TrueAllele has an extensible rule system for representing knowledge of STR artifacts. These rules detect potentially incorrect allele designations, and can report them to the analyst. In processing genotype data, TrueAllele records dozens of variables for each genotype that measure the deviation of observed data from expected behavior. A given rule can use any of these variables, drawn from one or more genotypes. For example, lane-to-lane artifacts require information from two adjacent genotypes.

After designating the alleles, TrueAllele applies all rules to every genotype and records the rule firings. When an analyst reviews the annotated results, *AlleleView* can list the rule firings for each problematic genotype. This information directs the analyst's attention to that genotype's specific artifacts.

Quality ranking. The vast majority of allele designations do not fire rules. TrueAllele ranks these good data according to a heuristic that incorporates the most important "quality" features. For example, an informative forensic measure would include peak height and floating window deviations. TrueAllele's *AlleleView* interface presents this good data set in a worst-first ordering, so that the analyst's attention is centered on the problematic data.

Adaptability is a key feature of TrueAllele's knowledge base. TrueAllele's use of flexible rule encodings and adjustable quality measures enables customization of the system over time in an ever-changing STR technology environment.

Example

A tutorial TrueAllele evaluation package can be downloaded from the Cybergenetics web site (http://www.cybgen.com). Example forensic data and forensic panel templates are also available. A detailed visual presentation (over fifty screen snapshots) is provided in the Tutorial user documentation. The example presented here uses processed forensic data from the web site, with ABI/377 electrophoresis (15) of an SGM+ panel.

DataDisk Setup. TrueAllele organizes gel (or capillary) run data on a *DataDisk*. The DataDisk has a "data" folder, which contains a set of gels and their annotations. There is also a "common" folder, which contains reference files for panels, locus information, dyes, and size standards; we provide a reusable template DataDisk for the SGM+ forensic panel. A DataDisk can be set up manually (e.g., using Finder and Excel on the Macintosh), or automatically from a laboratory database.

Signal/Image Processing. After setting up a DataDisk with 5-10 gels, the analyst starts TrueAllele and loads the DataDisk. In a typical high-throughput installation, the analyst then runs *ImageCall*, which (a) performs signal and image analysis, and (b) tracks the vertical lanes and the horizontal size standards simultaneously. The result of this two dimensional (2D) tracking is a 2D grid that the *ImageView* interface can visually superimpose on the data. After inspecting (and possibly editing) the 2D tracking grid, TrueAllele extracts 1D electropherograms, and is ready to designate alleles.

Allele Calling. The *AlleleCall* program automatically processes the genotyping data across all the gels. *AlleleCall* quantitates peaks, derives allelic ladders, and designates alleles. The program then applies rules, and assigns a quality measure to every designated allele.

Quality Checking. The analyst can check TrueAllele's designation results in the *AlleleView* interface (Figure 2). Reviewing the genotypes of one locus across all the gels helps the analyst concentrate on locus-dependent features. The analyst starts by reviewing those genotypes which triggered a rule (i.e., bad data); each genotype's fired rules are listed in a pop-up menu. Editing can then continue on the worst data first, since the genotypes are sorted by quality score. With reasonably good STR data, an analyst views the questionable 10% of designated genotypes before moving on to the next locus.

The *AlleleView* program provides additional interfaces (*Electropherogram*, *Quantitation*, *Genotype*, *Lanes*, *Family*, etc.) that the analyst can use to inspect specific artifacts. For example, possible dye bleedthrough is best examined in the *AlleleView Electropherogram* window, which automatically displays every signal in a particular locus size range in its own dye color.

Output Formats. The analyst can generate results in many different arrangements (e.g., ordered by locus or by sample). The results are produced in a flexible tabbed text format that can be input to other computer programs, submitted to a database, textually examined in a spreadsheet, or visually presented in *AlleleView*. Since TrueAllele keeps an audit trail of the dozens of variables (observed and expected behavior) considered in designating each allele, any subset of this information can be programmed into the output files.

Software

TrueAllele is written in the MATLAB 5.2 visualization and numerical programming language; this enables rapid development and deployment of signal processing algorithms. MATLAB (hence TrueAllele) is cross-platform, running on Macintosh, Windows, and Unix computers.

