John H. Leamon, Andrew Moiseff and Joseph F. Crivello
University of Connecticut, Storrs, CT, USA

ABSTRACT

A method is described that uses the ABI PRISM™ 310 genetic analyzer in conjunction with custom-designed software to identify and classify RAPD products. This methodology will also work well with AFLPs and microsatellite analyses. The methodology uses the ABI PRISM 310’s high-throughput (>500 samples per week) capabilities and in-lane molecular weight standards to efficiently separate and size DNA products. Peak detection, locus classification and export of the data in a form accessible by several genetic analysis programs were accomplished through a custom-written software program (Peaks®). Various criteria used by the program to identify and classify loci are described, and their effect on population analyses is examined. Criteria providing an effective, robust determination of population structure are presented.

INTRODUCTION

Gene flow assessment and quantification have become increasingly important to the members of aquaculture, fisheries management and agricultural biotechnology communities. Gene flow assessments are commonly used to determine whether populations of game fish are reproductively distinct (27), to quantify the genetic effects of farmed fish on wild stocks (6) or to estimate the rate of hybridization of genetically manipulated crops with native strains (20).

RAPD PCR has been shown to provide the basic information necessary to distinguish population structure and determine relative amounts of gene flow (18). The greatest drawback to the use of RAPDs has been the difficulty in reproducing conditions from one laboratory to another. Nonetheless, this approach has been successfully used with representatives from all kingdoms and with genomes of highly varying GC content. Products that vary between samples may result from genomic differences accumulated over evolutionary history or from mutations occurring within the life span of the organism. Mendelian inheritance of these products has been demonstrated for many organisms, including humans (21), insects (7), plants (14) and several species of fish (4,5,10,16). These polymorphisms serve as dominant genetic markers and represent sequences from both coding and noncoding regions of the template genome.

Because RAPD priming sites are randomly scattered throughout the template genome, they may be very accurate indicators of genetic distance (19). RAPDs have been routinely used in fish to differentiate species (13), subspecies (3), feral full-sibling families (16) and laboratory-reared strains (8). These markers have successfully detected polymorphisms in species that exhibit low levels of allozyme variation, such as Salvelinus namaycush (lake trout) (16), as well as blue Pinnace glauca and mako Isurus oxyrinchus sharks (preliminary data).

Since the DNA products produced by RAPDs are random, they require no prior knowledge of the genotype sequence, and simple trial and error can be used to determine the primers. Thus, time spent on protocol optimization and sequence determination is kept to a minimum. This advantage is not unique to RAPDs; other methods such as RFLP analysis do not require the determination of specific sequences. However, RAPDs are faster and cheaper than RFLP analyses and can obtain much higher throughput. With the benefits of high resolution, low cost and high throughput provided by RAPD analysis, more samples can be analyzed than with any other method.

Capillary electrophoresis of fluorescent RAPDs provides automated analysis with high throughput, low DNA requirements and a high degree of accuracy. This improves upon the more labor-intensive analysis of RAPDs that uses ethidium bromide or silver staining to reveal products that are separated by agarose or acrylamide electrophoresis and sized by visual comparison with...
known molecular weight standards. Automated detection of PCR products via fluorescently labeled primers or dNTPs has been shown to be effective for polymorphism detection (2) and resolution of genetic difference (13). However, both of these studies used the ABI PRISM™ 373 DNA sequencer rather than the ABI PRISM 310 genetic analyzer (both from PE Biosystems, Foster City, CA, USA) and analyzed low numbers (sample size of 4 and 8, respectively) of samples. Because of the low number of samples analyzed in these studies, manual techniques were sufficient to analyze the data. As the size and scope of population studies increase, automated detection and analysis become an increasingly attractive alternative in terms of accuracy, reproducibility and efficiency.

Here, we describe a high-throughput (>500 samples/week) process for detecting and screening genetic polymorphisms using fluorescently labeled primers, the ABI PRISM 310 genetic analyzer and novel data extraction protocols.

**METHODS AND MATERIALS**

**Genomic DNA Isolation**

*Fundulus heteroclitus macrolepidotus* (the common mummichog) were collected along the southeastern coast of Connecticut. From a total of 13 sub-populations located within three greater populations, 651 individuals were obtained and frozen at -80°C. Genomic DNA was isolated from frozen tissue samples using a modification of the Cockburn and Seawright protocol (1). A working concentration of 2.5 ng DNA/µL was prepared for each sample.

