Simple Sequence Repeat DNA Length Polymorphisms

P. B. Cregan, Research Geneticist
Soybean and Alafalfa Research Laboratory
USDA, Agricultural Research Service
Beltsville, MD

The suggestion by Botstein et al. (1980) that Restriction Fragment Length Polymorphisms (RFLP) provide the basis of a new type of genetic linkage map has led to the construction of such maps in numerous animal and plant species.

The markers on these maps have had broad application, ranging from the localization of genetic loci controlling human disease to the improvement of plant varieties by plant breeders. While not always the case, RFLP is often the result of the absence or presence of an endonuclease restriction site. Thus, in many instances only two alleles exist at a genetic locus. However, the likelihood that a particular molecular marker locus will be informative is positively related to the number of alleles at that locus. Thus, the probability that two plant inbreds will be polymorphic increases if the possibility of multiple alleles exists. Similarly, in an open-pollinated population the likelihood of heterozygosity at a locus will increase with the number of possible alleles.

The report of RFLP loci in humans with as many as eight different alleles (Wyman and White, 1980) suggested the possibility of greatly enhanced informativeness per locus. These so-called Variable Number Tandem Repeat (VNTR) loci (Nakamura et al. 1987) consist of sets of tandemly repeated DNA core sequences and have been referred to as "minisatellite" sequences by Jeffreys et al. (1985). The core units vary in length from 11 to 60 base pairs and the repeat region is flanked by conserved endonuclease restriction sites. Thus, the length of the restriction fragment produced by this type of genetic locus is proportional to the number of oligonucleotide core units it contains.

To complement RFLP markers, a second type of molecular marker based upon Polymerase Chain Reaction (PCR) technology (Mullis et al., 1986) has recently been widely used in plants. Williams et al. (1990) proposed the use of single arbitrary 10 base oligonucleotide PCR primers for the generation of molecular markers. These Random Amplified Polymorphic DNA (RAPD) markers are easily developed and because they are based on PCR amplification followed by agarose gel electrophoresis are quickly and readily detected. As a result, RAPD's may permit the wider application of molecular maps in plant science. Most RAPD markers are dominant and therefore, heterozygous individuals cannot be distinguished from both homozygotes. This contrasts with RFLP markers which are co-dominant and therefore, distinguish among the heterozygote and homozygotes. Thus, relative to standard RFLP markers, and especially VNTR loci, RAPD markers generate less information per locus examined.

PCR and Repetitive DNA Sequences

Alec Jeffreys and colleagues (Jeffreys et al., 1988) suggested combining the specificity and rapidity of PCR with the informativeness of VNTR loci in humans. Primers to the conserved flanking regions of VNTR loci were developed allowing PCR amplication of an entire VNTR locus. The resulting PCR products possess electrophoretic mobilities that differ according to the number of repeated DNA units in the VNTR allele(s) present.

This approach was recently extended to a different type of repetitive DNA in humans (Litt and Luty, 1989; Weber and May, 1989; Tautz, 1989). Rather than repeat units in the range of 11 to 60 base pairs in length, these workers suggested that high levels of length polymorphism exist in dinucleotide tandem repeat sequences. A dinucleotide repeat such as $(dC\text{-}dA)_n$. $(dG\text{-}dT)_n$ was reported to occur in the human genome as many as 50,000 times with n varying from 10 to 60. This type of reiterated sequence has been termed a Short Tandem Repeat Edwards et al. (1991), microsatellite (Litt and Luty, 1989) or a Simple Sequence Repeat (SSR) (Jacob, et al., 1991).

As is generally the case with VNTR loci, the DNA sequences flanking SSR's are conserved, allowing the selection of PCR primers that will amplify the intervening SSR in all genotypes of the target species. As initially reported, the PCR reaction includes a small amount of one ^{32}P-labeled nucleotide or one or two ^{32}P end-labeled primers to allow visualization of amplification products via autoradiography after electrophoresis on a standard sequencing gel. Variation in PCR product length is a function of the number of SSR units.

Figure 1 illustrates the detection of SSR length polymorphism using three genotypes, including two inbred parents and their F_1. Parent 1 is homozygous for the $(CA)_n$ allele and Parent 2, the $(CA)_{n-2}$ allele and each produces single PCR products. The F_1, being heterozygous, produces products corresponding to both alleles. Markers resulting from SSR length polymorphisms are placed on genetic maps in relation to other SSR, RFLP, RAPD, and phenotypic markers in a manner identical to that used with RFLP or RADP markers.

