DNA FINGERPRINTING TECHNIQUES AND CULTIVAR IDENTIFICATION

K.V. Bhat
National Research Centre on DNA Fingerprinting, NBPG, New Delhi-110 012

1. Introduction
Establishing the identity of crop varieties has assumed greater importance for protecting plant breeders’ and farmers’ rights in the post-CBD scenario particularly in the developing countries. An unambiguous characteristic pattern of cultivars obtained using DNA markers has been termed as a DNA fingerprint. The use of DNA markers to obtain a genotype specific profile has distinct advantages over morphological and biochemical methods. The morphological markers are influenced by the environmental conditions, are labour intensive and above all time consuming. Biochemical markers such as isozyme and protein patterns though minimally influenced by the environment offer limited polymorphism, and often do not allow discrimination between closely related genotypes. DNA markers overcome most of these disadvantages of morphological and biochemical markers.

Alec Jeffery et al. (1985) developed the technique of DNA fingerprinting in human beings for the first time. The usefulness of DNA fingerprinting technique for cultivar identification was demonstrated first by Dallas (1988) in rice. DNA identification analysis, identity testing, profiling, fingerprinting, typing of genotypes all refer to the characterization of one or more relatively rare features of an individual’s genome or hereditary make-up (Kirby, 1990).

Besides helping in varietal identification, DNA fingerprinting has numerous other applications in crop genetics such as: study of genetic diversity within a taxa and for the study of evolutionary and genetic relationships; tagging of economically useful traits; assessment of genetic purity of inbred lines and varieties; selection of recurrent parental genome in back cross and segregating generations thereby decreasing the number of generations required to breed pure line varieties; study breeding behaviour of crop plants and determine isolation distances in seed production programmes; selection of parents in a breeding programme to a) obtain higher heterosis, b) select parents with higher GCA for a polycross programme, c) select parents with wider genetic base; paternity analysis in perennials and tree crops; identification of hybrids and maternal seedlings in plants with polyembryony, e.g. Citrus, mango etc.; monitor genetic stability of germplasm conserved in form of seeds or tissues; for the detection of somaclonal variants; screening of duplicate accessions in gene banks and also for crime solving in forensics.

A number of molecular marker systems used routinely, either alone or in combinations, is available now for DNA fingerprinting and cultivar identification. Some of these are listed below:
- Restriction Fragment Length Polymorphism (RFLP)
- Random Amplified Polymorphic DNA (RAPD)
• Simple Sequence Repeats (SSR) Analysis
• Amplified Fragment Length Polymorphism (AFLP)
• DNA sequencing
• DNA micro-array analysis

2. Molecular Marker Systems

2.1 Restriction fragment length polymorphism (RFLP) analysis
In restriction fragment length polymorphism (RFLP) analysis, DNA is digested with restriction enzymes (which cut long strands of genomic DNA into smaller fragments at specific sites known as recognition sites), the resultant fragments are separated by gel electrophoresis, transferred to a filter by blotting (Southern blotting) and then specific labeled probes are hybridized to the filter (Southern hybridization) and subsequently leading to the detection of fragments by autoradiography (refer Flow chart 1). Specific probe/enzyme combinations give highly reproducible patterns for a given individual but variation between individuals can arise when mutations alter restriction sites or the fragment lengths between them (Burr et al. 1983). RFLPs are highly reproducible between laboratories. They are codominant markers and therefore allow the discrimination between heterozygotes from homozygotes. Availability of probes that can reliably detect variation at the level required and finding sufficient polymorphisms can be a problem at the cultivar level or population level in some species. If it is not possible to utilize probes from other related species (heterologous), new probes must be isolated from cDNA or genomic DNA libraries, which requires additional skill and investment of resources. RFLPs are time consuming and they are not amenable to automation. Once probe/enzyme combinations have been selected, throughput will depend on the number of gels that can be run each day in a laboratory. RFLPs require high quantities of good quality DNA and where limited amounts of source material are available, this requirement alone may preclude their application.

A variation of traditional RFLPs uses synthetic oligonucleotides of simple sequence repeat complexities to detect variations in hyper variable regions of the genome and is best suited for identification of genotypes. The most commonly used sequences are (AT)$_s$, (AG)$_s$, (CT)$_s$, (GACA)$_4$ and (GATA)$_4$ repeats. Polymorphism results from changes in the copy number of the basic repeat and is often referred to as Variable Number of Tandem Repeats (VNTRs). Hybridization of these probes essentially produces a multilocus pattern but has been found useful in generating individual or genotype specific hybridization patterns. They can also be used for cultivar identification in the same way as RFLPs.