**RAPD PCR and Gel Conditions**

The RAPD PCR was performed under the following conditions: 10 mM Tris (pH 8.3), 50 mM KCl, 2.5 mM MgCl₂, 0.8 µM primer, 0.2 mM dNTPs, 0.1 U Stoffel fragment DNA polymerase (PE Biosystems) and 25 ng genomic DNA in a Perkin-Elmer 9600 thermal cycler (9). The samples were cycled using a ramped annealing temperature; the reaction profile is detailed in Table 1.

Three separate reactions were performed for each sample, with one of three fluorescently labeled primers. When completed, the reactions from each sample were combined and diluted fivefold. Five microliters of the product mixture were then combined with 0.5 µL (4 fmol) of fluorescent size standard [GeneScan®-2500 TAMRA (Carboxytetramethyl-rhodamine)] (PE Biosystems) and 19 µL of double-diluted water and analyzed on an ABI PRISM 310 genetic analyzer. Samples were run under native conditions prescribed for low-resolution separation of large molecular weight products (26) in a 96-well format. Both the cathode and anode solutions contained 2.5% GeneScan polymer, and the run was conducted with a 31-mm length to detector (41-mm total length) capillary tube at 14 kV at 30°C for 15 min. Data were collected using GeneScan software (version 2.1) and saved as a separate sample file for each reaction. The GeneScan software was used to create an individual size standard for each file.

**Data Extraction, Peak Detection and Comparison**

Using the Ripper® software (25) based on the ABI toolkit generously supplied by PE Biosystems, each sample file was converted into a text matrix file. These files were then used by a second program, Peaks® (24), which detected peak locations based on the transition of the derivative from a positive to a negative number. This transition was taken to define a local maximum (a peak); the following derivative transition from negative back to positive defined a local minimum (a trough). A valid peak was detected when the difference between the local maximum and local minimum exceeded a designated threshold. The threshold value is a dynamic parameter and changes in response to a user-defined sensitivity setting. The sensitivity determines the minimum number of peaks that must be identified in the products generated by each primer. Peaks that met the threshold and sensitivity criteria were identified by molecular weight and stored in individual matrix files for later comparison with the total sample pool (Figure 1).

Each peak (locus) indicated the presence of a PCR product of a specific molecular weight. All loci produced by the same primer with identical molecular weights were considered to be copies of the same loci, while loci differing in molecular weight were considered separate loci. However, minor variations resulting from variability in sequencing analysis (such as PCR amplification and standards calibration) required a more robust procedure.

---

Table 1. Thermal Cycles Used for Arbitrarily Primed PCR Amplification of RAPD

<table>
<thead>
<tr>
<th>Number of Cycles</th>
<th>Denaturation Phase</th>
<th>Annealing Phase</th>
<th>Elongation Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>60 s at 94°C</td>
<td>60 s at 35°C</td>
<td>45 s at 72°C</td>
</tr>
<tr>
<td>5</td>
<td>60 s at 94°C</td>
<td>60 s at 37°C</td>
<td>45 s at 72°C</td>
</tr>
<tr>
<td>5</td>
<td>60 s at 94°C</td>
<td>60 s at 39°C</td>
<td>45 s at 72°C</td>
</tr>
<tr>
<td>5</td>
<td>60 s at 94°C</td>
<td>60 s at 41°C</td>
<td>45 s at 72°C</td>
</tr>
<tr>
<td>5</td>
<td>60 s at 94°C</td>
<td>60 s at 43°C</td>
<td>45 s at 72°C</td>
</tr>
<tr>
<td>5</td>
<td>60 s at 94°C</td>
<td>60 s at 45°C</td>
<td>45 s at 72°C</td>
</tr>
<tr>
<td>5</td>
<td>60 s at 94°C</td>
<td>60 s at 47°C</td>
<td>45 s at 72°C</td>
</tr>
</tbody>
</table>

Reactions conducted in 20 µL reaction mixture in a Perkin-Elmer 9600 thermocycler.
After being processed, each sample was characterized by a binary string that indicated the presence or absence of loci identified within the population. These binary strings were combined and outputted as a single text file. In this format, the binary strings can be imported into many available programs used for analysis population genetics, including POPGENE 32 (28), Genetic Data Analysis (22), and Tools for Population Genetics (TFGPA) (23). For studies with high sample numbers, large numbers of loci and considerable population structure (populations, subpopulations and sub-subpopulations), our laboratory has found TFGPA extremely useful.

Data Analysis

The data file that contained the binary strings for each sample was then imported into the TFGPA software package (23) that was used for the remaining analysis. The threshold, sensitivity and resolution of the Peaks program were modified to determine its effects on the analysis results. Strings were created using non-dynamic threshold settings ranging from 25–500, and the number of loci recorded were compared by one-way ANOVA with the Student-Newman-Keuls method for pairwise multiple comparison.

