

David N. Kuhn¹
 James Borrone¹
 Alan W. Meerow²
 Juan C. Motamayor³
 J. Steven Brown²
 Raymond J. Schnell²

¹Department of Biological Sciences, Florida International University

²Plant Sciences and National Germplasm Repository, US Department of Agriculture, Agricultural Research Service, Subtropical Horticulture Research Station, Miami, FL, USA

³MasterfoodsUSA, Mars Inc., Hackettstown, NJ, USA

Single-strand conformation polymorphism analysis of candidate genes for reliable identification of alleles by capillary array electrophoresis

We investigated the reliability of capillary array electrophoresis-single strand conformation polymorphism (CAE-SSCP) to determine if it can be used to identify novel alleles of candidate genes in a germplasm collection. Both strands of three different size fragments (160, 245 and 437 bp) that differed by one or more nucleotides in sequence were analyzed at four different temperatures (18°C, 25°C, 30°C, and 35°C). Mixtures of amplified fragments of either the intron interrupting the C-terminal WRKY domain of the Tc10 locus or the NBS domain of the TcRGH1 locus of *Theobroma cacao* were electroinjected into all 16 capillaries of an ABI 3100 Genetic Analyzer and analyzed three times at each temperature. Multiplexing of samples of different size range is possible, as intermediate and large fragments were analyzed simultaneously in these experiments. A statistical analysis of the means of the fragment mobilities demonstrated that single-stranded conformers of the fragments could be reliably identified by their mobility at all temperatures and size classes. The order of elution of fragments was not consistent over strands or temperatures for the intermediate and large fragments. If samples are only run once at a single temperature, small fragments could be identified from a single strand at a single temperature. A combination of data from both strands of a single run was needed to identify correctly all four of the intermediate fragments and no combination of data from strands or temperatures would allow the correct identification of two large fragments that differed by only a single single-nucleotide polymorphism (SNP) from a single run. Thus, to adequately assess alleles at a candidate gene locus using SSCP on a capillary array, fragments should be ≤ 250 bp, samples should be analyzed at two different temperatures between 18°C and 30°C to reduce the variability introduced by the capillaries, data should be combined from both strands and both temperatures, and undenatured double-stranded (ds)DNA molecular weight standards, such as ROX 2500, should be included as internal standards.

Keywords: Candidate genes / Capillary array electrophoresis / Single-strand conformation polymorphism
 DOI 10.1002/elps.200406106

1 Introduction

Single-nucleotide polymorphisms (SNPs), both as single base substitutions and single base pair insertion/deletions (indels), are the most common sequence differences found between alleles. For example, in the human genome there are approximately one-and-a-half million SNPs (<http://www.ncbi.nlm.nih.gov/SNP>). They occur on average once every 1000 bp, and, although found more

frequently in noncoding regions, are present in coding regions of the genome as well. In plants, SNPs in candidate genes have been found at a frequency of one every 139 bp [1]. The frequency, stability, distribution, and presence within coding regions make SNPs attractive as markers for detecting intraspecific sequence diversity. Potential applications of SNP-based markers include developing saturated genetic maps, mapping ESTs, detecting the genetic associations of phenotypes controlled by multiple loci, studying genetic diversity, and screening for disease susceptibility [2].

Methods have been developed for high-throughput detection of SNPs but these methods require *a priori* knowledge of the SNP being assayed or sequence information surrounding the SNP [3, 4]. Typically, SNP discovery requires either an extensive investment in generating sequence information from genetically unrelated

Correspondence: Dr. David N. Kuhn, Department of Biological Sciences, Florida International University, Miami, FL 33199, USA
E-mail: kuhnd@fiu.edu
Fax: +305-348-1986

Abbreviations: CAE, capillary array electrophoresis; DP, data point; 6FAM, 6-carboxyfluorescein; HEX, 6-hexachlorofluorescein; MW, molecular weight; nt, nucleotide/s; SNP, single-nucleotide polymorphism

individuals, or data-mining sequence information available from genomic and/or EST sequencing projects. This is an effective strategy for organisms with well-characterized genomes, *i.e.*, human and *Arabidopsis*, or when large EST libraries have been created from genetically distinct individuals.

Our research involves tropical plants that are minor crops in the US such as *Theobroma cacao*, which produces cocoa beans used for the production of chocolate. Typically, their genomes are not well characterized, the genetic diversity of collections is based predominantly on phenotypic data, and few or no large, well-defined populations, families, or inbred lines exist. An efficient method for SNP discovery and characterization in such organisms would be one that (i) reliably and reproducibly detects polymorphisms based solely on mobility; (ii) can be accomplished on a high-throughput platform; (iii) can be performed at different temperatures to capture the majority of polymorphisms; and (iv) can automate the identification of alleles from mobility data.

A method readily applied to detect novel polymorphisms without *a priori* knowledge is single-strand conformational polymorphism (SSCP). SSCP is a sensitive, economical procedure that indirectly detects sequence differences to a single base in amplified DNA fragments of the same length [5]. Polymorphisms are detected as alterations of mobility induced by nucleotide (nt) differences that cause stable changes in conformation of the ssDNA. Although the exact identity of the polymorphism cannot be determined, SSCP can be employed to detect and map unknown polymorphisms as they are codominant and can be as robust as other sequence specific markers. Because of their frequency of occurrence, they can be employed for fine mapping [6, 7], candidate gene analysis [8, 9], and estimation of allele frequencies in populations [2]. SSCP has been used successfully in plants to identify sequence polymorphisms without sequencing [10].

Initially developed for polyacrylamide gel electrophoresis, SSCP has been adapted to capillary electrophoresis (CE) [11, 12] and now to capillary array electrophoresis (CAE) [13–16] allowing the high-throughput analysis of hundreds of samples in a short time period. The application of SSCP to CAE has been relatively straightforward with the exceptions of (i) establishing a convenient internal standard to control for capillary-to-capillary and run-to-run variations in allele mobility at different temperatures and (ii) overcoming the limitations of readily available software to allow automated analysis from mobility data alone.

To account for capillary-to-capillary and run-to-run variations in standard fragment analysis performed by CAE, molecular weight (MW) standards of known lengths,

which are labeled with a different fluorescent dye than the sample, are coelectrophoresed with each sample. The mobility of the size standards is used to assign an MW to the sample and this data is directly imported into available software, *e.g.*, Genotyper (Applied Biosystems), for automatic allele determination.

Under the conditions employed for SSCP, the mobility of the ssDNA is not linearly related to the logarithm of its MW, thus the mobility of individual fragments cannot be reliably predicted or linked with size [17]. Several approaches have been developed to account for capillary-to-capillary and run-to-run variations for CE- and CAE-SSCP [18, 19]. Typically, commercially available MW standards are added and denatured in the same manner as the samples. For a single run in one capillary, the scan number is assigned to each MW peak and these values are then used to normalize the mobilities of the samples in all capillaries and runs. In addition, alleles of known sequence composition are often added to each sample [20, 21]. Such internal standards do reduce the amount of variability in mobility across capillaries and runs. However, for loci with large numbers of alleles, allele determination must be done manually based upon pattern differentiation or mobility differences as compared with the internal control, and cannot be readily automated. For loci with a limited number of mutations where all alleles can be run as internal standards, automation of allele calling is possible [21]. Several articles [19, 11, 14] stress the need to vary the temperature of the analysis to be able to detect all alleles because alleles may display identical mobilities at one temperature but different mobilities at another. Since ssDNA mobility is strongly and inconsistently affected by temperature, mobilities for the internal standards are also different at each temperature. Thus, data from samples cannot be readily combined or compared from two different temperatures, unless allelic internal standards are used [21].