The TrueAllele development process includes a version control system (CVS on Unix). The Cybergenetics web site (www.cybgen.com) provides a form for user feedback (bug reports, new features, etc.), and visitors can download evaluation software, documentation, and other resources.

Automated Unix scripts are used to test the software, and compile it into pcode files. The scripts then assemble the RunTime software for all three computer platforms, with packaging for different user needs (e.g., tutorial, program, update). TrueAllele documentation is distributed as bookmarked PDF files, generated using Abobe FrameMaker.

We tested TrueAllele on a suite of ten SGM+ ABI/377 gels. We ran TrueAllele's *SizeStdView* interface on the first gel to calibrate the GS500 size standards to the electrophoresis run conditions. We ran TrueAllele on a 266 MHz iMac computer with 160 MB total RAM. We observed the following results:

- *Signal/Image Processing*. After automatically tracking lanes and sizes, half the gels did not require grid editing. The other five gel grids were each edited in under five minutes.
- *Allele Calling*. TrueAllele designated all the samples in the ten gels for the SGM+ panel in an overnight run on the Macintosh computer.
- *Quality Checking*. TrueAllele's designations were reviewed by a human operator using the *AlleleView* navigator. The analyst detected all the miscalls either in the rule firings, or in checking the low scoring data. No designation errors were found in the (vast majority) of remaining "good" data.

Conclusion

Manual review of forensic data is a critical bottleneck in the construction of criminal offender databases. This labor-intensive task consumes considerable resources (time, people, cost, error, effort) that might be better applied elsewhere in the criminal justice system. We have developed a knowledge-based computer program for STR analysis that automates most of this human review process. Our TrueAllele program works with most DNA sequencers, and runs on most computers.

We are continuing to develop and refine the TrueAllele software. We have recently incorporated new modules for capillary processing and 96-lane tracking. We have developed new data representations that permit much faster processing of STR sizing data and more customizable display of genotyping results. These improvements in performance and adaptability are essential if software automation is to keep pace with the continuing evolution of STR technology. We expect that this improved TrueAllele will be a useful starting point for developing an intelligent casework software assistant.

Acknowledgments

The STR data were provided by Richard Pinchin and Declan O'Grady of the Forensic Science Service. TrueAllele programming and testing were done by Meredith A. Clarke and Michael Breen of Cybergenetics. This research was supported in part by SBIR Phase II grant award 2R44 HG01568-03 from the National Institutes of Health.

References

[1] 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.

[2] Fregeau CJ, Fourney RM (1993). DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification. *Biotechniques*, **15**(1)**:** 100-119.

[3] Kimpton CP, Gill P, Walton A, Urquhart A, Millican ES, Adams M (1993). Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Meth. Appl.*, **3:** 13-22.

[4] Urquhart A, Kimpton CP, Downes TJ, Gill P (1994). Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers. *Int. J. Leg. Med.*, **107:** 13-20.

[5] McEwen JE (1995). Forensic DNA data banking by state crime laboratories. *Am. J. Hum. Genet.*, **56:** 1487-1492.

[6] Gill P, Urquhart A, Millican ES, Oldroyd NJ, Watson S, Sparkes R, Kimpton CP (1996). Criminal intelligence databases and interpretation of STRs. *Advances in Forensic Haemogenetics*, **6:** 235-242.

[7] Perlin MW, Lancia G, Ng S-K (1995). Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. *Am. J. Hum. Genet.*, **57**(5)**:** 1199-1210.

[8] Ziegle JS, Su Y, Corcoran KP, Nie L, Mayrand PE, Hoff LB, McBride LJ, Kronick MN, Diehl SR (1992). Application of automated DNA sizing technology for genotyping microsatellite loci. *Genomics*, **14:** 1026-1031.

[9] Richards DR, Perlin MW (1995). Quantitative analysis of gel electrophoresis data for automated genotyping applications (Abstract). *Amer. J. Hum. Genet.*, **57**(4 Supplement)**:** A26.