Developments

SSR Loci in Humans

Human geneticists first demonstrated the highly polymorphic nature of SSRs in 1989. In a $(TG)_n$ repeat in the human cardiac muscle actin gene locus, Litt and Luty (1989) detected 12 length variants (alleles) in only 37 individuals. Likewise, Weber and May (1989) reported successful amplification of products from 10 dinucleotide SSR loci and found from 4 to 11 alleles at each by typing a maximum of 78 individuals. Numerous reports support the high frequency of polymorphic SSR loci in the human genome. It is also important to note that SSR loci appear to be randomly spaced throughout the human genome (Hamada et al. 1982; Stallings et al. 1990). As defined by Edwards et al. (1991), SSR loci include tri- and tetrameric repeats such as $(AAT)_n$ and $(AGAT)_n$, respectively. According to these authors, the combined frequency in the human genome of all 44 possible unique tri- and tetrameric SSRs with n = 7 or greater is estimated to be 400,000 or one SSR per 10 kbp. If even a small fraction of these loci were polymorphic, they would provide ample markers for a saturated genetic map.

SSR Loci in DNA Fingerprinting

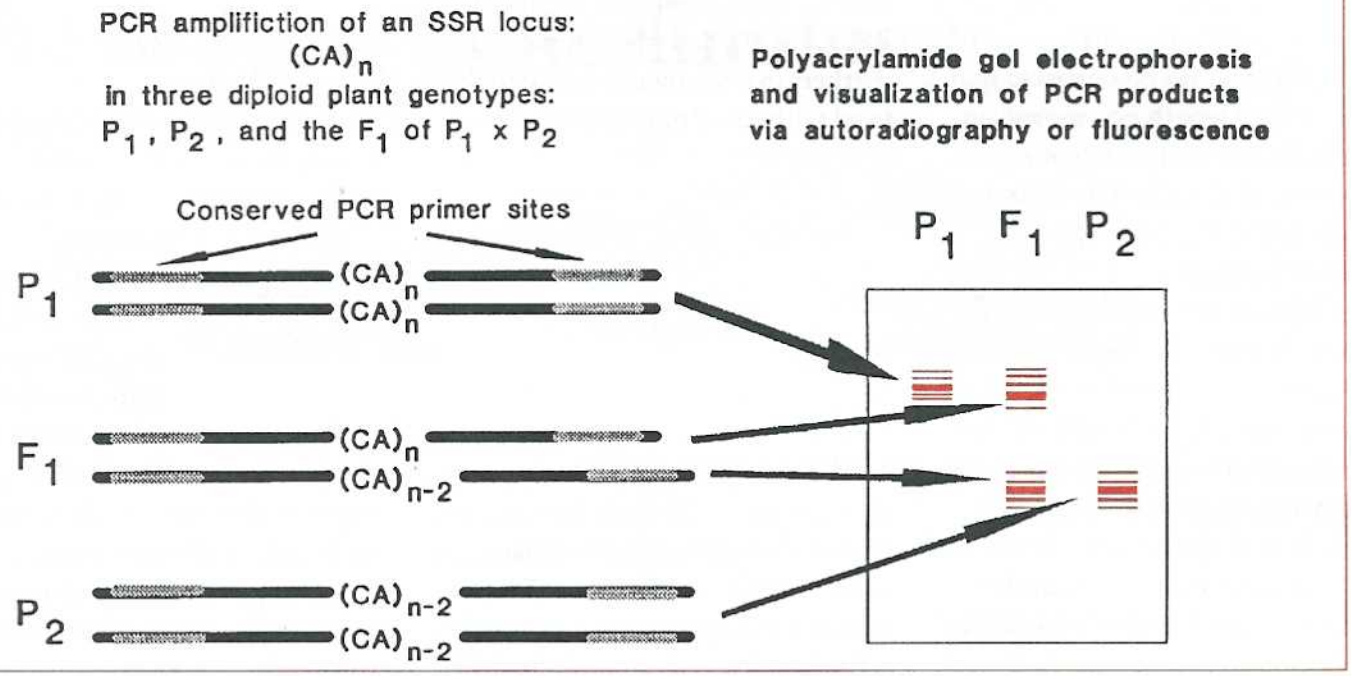
SSR markers can be employed in the development of unique allelic profiles for establishing individual identity. Distinctive profiles can be readily generated by defining the allelic constitution of individuals at relatively few loci, each of which is multi-allelic. Such a system is open-ended in that additional loci can be added if those already in use are inadequate to produce a unique profile for all individuals. With the advent of the Plant Variety Protection Act such a definitive cultivar identification system would be extremely useful.

Do SSRs Occur in Plants?