The research described here was undertaken to determine if CAE is a reliable and reproducible method to detect sequence polymorphisms in organisms with little available sequence information. We investigated if dsDNA MW standards could be used to reliably and reproducibly assign mobility values to single-stranded sample fragments and allow the alignment of samples at different temperatures. We also investigated the effect of the size of the fragments (160, 245 and 437 nt) on reliably identifying SNPs. The small fragments are a mixture of cloned alleles of the WRKY domain of the Tc10 locus and the medium and large fragments are a mixture of cloned alleles of the NBS domain of the TcRGH1 locus of *T. cacao*. We measured the mobility of the ssDNA of both strands of the three size classes in each capillary of a 16-

capillary array at four different temperatures (18°C, 25°C, 30°C, and 35°C) with three runs at each temperature. We performed a statistical analysis of the mobility data to determine the 95% confidence limits to distinguish alleles. We also estimated the amount of variance due to allele, capillary and run at each temperature. Finally, we investigated under which conditions (size, strand and temperature) all alleles could be identified by a single measurement at a single temperature.

2 Materials and methods

2.1 DNA samples

Three cloned alleles of the WRKY domain for the Tc10 locus of *T. cacao* L. were used to generate the small PCR fragments (1, 2, 3). Four cloned alleles of a portion of the resistance gene homologue RGH1 from *T. cacao* [13] were used to generate the medium (A, B, D, E) and large PCR fragments (F, G, I, K). GenBank accession numbers for each of the cloned alleles are given in Table 1.

2.2 Amplification of cloned alleles

The Tc10 clones were amplified as described in [22], generating the small fragments of 160 bp in length. The RGH1 clones were amplified as described in [13], generating the medium fragments of 245 bp and large fragments of 437 bp in length, respectively. The medium fragments and large fragments were amplified from the same cloned templates. The primers generating the medium fragments were nested within the primer sites of the large fragments. Primers for amplification were labeled differentially to identify each strand and are given in Table 1. The number of SNPs differentiating each PCR fragment is given in Table 2.

2.3 SSCP conditions

Samples were analyzed on an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). All runs were conducted on a 36 cm, 16-capillary array containing 5% GeneScan with 10% glycerol in 1 × Tris-borate-EDTA (TBE) as polymer and 1 × TBE as running buffer. Injection

Table 1. Alleles and primers used to generate fragments

Genbank accession	Alleles	Fragment	Size	PCR product (bp)	PCR primers Name Fluorescent label Sequence
AY331168	WRKY 10.1	1	Small	160	TcWRKY10for
AY331169	WRKY 10.2	2			6FAM CCCTTCACCTAATTGTTTCAGGA
AY331170	WRKY 10.3	3			TcWRKY10rev HEX CCCTCAAATCATGGGATGCT
AF402697	RGH1_3	A	Medium	245	TcRGH1for245
AF402709	RGH1_4	B			HEX GCTGTTGTCTCTCAGACTCC
AF402720	RGH1_2	D			TcRGH1rev245
AF402729	RGH1_6	E			6FAM TGAAGTCGTGTTGTCAGAAG
AF402697	RGH1_3	F	Large	437	TcRGH1for437
AF402709	RGH1_4	G			6FAM CATGGCAAAGAAGTTGGAAAG
AF402720	RGH1_2	I			TcRGH1rev437
AF402729	RGH1_6	K			HEX CATCAATCAATTCCTGTGGC

Table 2. Number and position of SNPs

Size	Frag-ments	SNPs Position and change	Total No. of SNPs
Small	1 and 2	695 (G/A)	1
	1 and 3	700 (C/A)	1
	2 and 3	695 (G/A), 700 (C/A)	2
Medium	A and B	223 (T/C), 241 (G/A)	2
	A and D	223 (T/C)	1
	A and E	175 (G/T), 223 (T/C), 241 (G/A)	3
	B and D	241 (G/A)	1
	B and E	175 (G/T)	1
Large	D and E	175 (G/T), 241 (G/A)	2
	F and G	223 (T/C), 241 (G/A), 322 (C/T), 324 (A/T)	4
	F and I	223 (T/C)	1
	F and K	175 (G/T), 223 (T/C), 241 (G/A), 326 (T/C)	4
	G and I	241 (G/A), 322 (C/T), 324 (A/T)	3
	G and K	175 (G/T), 322 (C/T), 324 (A/T), 326 (T/C)	4
I and K	175 (G/T), 241 (G/A), 326 (T/C)	3	

The positions of the SNPs are indicated from the clones nt sequence and not from the PCR fragments.

was for 22 s at 1 kV and electrophoresis was at a constant voltage of 15 kV. The electrophoretic runs were conducted at four temperatures: 18°C, 25°C, 30°C, and 35°C. These temperatures were chosen as 18°C is the lowest temperature the ABI 3100 can achieve and 30°C was the lowest temperature that the ABI 310 could achieve which allowed us to compare our current results with previous results from the ABI 310. Each amplification product was diluted 1:400 with distilled water, denatured at 95°C for 5 min and snap-cooled on ice. Equal volumes of the amplification products and undenatured ROX 2500 internal standards (Applied Biosystems), also diluted 1:400 in distilled water, were combined in a 1.5 mL Eppendorf tube and gently mixed. Twenty μ L of this mixture was added to each of 16 wells in a 96-well plate so that, for each electrophoretic run, the identical sample was injected onto all 16 capillaries. Initially, the fragment from each allele was run individually at each temperature to determine the mobility and pattern of the individual DNA strands. Based upon these results, the amplification products were combined to give approximately equal intensities of fluorescence into three sample sets: (1) the small fragments (1, 2, and 3) representing the three Tc10 WRKY alleles; (2) the medium fragments A and D and large fragments F and I representing RGH1_2 and RGH1_3 alleles, respectively; and (3) the medium fragments B and E and large frag-

ments G and K representing the RGH1_4 and RGH1_6 alleles, respectively. Each set of samples was electrophoresed three times at each temperature for a total number of 36 runs.

2.4 Data analysis

GeneScan 3.1 (Applied Biosystems) software was used to analyze all runs with the Large Fragment Analysis option enabled. The local Southern method was used for all alignments. The ROX 2500 MW standard peaks were assigned either their actual MW or a pseudoMW to align the data. The actual MW is the size, in bp, of each standard peak, based upon a comparison with the profile of the ROX 2500 standards, run undenatured at 30°C on 3% GeneScan polymer, supplied by the manufacturer (Applied Biosystems, product information). The pseudoMW was generated by the standard method currently used to analyze CE-SSCP (Applied Biosystems, GeneScan product information). A single sample file (data from one capillary) from each temperature was chosen and a value assigned to each molecular weight standard peak based upon the scan number of that peak for that run. These assignments were used as a "size standard" to analyze all the sample files run at the same temperature. Enabling the Large Fragment Analysis Update (Applied Biosystems) allowed the actual scan number to be applied to each peak. The MW standards were divided into two size ranges to analyze the data: (1) the 55–1199 bp range for the small and medium-length PCR fragments, and (2) the 1740–14 097 bp range for the large PCR fragments. In the 55–1199 bp range, MW values were not applied to two standard peaks, the 508 bp and 554 bp peaks, for alignment purposes. The 554 bp standard produces an odd shaped peak with a large shoulder, that, when included as a standard, often resulted in the incorrect assignment of MW to the surrounding standard peaks. The 508 bp standard peak is described as migrating anomalously in the product literature (Applied Biosystems) due to the presence of a hairpin loop (Lisa M. Davis, personal communication). Additional decisions regarding assigning values to the standard peaks based upon the empirical data are described in Section 3.2. The aligned data was imported into Genotyper (Applied Biosystems) to facilitate the production of the data tables. The individual single-stranded product peaks were placed into categories by their order (peak 1, peak 2) and by the dye used to label them (6-hexachloro-fluorescein, HEX, represented as green, 6-carboxy-fluorescein, 6FAM, represented as blue). For each individual product peak three values were collected: the MW, in bp, as determined from the MW of the internal standard peaks, the pseudoMW as determined by assigning the

scan number to the internal standard peaks, and the raw data point (rawDP), the unprocessed scan number for that peak in that particular electrophoretic run. Tables of data were created and exported to Microsoft Excel. Scan number can be converted into retention time using a conversion factor of ~ 6.25 scans per s.