In other tests, the initial dynamic threshold was set at 2500 U while the sensitivity was set at 5, 10, 15 or 30. For each sensitivity setting, binary strings were created with the resolution set at 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90 and 100. Number of loci, as well as values for \( \theta_P \) and \( \theta_S \), were derived for the entire data set under each of these parameter combinations and statistically compared by ANOVA. Similar tests were also run to examine the effect of varying the sensitivity on \( \theta \) values.

RESULTS

Of the original 651 individuals, 36 samples were discarded because of poor data quality or sequencer malfunction. A molecular weight calibration curve was calculated separately for each sample. Each of these curves effectively calibrated scan position to molecular weight (mean \( r^2 \) value for all 615 standards was 0.995, with an SD of 0.003). Within each sample, products were sized according to the standard curve generated for the respective file.

Loci detected by RAPD were analyzed as alleles under four assumptions. First, RAPD products segregate in a Mendelian fashion as dominant alleles. Second, all of the genotype frequencies at the loci are in Hardy-Weinberg proportions. Third, all of the homozygous
recessive alleles in an individual are considered to be from identical mutations. Fourth, dominant alleles have also arisen from identical mutations.

Signal threshold, sensitivity and resolution were used as criteria for extracting data. For a given sample, the number of peaks ranged from less than 100 using stringent criteria to an excess of 3000 when using the least stringent criteria. To address relevance of added loci in discerning population structure, TFPGA was used to calculate Weir and Cockerham’s θ values derived from the same data set under different resolution and sensitivity settings (17). F-statistics, such as Weir and Cockerham’s (17) θ values, appraise population structure by comparing allele frequencies observed in the population. Similarly, θ values estimate the correlation of allele pairs between individuals within a subpopulation through analysis of allele frequency variance partitioning into population (θP) and subpopulation (θS) components.

Differences between the θ values generated from binary strings using different sensitivities were assessed by ANOVA and found to be significant for θS (P<0.001). The highest θS or θP values were produced with a sensitivity of 30; pairwise comparisons were significant for all combinations of sensitivities except those between 10 and 15. The θ values produced by the different sensitivities were not significantly different for θP (P>0.05).

Resolution was a significant source of variance in determining θS or θP values (P<0.0001 for both values). The vast majority of the θS or θP values derived from identical sensitivities but differing resolutions were significantly different from each other (Figure 2). While values for θP differed only slightly with varied resolution and sensitivity combinations, more dramatic differences were seen between θS values. The local maximum for θS (0.1440) was generated at a sensitivity of 30 and a resolution of 40 bp, and identified 167 loci. The global maximum for θS (0.1669) occurred at a sensitivity of 30 and a resolution of 100 bp and produced 88 loci. At each of the sensitivities examined, a decrease in resolution resulted in an initial, dramatic increase in θS values. The rate of increase slowed after a resolution of approximately 40 bp had been reached, with subsequent decreases in resolution producing θS values that loosely vacillated around those obtained with the 40-bp resolution. The 40-bp resolution produced a local maximum when sensitivity was set at 30, and the global maximum among θS values produced with a sensitivity of 15 (0.1329 and an SD of 0.0063). The numbers of loci generated by the two methods were very similar (167 for the 30 peaks, 166 for the 15 peaks).

As loci number increased, θP rose rapidly and then achieved a maximum value at samples with 88, 96 and 167 loci (Figure 3). The θP then entered a sustained asymptotic decline as loci number increased. The θS value did not increase dramatically during the initial increase in loci number, although it did exhibit a similar sustained decrease as loci number increased (Figure 4).

The frequency of dominant loci (where a band is detected) that were generated under different resolving powers exhibited informative trends (Figure 5). For the two sensitivities examined [5 and 30], a high resolution (1 bp) resulted in extremely low average dominant frequencies (0.76% and 0.19%, respectively), and a high number of loci were detected (1891 and 3284, respectively). When resolution was set near the middle (40 bp) of the total range examined (1–100 bp), the average dominant frequency reached 24.8% with 128 loci at a sensitivity of 5,
and 34.5% with 167 loci at a sensitivity of 30. At extremely low resolution (100 bp), the frequency climbs to 34.9% with 64 loci and 39.2% with 88 loci for the respective sensitivity settings.

Based on the data, a sensitivity of 30 and a 40-bp resolution were chosen as optimal settings for data generation and used to create the binary strings used for the remainder of the analysis. This resulted in a 167 loci character string for each of the 615 individuals used in the genetic analysis.