[10] Pálsson B, Pálsson F, Perlin M, Gubjartsson H, Stefánsson K, Gulcher J (1999). Using quality measures to facilitate allele calling in high-throughput genotyping. *Genome Research*, **to appear**.

[11] Puers C, Hammond H, Jin L, Caskey C, Schumm J (1993). Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01[AATG]n and reassignment of alleles in population analysis by using a locus-specific allelic ladder. *Am. J. Hum. Genet.*, **53**(4)**:** 953-8.

[12] Griffiths RAL, Barber MD, Johnson PE, Gillbard SM, Haywood MD, Smith CD, Arnold J, Burke T, Urquhart A, Gill P (1998). New reference allelic ladders to improve allelic designation in a multiplex STR system. *Int. J. Legal Med.*, **111**(5)**:** 267-272.

[13] Hauge XY, Litt M (1993). A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Hum. Molec. Genet.*, **2**(4)**:** 411-415.

[14] Perlin MW (1999). Method and system for genotyping, U.S. Patent, #5,876,933

[15] Frazier RRE, Millican ES, Watson SK, Oldroyd NJ, Sparkes RL, Taylor KM, Panchal S, Bark L, Kimpton CP, Gill PD (1996). Validation of the Applied Biosystems Prism™ 377 automated sequencer for forensic short tandem repeat analysis. *Electrophoresis*, **17:** 1550-1552.

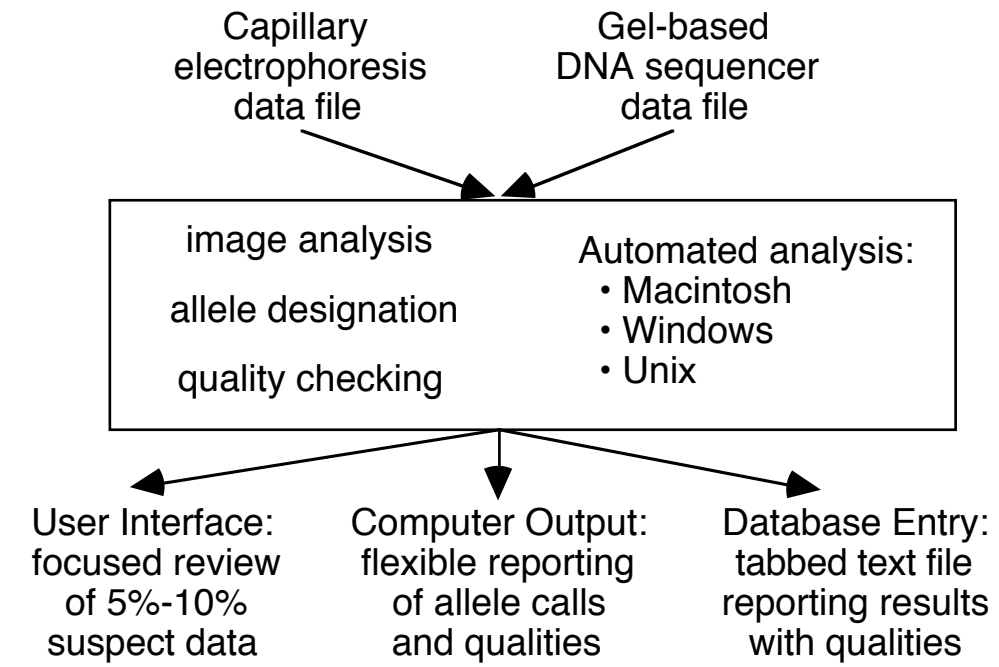

Figure 1. TrueAllele is a flexible automated genotyping system. The program can input data from capillary or gel DNA sequencers, and process these data on most computer platforms. Processing entails image/signal analysis, allele designation, and quality checking. The output can be visually presented for user review, or in file formats suitable for downstream computer analysis or database entry.