2.5 Statistical analysis

The statistical analysis was conducted by group: small (160 bp), medium (245 bp), large (437 bp) fragments and by temperature. The factors “fragment”, “capillary”, and the “fragment-by-capillary” interaction were considered to be fixed effects while the factors “replication”, “replication-by-capillary”, and “replication-by-fragment” were considered to be random effects. All effects and interactions of interest were determined *a priori*. The statistical model used to analyze each group can be written as follows:

$$Y = \mu + A_i + R_j + C_k + AR_{ij} + RC_{jk} + AC_{kj} + \varepsilon_{ijk}$$

where μ is the experimental grand mean, A_i is the effect of the i^{th} fragment, R_j is the effect of the j^{th} replication, C_k is the effect of the k^{th} capillary, AR_{ij} is the interaction of the i^{th} fragment with the j^{th} replication, RC_{jk} is the interaction of the j^{th} replication with the k^{th} capillary, AC_{kj}

is the interaction of the i^{th} fragment with the k^{th} capillary, and ε_{ijk} is the error term. Proc Mixed of Statistical Analysis Software (SAS) was used with restricted maximum likelihood (REML) fitting of variance components from model-based random effects and least-squares estimation of fixed effects [23]. The least-square means statement was used to obtain estimates of fragment means, standard errors, and confidence intervals. For statistical testing of allele mean comparisons, the Bonferroni adjustment was used to correct the error term for multiple mean comparisons. In cases where data from a single capillary within a run was missing or had obvious outliers, an average of the results from the other capillaries was used in place of outliers or missing data to balance the data set. The complete dataset is available upon request.

3 Results

3.1 Electrophoretic mobility of the undenatured molecular weight standards

Figures 1 and 2 show the relative retention time of the undenatured ROX 2500 internal standards in the two different size ranges (55–1199 bp, Fig. 1 and 1740–

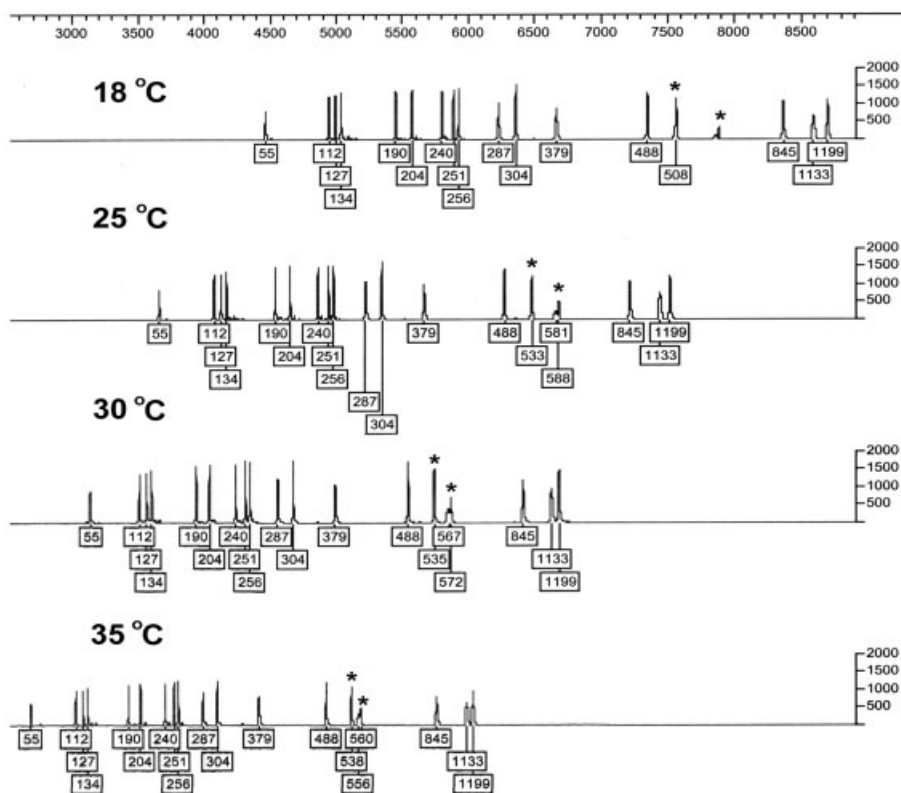


Figure 1. Electropherograms of undenatured ROX 2500 internal standards, 55–1199 bp range. Stars (*) indicate MW standard peaks not assigned values for alignment purposes. The temperature is indicated above the electropherogram. Electropherograms are aligned by scan number which appears at the top of the figure. Scan number can be converted to retention time using ~ 6.25 scans/s as a conversion factor. Values on the right vertical axis are relative fluorescence units.

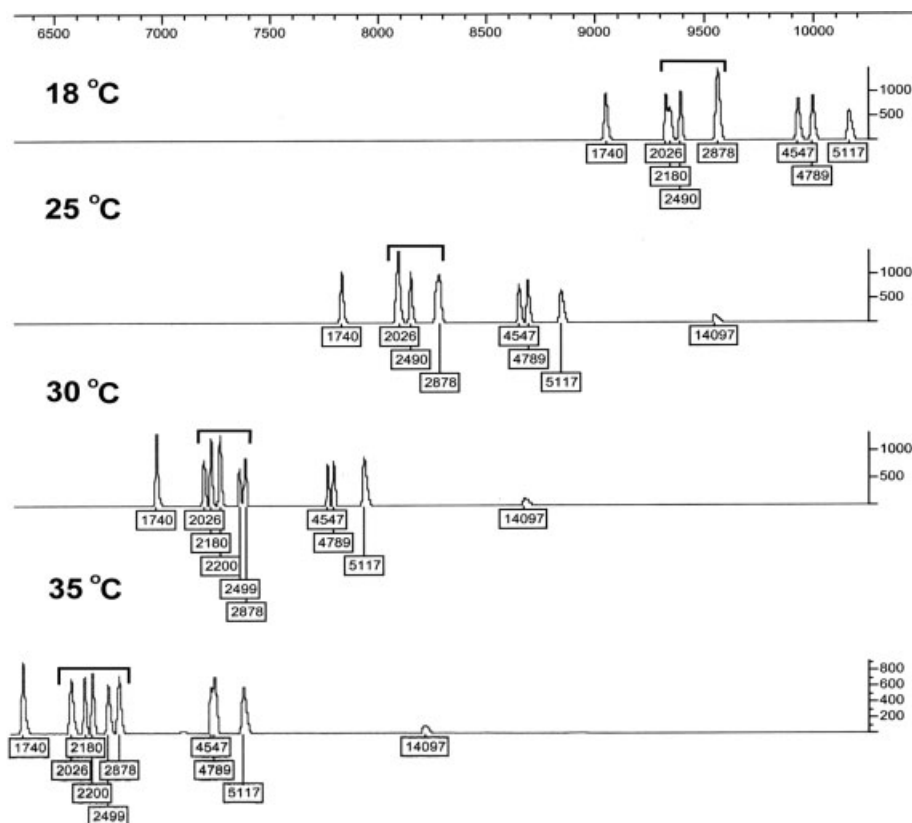


Figure 2. Electropherograms of undenatured ROX 2500 internal standards, 1740–14 097 bp range. Brackets indicate the MW range in which the standard peaks showed temperature-dependent mobility anomalies. For details, see Fig. 1.