DISCUSSION

Detection of polymorphisms from a broader spectrum of the genome increases the precision with which population substructure can be discerned. As sensitivity increases, the concentration of product required for detection decreases. The ABI PRISM 310 genetic analyzer resolves labeled products at picomolar concentrations. As detectable concentrations become smaller, they approach baseline values. This makes detection of fluorescently labeled PCR product (signal) from the background (noise) increasingly problematic and emphasizes the need to differentiate between baseline noise and labeled product using objective criteria.

Several different strategies can be used to extract signal from the noise. One common method attempts to normalize all signals to a standard intensity (such as the highest signal within the sample). Unfortunately, normalization of a low signal-to-noise data set increases the noise as well as the signal. Another method used to separate signal from noise relies on a predetermined threshold value. However, the magnitude of the threshold depends on the signal and noise characteristics of the data set, and the proper threshold for one data set may not be appropriate for another. For example, a high threshold required for identification of a signal in a data set with high signal and noise might exclude all data in a set characterized by low signal and noise.

The Peaks program circumvents these problems by combining a dynamic threshold with a sensitivity parameter. The software allows the user to assign a minimum intensity threshold that the peaks must exceed to be recorded as loci. Based on analyses of many different threshold values, a minimum threshold of 25 U was empirically chosen. This value was high enough to be insensitive to inherent system electronic noise, yet low enough to detect the presence of minor peaks.

However, with the benefit of added sensitivity, the question arises whether all bands produced by all individuals are equally valid, or if a subset consisting of only highly amplified bands should be analyzed. At the most sensitive software settings and a threshold of 25, each primer produced 28 ± 8 bands per sample. A sensitivity of 30 incorporates almost all peaks present in an individual, while sensitivity settings of 15, 10 and 5 remove low amplification bands with increasing severity.

Discarding low-intensity bands has the potential to lose valuable data and increase the possibility of false negatives. It also reduces the total number of loci studied, thereby reducing the number of locations at which variation can occur. A RAPD study of genetic variation in *Salvelinus alpinus* (arctic char) discarded bands that stained poorly or were deemed inconsistent, which resulted in the loss of approximately 40% of the original bands (12). Four potentially useful markers were identified, but these were unable to discriminate between two suggested subspecies.

**Figure 3. Interaction between $\theta_P$ values and loci number.** Errors indicate sd.

**Figure 4. Interaction between $\theta_S$ values and loci number.** Errors indicate sd.
On the other hand, if all detected products are used, no information is lost and no false negatives result, although the possibilities of false positives increase. The total number of loci increases relative to processed samples and with it, the number of potential polymorphisms. This all-inclusive method for product scoring has been used successfully in the past (15), if robust similarity coefficients that are insensitive to false positives (e.g., Nei and Li’s estimate) were used for analysis.

Accurate assessment of genetic variation requires maximization of loci identification while minimizing sensitivity to system noise and error. A balance must be struck between the system’s potential resolution and robustness. In this case, classification of detected products as either unique loci or alleles of the same locus is dependent on the molecular weight. Thus, resolution depends on the accuracy of locus sizing due to the presence of individual molecular weight markers and the resolution inherent in the system. Robust locus classification avoids overreliance on the internal sizing standards by including cautious confidence limits into the analysis. Our goal is to maximize the benefits of resolving power inherent in the 310 capillary electrophoresis system while accounting for variability and noise.

Standardized molecular weight markers analyzed on the ABI PRISM 310 revealed a high degree of resolution within each run and routinely differentiated between calibrator peaks separated by five bases or less. Agreement between runs was high as well; a comparison of all standard curves in the study resulted in an SD of 0.003. As the maximum size of any fragment analyzed in this study was 1400 bp, the 0.3% deviation translates into a minimum resolvable difference of 4 bp under ideal circumstances. This represents the theoretical resolving power; the actual resolving power is much less since it is degraded by variation induced from other sources. Slight variations in PCR amplification, as well as characteristics inherent in the data, such as variable peak width or noise, can induce errors and variability between runs, even though electronic system noise is extremely low (unpublished data).

Overreliance on the system’s resolution denies the influence of other sources of variability and error and is represented in our system by analysis using high-resolution settings (1–15 bp). When loci are classified at high resolution, large numbers of loci are recorded, of which many are nonexistent, unique to an individual or with a low frequency of dominant occurrence.