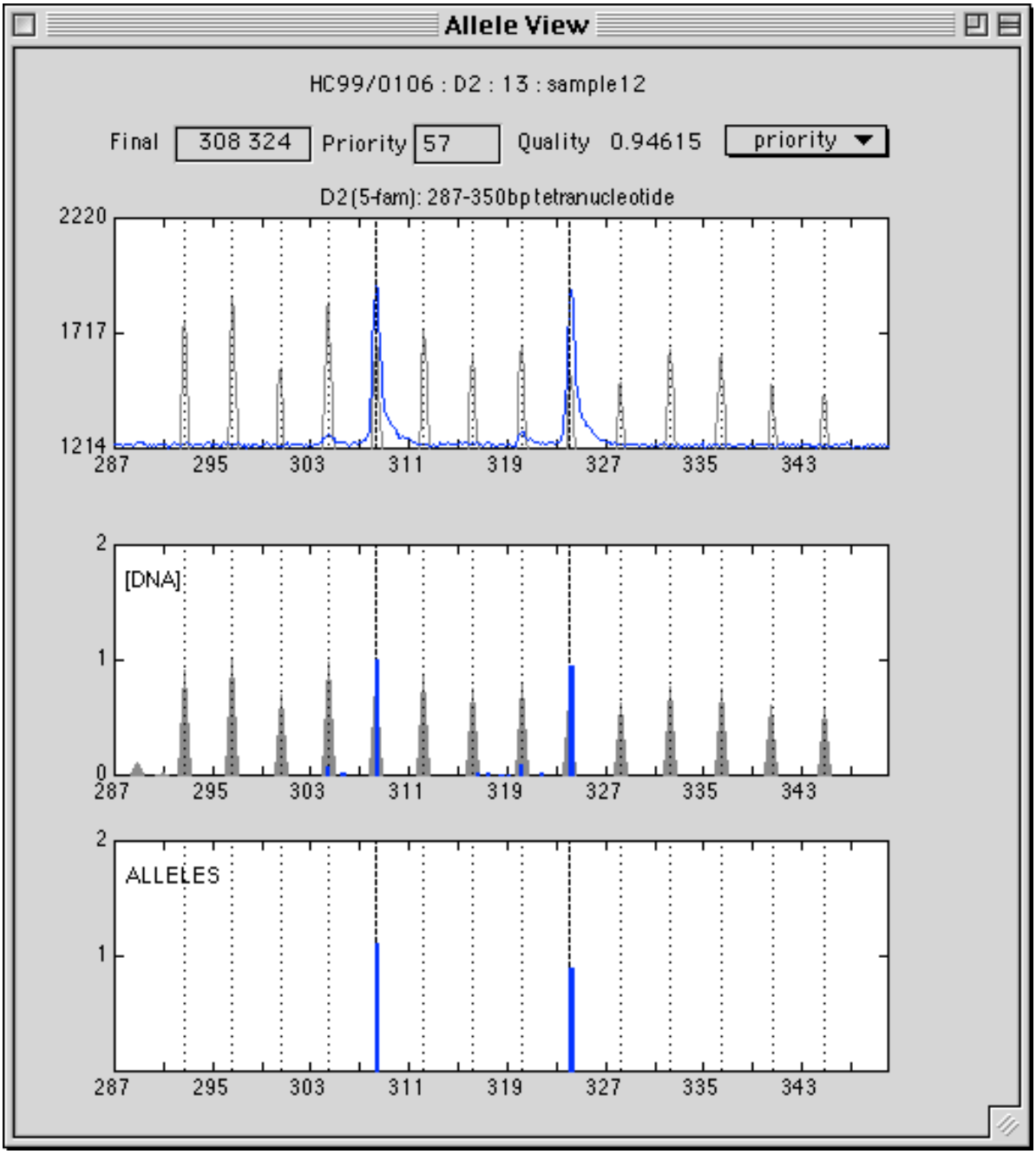

Figure 2. The AlleleView navigator program automatically displays allele designation information. Shown here is a D2 designation of (308, 324) for a sample. The top text region provides navigation data, including the gel, locus, lane, sample, designation, review priority, quality score, rule firings, dye and size range. The first electropherogram pane shows the D2 sample signal superimposed on the odd D2 allelic ladder signal. The second quantitation pane shows the results of DNA quantitation and peak sizing. The third designation pane shows the final (editable) allele designations. The allelic ladder sizes are drawn as dashed vertical lines throughout. Each pane can be opened to reveal additional relevant visual information.