14 097 bp, Fig. 2) over the four temperatures used in this study (18 °C, 25 °C, 30 °C, 35 °C). The mobility of the dsDNA decreased (the retention time increased) as the temperature decreased. Regression lines from a plot of the logarithm of the assigned MWs to the standard peaks *versus* the scan number at each temperature (Fig. 3) have high R^2 values between 0.976 and 0.984. Thus, the undenatured dsDNA MW standards migrated in a predictable manner under SSCP conditions, and could be used to assign MW values to the ssDNA.

3.2 Assignment of values to undenatured molecular weight standard peaks

3.2.1 Assignment of molecular weights

The ROX 2500 internal standard consists of 28 fragments ranging from 55 to 14 097 bp (Applied Biosystems, product information). From 55–1740 bp, the electrophoretic profile of the undenatured MW standards at each temperature was identical to the profile provided by Applied Biosystems although different polymers, temperatures, and instruments were used: 3% Genescan at 30 °C on an ABI 310 Genetic Analyzer (product literature) *versus* 5% Genescan with 10% glycerol at 18 °C, 25 °C, 30 °C, or 35 °C on an ABI 3100 Genetic Analyzer (this

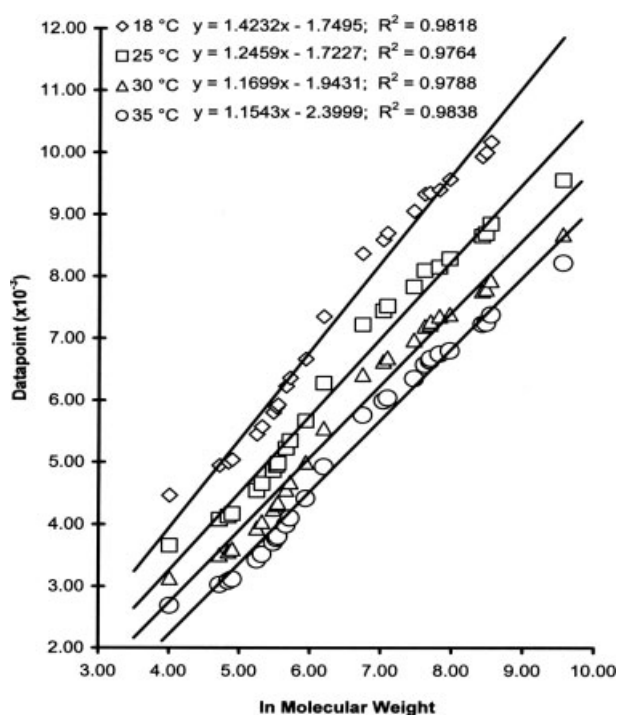


Figure 3. Plot of assigned MW in ln bp *versus* scan number at different temperatures for undenatured ROX 2500 internal standards. Retention time can be calculated using ~ 6.25 scans/s.

study). The assignment of MW sizes to the standard peaks from 55 to 1199 bp was unequivocal across all temperatures (Fig. 1).

For the higher MW standards (1740–14 097 bp range), the assignment of MW size for the 2026–2499 bp range was problematic at all temperatures. At all temperatures tested, the electrophoretic profiles of the higher MW standards were not identical with one another, nor were they identical with the profile presented in the product literature. The electrophoretic profile at 30°C (Fig. 2) most closely resembled the product literature. The higher percentage polymer (5% Genescan) did allow better resolution of the three fragments between 4547 and 5117 bp at all temperatures. However, the fragments from 2026 to

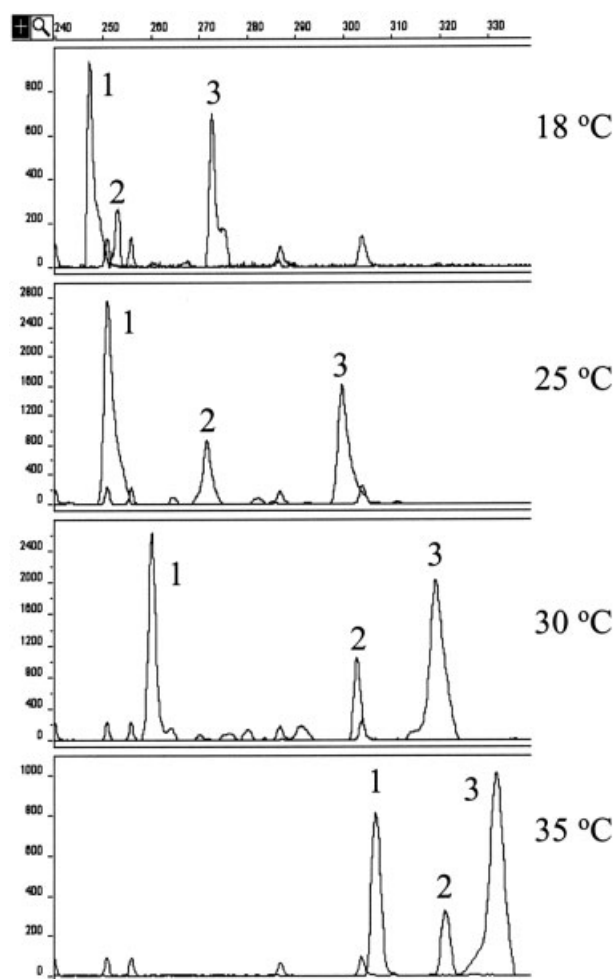


Figure 4. Effect of temperature on mobility of the single strands in relation to the internal standard. Shown are the electropherograms for the forward strands of the small fragments. The individual fragments are indicated by number. The electropherograms have been aligned by size (bp) which appears at the top of the figure to allow alignment across all four temperatures. Values on the left vertical axis are relative fluorescence units.

2878 bp exhibited variable mobilities at each temperature. At different temperatures, different fragments displayed identical mobilities (Fig. 2). In the product literature, the fragments of 2026, 2180 and 2483 bp have essentially identical mobilities while the next fragment separated by baseline is 2499 bp. Thus, two fragments differing by 16 bp (2483 and 2499) are easily resolved, but three fragments differing by >400 bp (2026, 2180 and 2483) are not. The assignment of the correct MW to the correct peak was equivocal within this range.

In instances where two peaks overlapped, the smaller MW value was assigned to the peak. For example, the 2026 and 2180 bp standards showed coincident mobilities at 25°C, and the peak was assigned a value of 2026 bp (Fig. 2). Because of uncertainty in correct assignment at any temperature, the 2483 bp fragment peak was not labeled with a value for alignment purposes. At 30°C and 35°C, this led to the miscalling of the surrounding peaks. At these two temperatures, a value of 2200 bp, the “approximate size” of the peak as determined from plotting its mobility *versus* the regression line obtained at that temperature, was applied (Fig. 2, 30°C and 35°C). Additionally, the 14 097 bp standard peak was low and broad at all temperatures, and the peak was absent in more than half of the electropherograms at 35°C. The absence of the 14 097 bp standard peak prevented the analysis of the large fragments (F, G, H, and I) that migrated between the 5117 bp and the 14 097 bp standards at 35°C. The assignment of actual MWs to the standard peaks made it possible to align and directly compare samples electrophoresed at different temperatures (Fig. 4).

3.2.2 Assignment of pseudomolecular weights

The assignment of a pseudoMW was readily accomplished across both size ranges for all temperatures. For the 1740–14 097 bp range, the pseudoMW values were easily assigned, because each individual peak did not need to be identified with a “correct” value, but only that the same peak across all samples was labeled with the same value. As the pseudoMW values were derived from the mobility of the undenatured standard peaks within a single temperature and these mobilities differed at the different temperatures, samples electrophoresed at different temperatures could not be aligned or directly compared when analyzed by pseudoMW.