In intraspecific comparisons conducted over relatively small geographical scales, a large proportion of bands are often shared between individuals
because of a typically high degree of relatedness. The presence of the shared bands is crucial, as Weir and Cockerham's $\theta$ value employs the correlation between shared alleles and location to estimate population structure (17). When large numbers of low-frequency alleles are introduced, shared alleles represent a substantially lower proportion of the total gene pool and contribute correspondingly less to $\theta$ value. Therefore, a uniformly high level of dissimilarity from each other characterizes individuals and populations. The $\theta_P$ and $\theta_S$ values drop accordingly, indicating low levels of differentiation and high levels of genetic exchange.

As resolution decreased, the frequency of the dominant loci increased dramatically. When taken to extremes, this also results in the loss of informative data. Where excessively high resolution lost data by splitting existing peaks into random noise, excessively low resolution created frame shift errors in which unique loci between individuals were erroneously classified as identical. As the resolution decreased, the number of low frequency alleles decreased as well, and individuals became increasingly similar. This is not reflected by changes in $\theta_P$ and $\theta_S$ values, however, as loci number decreases in a similar fashion. Although the number of low-frequency alleles dropped as they were combined with existing, previously unique loci, a few peaks retained their polymorphic status regardless of the resolution. As loci number dropped, these peaks comprised a larger portion of the overall population, and $\theta_P$ and $\theta_S$ values remained roughly the same.

In this study, we made a decision to use a sensitivity of 30 peaks to include all detectable loci. We chose a resolution of 40 bp to avoid overreliance on the inherent accuracy of instrument-generated sizing algorithms, but still availed ourselves of the reproducibility and superior accuracy that sizing standards confer relative to traditional methods. Additionally, the 40-bp resolution represents the lowest resolution that produces a $\theta_S$ value comparable to the upper limit convergence values seen in Figure 2. The 40-bp resolution represents the point of diminishing returns, where additional decreases in resolution do not produce comparable increases in differentiation. Beyond 40 bp, a 2- to 2.5-fold reduction in resolution is required to achieve statistically greater $\theta_S$ values resulting in a twofold decrease in the number of loci identified. The empirically determined resolution value used in this study was surprisingly close to the 37.5 bp variability experimentally determined for the ABI 373 sequencer and GeneScan software package (11). Admittedly, the experimental work was conducted on an ABI 373 vertical gel system rather than through capillary

![Graphs showing the frequencies of dominant loci at various settings](image)

**Figure 5.** The frequencies of dominant loci at various settings. (A) Sensitivity 5, resolution 1; (B) sensitivity 5, resolution 40; (C) sensitivity 5, resolution 100; (D) sensitivity 30, resolution 1; (E) sensitivity 30, resolution 40; (F) sensitivity 30, resolution 100.
electrophoresis, but the close approximation of the two values is reassuring.

The ABI PRISM 310 genetic analyzer can rapidly and effectively detect genetic polymorphism in large-sample numbers. This methodology is useful for RAPDs, AFLPs and microsatellite analysis. Used in conjunction with the Peaks software program, loci can be identified and classified with high resolution while remaining relatively insensitive to associated error and variance. Analysis of the relationship between threshold, resolution, sensitivity criteria and the resulting $\theta$ values revealed the appropriate settings needed for maximum information. Criteria were chosen that produced the greatest estimation of population differentiation with the smallest number of loci, revealed relatively high band sharing at relatively low resolutions and represented the lowest resolution at which $\theta_P$ and $\theta_S$ values remain nearly constant.

Above all, while manipulation of the sensitivity and resolution can produce statistically different values for $\theta_P$ and $\theta_S$, these differences are not biologically significant. In all cases, the $\theta_S$ value remains below the important 0.05 level at which significant population structure is deemed to occur, while the $\theta_P$ value remains above this level in all but a few cases.

ACKNOWLEDGMENTS

We would like to thank Ali R. Zareh, Product Manager, Genotyping Software, PE Applied Biosystems for the GeneScan source code. This work was done in partial fulfillment of the Doctor of Philosophy Degree requirements for J.L.

REFERENCES


OTHER SOURCES

22. Lewis, P.O. and D. Zaykin. 1999. Genetic data analysis: computer program for the analysis of allelic data. Free program distributed by the authors over the Internet from the home page at http://chee.unm.edu/gda/.
28. Yeh, F.C. 1997. POPGENE, the user-friendly shareware for population genetic analysis. Alberta, Canada, Molecular Biology and Biotechnology Centre, University of Alberta, Canada.

Received 23 August 1999; accepted 10 December 1999.

Address correspondence to:
Dr. Joseph F. Crivello
Department of Physiology and Neurobiology
U-156 Horsebarn Hill Road
University of Connecticut
Storrs, CT 06269
Internet: crivello@predator.pnb.uconn.edu