Can SSRs be used in plant genetic studies? Of particular interest in this regard is the occurence of SSR DNA in higher plants. Tautz et al. (1986) examined the European Molecular Biology Laboratory DNA sequence library for the presence of di- and trinucleotide repeats. A comparison of very limited plant and algal sequence data with those of vertebrates indicated a similar frequency of the two types of SSR's in the two groups of organisms.

Sarkar et al. (1991) reported a search of GenBank™ sequences for purine/pyrimidine repeats greater than 13 units in length. Calculations from their data indicate 2.3 such SSR

Figure 1. Generation and visualization of Simple Sequence Repeat Length Polymorphism.



loci per 100 kb in primates and 1.8 in yeast. While these reports are not clearly indicative of SSR's in higher plants, they do suggest the possibility of their occurrence. The only published report clearly documenting SSR's in higher plants is that of Condit and Hubbell (1991). They screened DNA libraries of five tropical tree species as well as *Zea mays* for the presence of clones containing $(AC)_n$ and $(AG)_n$ repeat sequences. They estimated a total of $(AC)_n + (AG)_n$ SSR sequences ranging from 5×10^3 to 3 to 10^5 among six species examined.

It seemed appropriate to undertake a further investigation of SSR DNA in plant genomes. Therefore, a search of GenBank™ was completed with the assistance of Dr. Susan McCarthy, Coordinator for the National Agricultural Library Plant Genome Data and Information Center. Despite the limited number of plant sequences available in GenBank™, a number of di- and tri-nucleotide SSR's were identified (Table 1). The data support the presence of SSR DNA in numerous plant species, including crop plants.

Do Plant SSR's Exhibit Length Polymorphism?

At this time no published information is available to answer this question. However, Weber (1990) suggested that human $(CA)_n$ sequences with n of 10 or less are unlikely to exhibit length polymorphism, whereas sequences with n greater than 15 are consistently polymorphic.

A more detailed look at the data obtained from the search of GenBank™, reported in Table 1, indicates at least one instance of a SSR with greater than 15 tandem repeats in each of the following higher plant species: *Arabidopsis thanliana, Daucus carota, Glycine max, Hordeum vulgare, Nicotiana tabacum, Pisum sativum, Oryza sativa, Solanum tuberosum,* and *Zea mays.* This finding suggests that plant SSR's have a high probability of exhibiting length polymorphism.

Technical Questions

One obvious drawback to developing a genetic linkage map generated using SSR markers is the time-consuming nature of the steps required to identify polymorphic loci. This is particularly true of SSR

Table 1. The number and average distance (in kilobase pairs) between all possible dimeric, trimeric, and tetrameric Simple Sequence Repeat DNA sequences in plant species determined from a search of GenBank™.

Plant species	Kilobases searched kbp	Dimeric repeats* No.	Dimeric repeats* Distance between repeats kbp	Trimeric repeats† No.	Trimeric repeats† Distance between repeats kbp	Tetrameric repeats‡ No.	Tetrameric repeats‡ Distance between repeats kbp
Saccharomyces cervisiae	2288	40	57	29	79	2	1144
Nicotiana tabacum	118	4	29	0	-	0	-
Glycine max	212	6	35	1	212	4	53
Lycopersicon esculentum	135	2	68	0	-	1	135
Triticum aestivum	151	1	151	43	4	0	-
Medicago sativa	30	0	-	1	30	0	-
Pisum sativum	129	3	43	3	43	2	64
Zea mays	368	3	123	4	91	6	61
Arabidopsis thaliana	247	4	62	1	247	0	-
Oryza sativa	137	2	68	2	68	6	23

*Dimeric repeats such as $(AT)_n$ with $n > 9$.
†Trimeric repeats such as $(ATT)_n$ with $n > 7$.
‡Tetrameric repeats such as $(AGAT)_n$ with $n > 4$.

markers as compared with the RAPD system. First, a DNA library must be developed and screened with a repetitive sequence oligonucleotide probe to identify desired clones. If the library is composed of relatively short clones, selected clones can be sequenced in their entirety to identify the SSR and determine flanking sequences. If the library is composed of longer sequences, subcloning and identification of the appropriate subclone would be required before sequencing. To expedite SSR isolation and sequencing, at least two PCR-assisted procedures have been suggested. Edwards et al. (1991) proposed a protocol that allows the amplification of the subclone containing the SSR followed by sequencing of flanking DNA regions. Browne and Litt (1992) suggested the use of a set of degenerate sequencing primers that anneal directly to the SSR. Both procedures should allow relatively rapid determination of sequences flanking SSR's and thereby expedite PCR primer selection.