3.3 Elution order of individual fragments

Each fragment was electrophoresed individually at each temperature to characterize the mobility of each strand and the data summarized in Table 3 as the order of elu-

Table 3. Resolution of fragments in order of elution at different temperatures by different scoring methods

Size	Strand	T (°C)	rawDP ^{a)}	MW ^{b)}	pseudoMW ^{c)}
Small	Reverse	18	1<2<3	1<2<3	1<2<3
		25	1<2<3	1<2<3	1<2<3
		30	1<2<3	1<2<3	1<2<3
		35	1<2<3	1<2<3	1<2<3
Medium	Forward	18	A=D<B=E	A=D<B=E	A=D<B=E
		25	E<B<A<D	E<B<A<D	E<B<A<D
	Reverse	18	A=D<B=E	A=D<B=E	A=D<B=E
		25	E=B<D<A	E=B<D<A	E=B<D<A
	Forward	30	E<B<D=A	E<B<D=A	E<B<D=A
		30	E=B<D<A	E=B<D<A	E=B<D<A
	Reverse	30	E=B<D<A	E=B<D<A	E=B<D<A
		35	E=B<D<A	E=B<D<A	E=B<D<A
Large	Forward	18	K<F=I<G	K<F=I<G	K<F=I<G
		25	K<G<F<I	K<G<F<I	K<G<F<I
	Reverse	18	G=I<F<K	G<I<F<K	G<I<F<K
		25	G<I<F<K	G<I<F<K	G<I<F<K
	Forward	30	K<G<F=I	K<G<F=I	K<G<F=I
		30	G<I<K=F	G<I<K<F	G<I<K<F
	Reverse	30	K<G<F=I	K<G<F=I	K<G<F=I
		30	G<I<K=F	G<I<K<F	G<I<K<F

a) Fragment peaks were scored by the scan number without correction by internal MW standards.

b) Fragment peaks were scored by correction with internal dsDNA MW standards that had been assigned their fragment length.

c) Fragment peaks were scored by correction with internal dsDNA MW standards that had been assigned their scan number.

=, fragments could not be distinguished by Genotyper and have identical mobilities or the means were not different at the 95% confidence limit.

≤, fragments could not be distinguished by a single measurement due to overlap of ranges of mobility values but the means were different at the 95% confidence limit.

<, fragments could be distinguished by a single measurement and the means were different at the 95% confidence limit.

tion. The small fragments showed marked variability in mobility across the four temperatures (Fig. 4). However, the order in which the forward and reverse strands eluted was consistent over all four temperatures (Table 3). Each reverse strand produced a single peak that was easily distinguished by mobility, and, when the fragments were combined, was resolved from the others at all four temperatures (Fig. 4). Each of the forward strands produced a pattern of two peaks, representing two stable conformers (data not shown). When the three fragments were combined together, the pattern produced by the forward strands was complex, many of the peaks overlapped with one another, and

correct automated assignment of the peaks to each fragment (allele) was difficult at each temperature. Thus, the forward strands for the small fragments were uninformative, and were not included in further analysis.

The patterns and mobilities of the medium and large fragments across the temperatures were also complex. Unlike the small fragments, the order in which the forward and reverse strands of the medium and large fragments eluted was not consistent either with one another at a single temperature nor was the order of elution maintained across the four temperatures (Table 3). For example, the order of elution for the forward strands of medium fragments A and D at 25°C is different than the reverse strands at the same temperature and different than the forward strands at 30°C. At every temperature, the order of alleles of the forward and reverse strands of the large fragments was different. In addition, the order of the reverse strands was different at 25°C and 30°C.

3.4 Mobility as a function of nucleotide differences

For the small fragments, 2 and 3 differed by two SNPs (Table 2) but were much closer in mobility at 30°C than were 1 and 3 which differed by only one SNP (Fig. 4). For the medium fragments, A and B differ by two SNPs and yet were closest in mobility at 25°C (forward strand: 486.5 and 484.3 bp MW), while fragments A and D, differing by one SNP, were quite different in mobility at 25°C (forward strand: 486.5 and 509.1 bp MW). For the large fragments, F and K differed by four SNPs while F and I differed by only one. However, at 25°C for the reverse strands, the difference in the mobilities of I (5238 MW) and F (5401 MW) was much greater than F and K (5474 MW).

3.5 Statistical analysis and contributions to variance

3.5.1 Pairwise comparison of the allele mobility means

To determine if fragments could be reliably distinguished by their mobility, with or without using undenatured dsDNA as MW standards, the 95% confidence limits for the means of the values for fragment peaks analyzed were calculated. The mobilities of the forward and reverse strands are presented as the MW (as determined from the MW in bp of the undenatured standards), the pseudoMW (based upon assignment of scan numbers from a single run to the undenatured standards), and the datapoint (the raw, unanalyzed scan number for the samples). A sample of the dataset is given in Table 4 for the four medium-size fragments at 25°C.

Table 4. Summary statistics for medium fragments at 25°C

Strand	Frag- ment	N	Variable	Lower 95% CL for mean	Upper 95% CL for mean	CV	Maximum	Mean	Minimum	Std. dev.
Reverse	A	48	rawDP	6511.65	6525.05	0.354	6566	6518.35	6470	23.07
			MW	563.51	564.39	0.268	566.74	563.95	559.88	1.51
			pseudoMW	6599.50	6602.41	0.076	6610.27	6600.96	6587.91	5.01
	B	48	rawDP	6457.80	6468.90	0.296	6510	6463.35	6423	19.11
			MW	545.75	546.43	0.214	548.23	546.09	542.54	1.17
			pseudoMW	6532.38	6535.04	0.070	6542.02	6533.71	6521.83	4.58
	D	48	rawDP	6460.82	6474.18	0.356	6516	6467.50	6419	23.00
			MW	549.50	550.28	0.245	552.32	549.89	546.35	1.35
			pseudoMW	6547.55	6550.29	0.072	6557.51	6548.92	6536.99	4.73
	E	48	rawDP	6457.80	6468.90	0.296	6510	6463.35	6423	19.11
			MW	545.75	546.43	0.214	548.23	546.09	542.54	1.17
			pseudoMW	6532.38	6535.04	0.070	6542.02	6533.71	6521.83	4.58
Forward	A	48	rawDP	6186.21	6194.87	0.241	6214	6190.54	6161	14.90
			MW	485.64	487.30	0.588	491.64	486.47	479.51	2.86
			pseudoMW	6263.53	6271.70	0.225	6293.42	6267.62	6236.58	14.08
	B	48	rawDP	6186.03	6196.05	0.279	6232	6191.04	6155	17.25
			MW	483.98	484.68	0.250	486.46	484.33	481.04	1.21
			pseudoMW	6253.17	6257.01	0.105	6266.81	6255.09	6237.63	6.60
	D	48	rawDP	6294.32	6304.97	0.291	6334	6299.65	6261	18.34
			MW	508.48	509.75	0.428	512.93	509.11	503.34	2.18
			pseudoMW	6374.79	6380.54	0.155	6394.84	6377.66	6351.40	9.90
	E	48	rawDP	6153.71	6163.87	0.284	6201	6158.79	6122	17.49
			MW	477.86	478.53	0.239	480.13	478.20	475.08	1.15
			pseudoMW	6220.62	6224.15	0.098	6233.03	6222.39	6205.34	6.08

CL, confidence limit

Std. dev., standard deviation

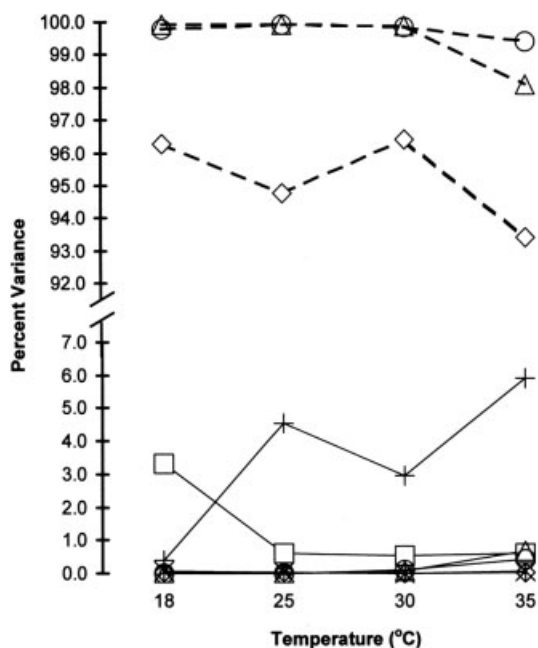
Pairwise comparisons of the fragment means at the 95% confidence limits were performed and are summarized in Table 3. For the small fragments, the means of fragment mobilities were distinct from one another at the 95% confidence limit at any temperature, using any size assignment method (MW, pseudoMW, or rawDP) (Table 3). For the medium and large fragments, whenever the fragment peaks could be resolved by Genotyper, the means of the fragment mobilities were significantly different at the 95% confidence level using either the MW or the pseudoMW values assigned to the peaks but not the rawDP (Table 3). For the medium fragments, all four could be discriminated by the mobility of the forward strands only at two temperatures, 25°C and 35°C. For the reverse strands, fragments B and E had coincident mobilities at every temperature. For the large fragments, the mobility of the forward strands was sufficient to discriminate among the four fragments only at 25°C, while the mobility of the reverse strands could discriminate the four fragments at 18°C, 25°C and 30°C.

3.5.2 Coefficient of variation of mobility values

For the small fragments, the coefficient of variation (CV) was always greatest for the rawDP, with the MW next greatest, and pseudoMW the least at all temperatures. There was no correlation between an increase in temperature and an increase in the CV. For the medium and large fragments, the pseudoMW CV was always the least but no other consistent correlation was observed.

3.5.3 Analysis of variance for allele, capillary and replication

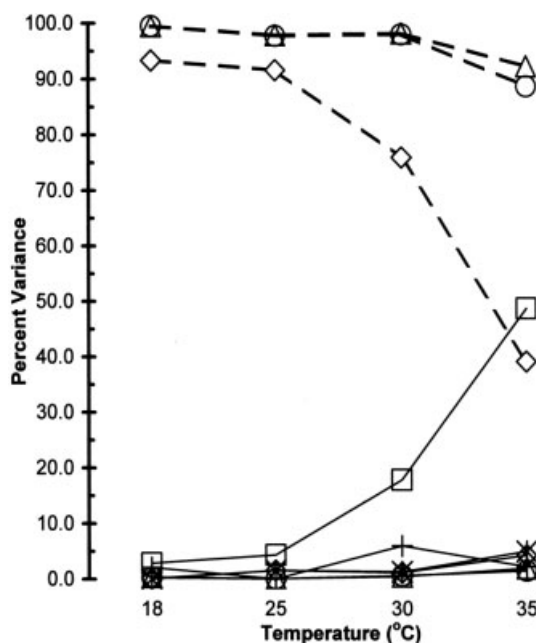
The relative contribution to observed variance by each factor was calculated in terms of percent from components for random effects and from sums of squares for fixed effects (Figs. 5–7). For the small fragments (Fig. 5), fragment differences were responsible for greater than 97% of the variance over all the temperatures when mobilities were assigned using the MW or pseudoMW, and



○— Fragment Molecular Weight
 △— Fragment Pseudo Molecular Weight
 ◇— Fragment Raw Datapoint
 *— Capillary Molecular Weight
 ◆— Capillary Pseudo Molecular Weight
 □— Capillary Raw Datapoint
 ○— Replication Molecular Weight
 △— Replication Pseudo Molecular Weight
 +— Replication Raw Datapoint

Figure 5. For small fragments (WRKY 1,2,3), each factor's contribution to the percent variation for each method of scoring is given at each temperature. Methods of scoring are described in Section 2.5.

greater than 90% for the rawDP. Contributions due to the capillary were negligible for the small fragments at any temperature when scored with the MW or pseudoMW. For the medium fragments (Fig. 6), fragment differences accounted for greater than 97% of the variance at 18°C, 25°C and 30°C for the MW and pseudoMW values, but at 35°C only contributed to 90% of the variance observed, with greater than 5% due to capillary-to-capillary variation. For the rawDP, contributions to variance due to the capillary was less than 5% at 18°C and 25°C but increased to 48% at 35°C, which was greater than the contribution due to the fragments (Fig. 6). For the large fragments (Fig. 7), the same trends as with the medium fragments were observed. The fragment contributed greater than 85% of the variation observed at the three temperatures when scored with either the pseudoMW or the MW with an increase in the variance due to the capillary at the higher temperatures. For all size fragments at all temperatures, the contributions to the variance by



○— Fragment Molecular Weight
 △— Fragment Pseudo Molecular Weight
 ◇— Fragment Raw Datapoint
 *— Capillary Molecular Weight
 ◆— Capillary Pseudo Molecular Weight
 □— Capillary Raw Datapoint
 ○— Replication Molecular Weight
 △— Replication Pseudo Molecular Weight
 +— Replication Raw Datapoint

Figure 6. For the medium fragments (A, B, D, E), each factor's contribution to the percent variation for each method of scoring is given at each temperature. Methods of scoring are described in Section 2.5.

replications were significant ($P > 0.05$), but accounted for a minor amount, less than 1%, of the total variation noted.

3.5.4 Capillary-to-capillary variation

As the capillary-to-capillary variation was the second largest effect on the variance at 30°C for the large fragments, the capillary means were calculated for MW, pseudoMW and rawDP and a pairwise comparison performed to determine if specific capillaries contributed predominantly to the variation noted. No consistent differences among the capillaries were noted (Fig. 8). In addition, capillary-to-capillary variation was not consistent on a particular capillary over all temperatures, runs, or between groups of runs (data not shown).

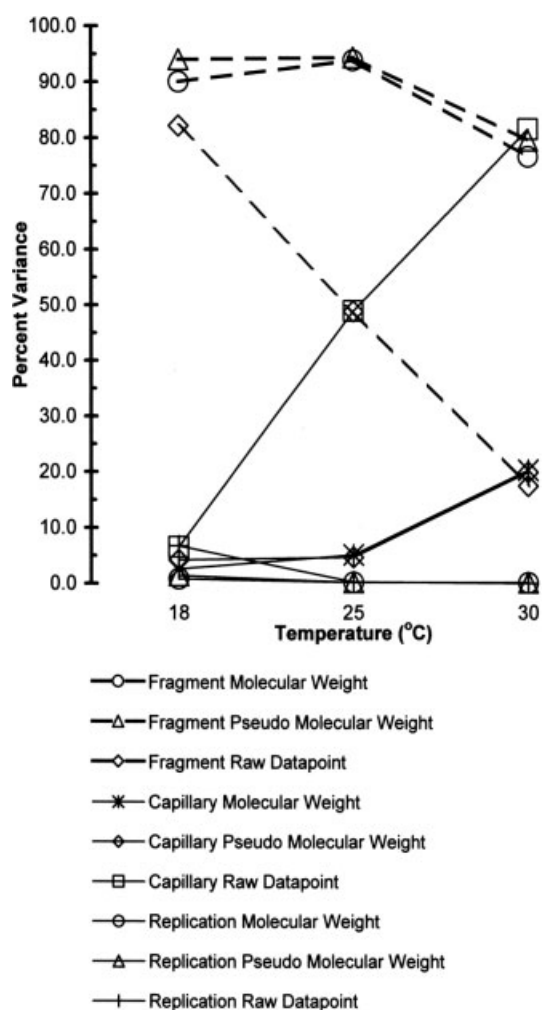


Figure 7. For the large fragments (F, G, I, K), each factor's contribution to the percent variation for each method of scoring is given at each temperature. Methods of scoring are described in Section 2.5.