A second possible impediment to the widespread use of SSR length polymorphisms by plant geneticists may be the perception that the routine detection of PCR products differing only slightly in length is difficult and laborious. Much of the work reported to date with SSR makers has used ^{32}P labeled PCR products and denaturing polyacrylamide gels. The use of fluorescent dye labeled PCR primers and fluorescent detection of multiplex PCR products (Edwards et al., 1991) offers the prospect of rapid determination of the allelic constitution of three SSR loci from one PCR reaction.

A less costly approach, but one that may be quite functional, would be the use of non-denaturing polyacrylamide gel separation followed by ethidium bromide or silver staining. The elimination of secondary structure via denaturation may not be necessary for distinguishing DNA fragments that only vary in composition by the presence or absence of a few internal SSR units.

Many technical questions remain as to the applicability and use of polymorphic SSR sequences in plant genetic studies. However, the informativeness of this type of marker, the rapid detection via PCR, and the potential of tens of thousands of SSR sequences per genome suggest that plant geneticists may wish to consider the use of polymorphic SSR loci as genetic markers. Genetic markers generated via variation in SSR length may provide a useful complement to the RFLP and RAPD markers currently in use.

REFERENCES

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32,314-331.

Browne, D. L. and Litt, M. (1992) Characterization of $(CA)_n$ microsatellites with degenerate sequencing primers. Nucleic Acids Res. 20,141.

Condit, R. and Hubbell, S. P. (1991). Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. Genome 34,66-71.

Edwards, A., Civitello, A., Hammond, H. A., and Caskey, C. T. (1991). DNA

typing and genetic mapping with trimeric and tetrameric tandem repeats. Am. J. Hum. Genet. 49,746-756.

Hamada, H., Petrino, M. G., and Kakunaga, T. (1982). A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Natl. Acad. Sci. USA 79,6465-6469.

Jacob, H. J., Lindpaintner, K., Lincoln, S. E., Kusumi, K., Bunker, R. K., Mao, Yi-Pei, Ganten, D., Dzau, V. J. and Lander, E. S. (1991). Genetic mapping of a gene causing hypertension in the stroke-prone hypertensive rat. Cell 67,213-224.

Jeffreys, A. J., Wilson, V., Neumann, R., and Keyte, J. (1988). Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. Nucleic Acids Res. 16,10953-10971.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). Hypervariable "minisatellite" regions in human DNA. Nature 314,67-73.

Litt, M. and Luty, J. A. (1989). A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. 44,397-401.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erilich, H. (1986). Specific enzymatic amplification of DNA *in vitro*: The polymerase chain reaction. Cold Spring Harbor Symp. Quant. Biol. 51:263-273.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto E., Hoff, M., Kumlin, E., and White, R. (1987). Variable number tandem repeat (VNTR) markers for human gene mapping. Science 235,1616-1622.

Sarkar, G., Paynton, C., and Sommer, S. S. (1991). Segments containing alternating purine and pyrimidine dinucleotides: patterns of polymorphism in humans and prevalence throughout phylogeny. Nucleic Acids Res. 19,631-636.

Stallings, R. L., Torney, D. C., Hildebrand, C. E., Longmire, J. L., Deaven, L. L., Jett, J. H., Doggett, N. A., and Moyzis, R. K. (1990). Physical mapping of human chromosomes by repetitive sequence fingerprinting. Proc. Natl. Acad. Sci. USA 87,6218- 6222.

Tautz, D. (1989). Hypervariablity of simple sequences as a general source of polymorphic DNA markers. Nucleic Acids Res. 17,6463-6471.

Tautz, D., Trick, M., and Dover, G. A. (1996). Cryptic simplicity in DNA is a major source of genetic variation. Nature 322,652-656.

Weber, J. L. (1990). Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. Genomics 7,524-530.

Weber, J. L. and May, P. E. (1989). Abundant class of human DNA polymorhisms which can be typed using the polymerare chain reaction. Am. J. Hum. Genet. 44,388-396.

Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A., and Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 18,6531-6535.

Wyman, A. R. and White, R. (1980). A highly polymorphic locus in human DNA. Proc. Nat. Acad. Sci. USA 77,6754-6758. ◆

CURD — continued from page 16

among experts, co-sponsors, and problem areas.

Availability

CURD is available online to anyone using a microcomputer and modem. After allowing users one free search, KSU charges a fee per online session. Individuals seeking access may contact the CURD office at the following address:

Corn Utilization Research Database
Food and Feed Grains Institute
Farrell Library, Room 419
Kansas State University
Manhattan, KS 66506-1200
Ph (913) 532-7452; FAX (913) 532-5861
INTERNET:
CURD@KSUVM.KSU.EDU ◆