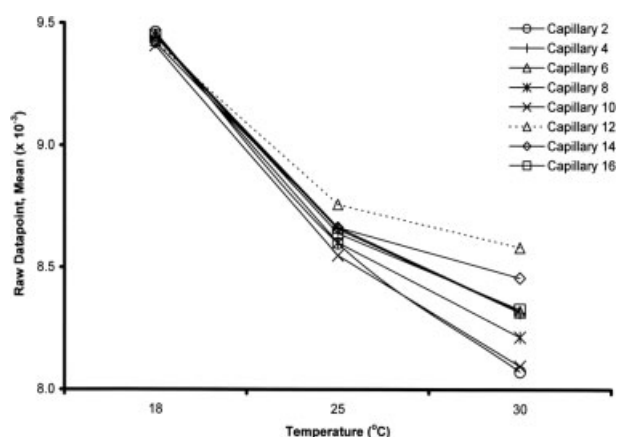


Figure 8. Means of the rawDP values for the reverse strand of the large fragments (F, G, I, K) for eight individual capillaries at 18°C, 25°C and 30°C.

3.6 Resolution of alleles based on a single measurement

As few biochemists or geneticists would ever measure any value 48 times, we have attempted to determine if alleles could be resolved by their mobilities after a single measurement. To this end, we have calculated both the mean and the range of values in Table 4. An example is presented here to determine if, from the mobility data, all four medium fragments could be discriminated by a single measurement independent of which capillary ran the samples. Because the order of elution of alleles cannot be predicted, the allele assignments to the mobilities were developed from the individual clone mobility data. For the forward strand (green), the range of calculated MW and pseudoMW values for D does not overlap with the range from any other allele. In addition, B can be distinguished from E and allele A from D and E. The only two fragments that could not be discriminated by their mobilities due to overlap of the ranges are A and B. However, the data from the reverse strand show that A can be distinguished from B, D or E and that there is no overlap in the range of the values. Thus, by considering the data from both strands, all four fragments can be resolved with a single measurement.

Table 3 is a summary of the ability of the fragments to be resolved by a single measurement. For the small fragments, the MW and pseudoMW were able to identify alleles at all temperatures for the reverse strand. RawDP was not capable of distinguishing all alleles at 18°C, 30°C and 35°C. For the medium fragments, as discussed above, data from both strands must be combined to identify all four fragments. No identification would be possible if fragments were measured a single time at 18°C. Combining of data for both strands for single measurements would identify all four alleles for pseudoMW at 25°C, 30°C and 35°C and for MW at 25°C and 30°C. RawDP could not be used to identify all four alleles at any temperature even if strand data is combined. For large fragments, all four fragments cannot be distinguished by a single measurement at any temperature, by any method or by combining strand data because F and I cannot be reliably resolved.

4 Discussion

4.1 Use of undenatured dsDNA molecular weight standards

A primary concern with the use of CAE with SSCP is the choice of a suitable internal MW standard. We investigated the use of commercially available dsDNA MW standards (ROX 2500) which vary in size from 55 bp to

14 097 bp. The ROX 2500 MW standards provide a means to compare SSCP runs between temperatures when the actual length in bp are assigned to the peaks. This method is only useful in the range from 55 to 1199 bp which is suitable for single strands of fragments 155 and 245 bp in length. The correct assignment of length values to the larger fragments is not possible due to fusion of peaks, anomalous migration, and reduction and broadening of the largest fragment peaks to a point where they are not reliably detectable. The clearest evidence of this was that the CV for raw datapoint was often smaller than that for the MW in the range from 1740 to 14 097 bp, especially when the datapoint values were over 9000. Application of the MW to the internal standards allowed comparison of mobilities across temperatures. The pseudoMW gave a smaller CV than the MW but does not allow direct comparison across temperatures.

4.2 Identifying novel alleles from mobility data

Our interest is in using this SSCP method to develop molecular markers from candidate genes and ESTs that can be used to estimate genetic diversity in a germplasm collection. Hence, we would like to use this method without knowing in advance the mobilities of all the possible alleles, but rather to discover alleles by measuring their mobilities. In addition, we wanted to use the currently available commercial software to automate the calling of alleles at any locus to make high-throughput analysis practicable. Several problems arose during the investigation that have complicated but not prevented the use of this method to characterize germplasm collections.

(i) It is clear that the mobility of ssDNA under SSCP conditions is inconsistent and cannot be predicted for either strand at one temperature from information gathered at another temperature. It would be tempting to conclude that reducing the length of the fragment would address this difficulty, but because only three alleles of the small fragments were analyzed, consistency of allele order over both strand and temperature cannot be concluded. However, we are currently developing a software method to describe alleles for heterozygotes by their mobilities from both strands and from different temperatures, without regard to allele order. Initial treatment of the data from the cocoa germplasm collection suggests that this can be accomplished using either MW or pseudoMW.

(ii) We assumed that fragments differing by only a single SNP should be the most difficult to separate. This assumption was important because it deals with the trade off between short fragments that may contain fewer SNPs but are easier to resolve and long fragments that may contain more SNPs but may be more difficult to resolve if

only a single SNP is present. With the current expense of fluorescently labeled primers, it is important to capture the greatest amount of diversity in a locus with a single set of primers. For the large fragments, differentiating fragments that differed only by one SNP (F and I) was not possible from a single measurement at any temperature-dye combination. Thus, increasing the fragment length to 437 nt from 245 nt decreased the resolution of the method.

(iii) A unique advantage of SSCP is to be able to collect data at several different temperatures for the same fragments. However, as the temperature was raised, the amount of variance due to the capillaries increased for all sizes of fragments investigated (Figs. 5–7). At the lower temperatures, the fragments remain in the capillary for a longer period and greater diffusion may be expected to lead to broader peaks. The effect was exactly the opposite. Although the fragments travel faster at high temperatures, less stable conformations of ssDNA or increased diffusion due to the higher temperature may have produced the greater capillary-to-capillary variation. Because there was no consistent alteration of mobilities detectable due to the capillary, small differences in capillary physical properties may not be important in this variation.

Although the fragments were retained on the capillaries longer at the lower temperatures, their mobility increased in relation to the mobility of the undenatured double-stranded MW standards. Thus, at the lowest temperature, 18°C, the product peaks had the lowest apparent MW (Fig. 1) suggesting that increased stability of the ssDNA conformation increases mobility. However, at 18°C, the medium and large fragments that differed by a single SNP had identical mobilities. At 35°C, the ssDNA conformers for all size fragments ran closer together. In addition, for the higher MW range, the 14 097 bp fragment was often undetectable at 35°C. Thus, both 18°C and 35°C are not appropriate temperatures for analysis.

4.3 Reliability of mobility measurements

Our primary concern was to determine if ssDNA mobilities under SSCP conditions were consistent across the capillary array so that we could automate the scoring of alleles similar to that for microsatellite loci. Our data demonstrate that wherever two peaks could be detected by Genotyper, the means of those measurements would allow accurate assignment of the allele. This assignment could be done using the MW or the pseudoMW for all size fragments at all temperatures, and the rawDP for the small fragments reverse strand at all temperatures. Although mobilities for a particular strand at a particular

temperature could not be predicted, each strand had a definable and reliable mobility at each temperature. Therefore, in analyzing mobility data from a germplasm collection, we may conclude that strands with altered mobility represent novel sequences and, hence, novel alleles, without having to sequence each allele. Thus, for highly replicated measurements, all alleles for all size fragments could be identified even if they differed by only a single SNP.

The range of mobility values for each strand at each temperature across all capillaries and replications was analyzed to determine if an allele could be reliably identified after a single run. Our criterion was that alleles could be reliably identified if there was no overlap of ranges. In this analysis, for the small fragments, (160 nt) a single run at any temperature in any capillary should provide a reliable means to assign an allele based solely on the calculated MW or pseudoMW mobility of the fragment. For the intermediate-size fragments, all four fragment means could only be resolved at 25°C for the forward strand. However, at 25°C for the forward strand, A and B (which were not coelectrophoresed in our experiment) were almost indistinguishable by mobility and could not have been reliably identified by a single measurement. However, combining of data from both strands at 25°C was sufficient to resolve all the medium fragments with a single measurement. In addition, it appears that combining data from different temperatures should be the best method to define novel alleles by their mobilities. Because the order of alleles may change between temperatures, it will be necessary to compile datasets of mobilities for individuals homozygous at a locus and use these combined mobilities to identify the initial set of alleles. These allele descriptions can be subtracted from the descriptions of heterozygous individuals to allow the identification of novel alleles without knowing the order of elution *a priori*. We are currently developing software to automate this process.

As seen from our data for intermediate and large fragments, there are numerous examples where alleles cannot be separated on a particular strand or temperature. B and E coelectrophoresed at all temperatures for the reverse strand. Thus, it becomes essential to combine data for both strands and two temperatures to avoid misidentifying heterozygous individuals as homozygous. This should increase the sensitivity of this method to allow detection of every allele. We are currently comparing the estimation of genetic diversity in a germplasm collection using this method and microsatellite analysis.

For the large fragments, combining data over two temperatures does allow G and K to be distinguished from each other and from F and I. However, there is no temperature-strand combination that would allow a com-

pletely reliable (*i.e.*, the ranges of the values from the control experiment do not overlap) assignment of F or I to a homozygous individual by a single measurement or by combining measurements at two different temperatures. Thus, unless F and I were present as a heterozygote (as in the control experiment where they can be correctly identified when coelectrophoresed), homozygotes of either F or I could be misidentified. For the large fragments, increased numbers of SNPs do not improve resolution and fragments differing by a single SNP are not resolved by combining measurements at two different temperatures. Therefore, we do not recommend using fragments of 437 nt or greater to estimate genetic diversity in germplasm collections.

4.4 Recommendations

To adequately assess alleles at a candidate gene locus using SSCP on a capillary array: fragments should be ≤ 250 bp, samples should be analyzed at two different temperatures between 18°C and 30°C to reduce the variability introduced by the capillaries, and undenatured dsDNA MW standards, such as ROX 2500, should be included as internal standards. For each sample, either the actual MW of the internal standard or a pseudoMW should be used to estimate the mobility of the ssDNA alleles. Data should be combined from both strands and both temperatures to completely resolve all alleles at the locus. Multiplexing of samples of different size range is possible, as intermediate and large fragments were analyzed simultaneously in these experiments. Slight changes in polymer solutions and changing of capillary arrays also has an effect on absolute mobilities of ssDNA, even when normalized (data not shown). Therefore, when possible, loci should be analyzed using the same polymer solution and capillary array.

We gratefully acknowledge the financial support of Masterfoods USA, MARS Inc. We thank Dr. Cuauhtemoc Cervantes-Martinez and Ms. Kyoko Nakamura for helpful discussions.

Received April 19, 2004

5 References

- [1] Schneider, K., Weisshaar, B., Borchardt, D. C., Salamini, F., *Mol. Breed.* 2001, 8, 63–74.
- [2] Baba, S., Kukita, Y., Higasa, K., Tahira, T., Hayashi, K., *Bio-Techniques* 2003, 34, 746–750.
- [3] Kirk, B. W., Feinsod, M., Favis, R., Kliman, R. M., Barany, F., *Nucleic Acids Res.* 2002, 30, 3295–3311.
- [4] Landegren, U., Nilsson, M., Kwok, P. Y., *Genome Res.* 1998, 8, 769–776.

- [5] Orita, M., Suzuki, Y., Sekiya, T., Hayashi, K., *Genomics* 1989, 5, 874–879.
- [6] Srinivasan, J., Sinz, W., Jesse, T., Wiggers-Perebolte, L., Jansen, K., Buntjer, J., van der Meulen, M., Sommer, R. J., *Mol. Genet. Genomics* 2003, 269, 715–722.
- [7] Fulton, R. E., Salasek, M. L., DuTeau, N. M., Black, W. C., *Genetics* 2001, 158, 715–726.
- [8] Plomion, C., Hurme, P., Frigerio, J. M., Ridolfi, M., Pot, D., Pionneau, C., Avila, C., Gallardo, F., David, H., Neutelings, G., Campbell, M., Canovas, F. M., Savolainen, O., Bodenies, C., Kremer, A., *Mol. Breed.* 1999, 5, 21–31.
- [9] Hongtrakul, V., Slabaugh, M. B., Knapp, S. J., *Mol. Breed.* 1998, 4, 195–203.
- [10] Slabaugh, M. B., Huestis, G. M., Leonard, J., Holloway, J. L., Rosato, C., Hongtrakul, V., Martini, N., Toepfer, R., Voetz, M., Schell, J., Knapp, S. J., *Theor. Appl. Genet.* 1997, 94, 400–408.
- [11] Ren, J. C., Ueland, P. M., *Hum. Mutat.* 1999, 13, 458–463.
- [12] Inazuka, M., Wenz, H. M., Sakabe, M., Tahira, T., Hayashi, K., *Genome Res.* 1997, 7, 1094–1103.
- [13] Kuhn, D. N., Heath, M., Wisser, R. J., Meerow, A., Brown, J. S., Lopes, U., Schnell, R. J., *Theor. Appl. Genet.* 2003, 107, 191–202.
- [14] Larsen, L. A., Christiansen, M., Vuust, J., Andersen, P. S., *Comb. Chem. High Throughput Screen.* 2000, 3, 393–409.
- [15] Kukita, Y., Higasa, K., Baba, S., Nakamura, M., Manago, S., Suzuki, A., Tahira, T., Hayashi, K., *Electrophoresis* 2002, 23, 2259–2266.
- [16] Larsen, L. A., Christiansen, M., Vuust, J., Andersen, P. S., *Hum. Mutat.* 1999, 13, 318–327.
- [17] Atha, D. H., Kasprzak, W., O'Connell, C. D., Shapiro, B. A., *Nucleic Acids Res.* 2001, 29, 4643–4653.
- [18] Higasa, K., Kukita, Y., Baba, S., Hayashi, K., *BioTechniques* 2002, 33, 1342–1348.
- [19] Andersen, P. S., Jespersgaard, C., Vuust, J., Christiansen, M., Larsen, L. A., *Hum. Mutat.* 2003, 21, 455–465.
- [20] Schnell, R. J., Olano, C. T., Kuhn, D. N., *Electrophoresis* 2001, 22, 427–432.
- [21] Wenz, H. M., Ramachandra, S., O'Connell, C. D., Atha, D. H., *Mutat. Res.* 1998, 382, 121–132.
- [22] Borrone, J. W., Kuhn, D. N., Schnell, R. J., *Theor. Appl. Genet* 2004, 109, 495–507.
- [23] SAS Institute Inc., *Stat Users Guide* 2000, Version 8, Cary, NC, USA.