2.2 Random Amplified Polymorphic DNA (RAPD)
Polymerase chain reaction (PCR) amplification of discrete loci with single, random sequence, oligonucleotide primers have become popular because of its simplicity and ease of use in modestly equipped laboratory (Williams et al. 1990 and Welsh and McClelland, 1990). The RAPD amplification reaction is performed on genomic DNA with an arbitrary oligonucleotide primer. It results in the amplification of several discrete DNA products. These are usually separated on agarose gels and visualized by ethidium
bromide staining. Each amplification product is derived from a region of the genome containing two short DNA segments with some homology to the primer, present on the opposite strands of DNA, and close enough to each other so that DNA amplification can occur (refer Flow chart 2). Short primers and low annealing temperatures during the amplification assure that several sites, randomly distributed in the genome, give rise to amplification products. The polymorphisms between individuals result from sequence differences in one or both of the primer binding sites. Such polymorphisms behave as dominant genetic markers. The reactions require small amounts of DNA (15 - 25 ng) and products can be detected by simple non-radioactive techniques.

2.3 Simple Sequence Repeats (SSR)
Presence of short tandem repeats of varying length is characteristic of microsatellite loci (Akkaya et al. 1992). The simple sequence repeats (SSR) that are similar in nature have been shown to be abundant and highly polymorphic in eukaryotic genomes. They may be dinucleotide repeats \((AC)_n\), \((AG)_n\), and \((AT)_n\) or trinucleotide repeats or tetranucleotide repeats. SSRs are analyzed by PCR-amplification of a short genomic region containing the repeated sequence, and size estimation of the repeat length by gel separation. Very little DNA is required for SSR analyses, if the primer sequences are known a laboratory set up to perform RAPD analyses could be used for the purpose. Microsatellites are highly useful markers for cultivar identification as they have been shown to be highly polymorphic and genotype specific. The technique is more robust and reproducible. Hence this is fast replacing RAPDs as a tool for cultivar identification.

2.4 Amplified Fragment Length Polymorphism (AFLP)
A new method has recently been developed which is universally applicable. It reveals very high levels of polymorphism and is highly reproducible (Vos et al. 1995). This procedure, termed amplified fragment length polymorphism (AFLP) is essentially intermediate between RFLP and PCR. The first step involves restriction digestion of the genomic DNA, followed by selective rounds of PCR amplification of the restricted DNA fragments. The amplified products are separated on a denaturing sequencing gel and visualized after exposure to X-ray film, but the technique has been adapted to use fluorescent-labeled primers and Automated Sequencers. AFLPs provide an effective means for covering large areas of the genome in a single assay. All the evidence so far indicates that they are as reproducible as RFLPs. But they require more DNA (1 µg) per reaction and are more technically demanding than RAPDs.

2.5 DNA micro-array analysis
The non-gel based techniques such as DNA micro-array analysis; which have been in use in human genome analyses are fast being developed in plant systems as well. The technique involves synthesis of an array of oligonucleotides specific to a locus on small micro-plates. Each nucleotide complementary to a locus differs from the adjacent one by a single base. Hybridization of labeled fragments of genomic DNA at specific conditions on to this chip and further treatment for ELISA-like color development provides for scoring of differences in the DNA base sequences at a locus among the cultivars. These arrays are reusable and the technique is amenable to automation, thereby increasing the sample throughput.
The oligonucleotide hybridization is very specific as a stable hybrid occurs only if the oligonucleotide is able to form a completely base-paired structure with the target DNA. If there is a single mismatch the hybrid does not form. The DNA micro-arrays are designed to allow much hybridization. With the technology available, a density of up to one million oligonucleotides per cm² has been achieved (Ramsay, 1998). This results in screening for half a million polymorphisms in a single experiment.

3. Choice of the technique
Selection of a technique in a particular taxa and in a laboratory would be governed by several criteria; the most important ones are availability of the technology for adoption, cost involved in development and analysis, level of polymorphism detectable, possibility of automation when the number of samples to be handled are large, reliability of the technique and the level of the skill required for the analysis. A comparison of the various techniques is presented in Table 1 in the Appendix.

4. Cultivar Identification
Analyses with these techniques (except micro-array analyses) produce X-ray autoradiograms or gel patterns comprising of ‘band profiles’. The objective of DNA fingerprinting is to identify molecular loci that are polymorphic enough to produce ‘band profiles’ which are ‘genotype specific’. Unlike human populations (which are highly heterozygous and heterogeneous), the cultivars sharing common parentage might not produce genotype specific band profiles with a single probe or primer with any of the techniques. Therefore use of a set of probes or primers might be necessary.

Several statistical parameters are considered before selection of a technique as the most suitable for genotype identification. One of the most useful parameter in the selection of probes or primers is the probability of chance identity of patterns of any two genotypes analyzed. This probability should be as low as possible for a probe to be useful for DNA fingerprinting. Use of a combination of such probes decreases the chance of false “inclusions”. Calculation of these essential parameters would be discussed in detail in the lecture.

A brief guide to calculation of various genetic diversity parameters and most commonly used procedures in cluster analyses are presented here.

4.1 Genetic diversity parameters

1. Polymorphism information content:
   \[ \text{PIC} = 1 - \left( \sum_{i=1}^{n} p_i^2 \right) - \left( \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} 2 p_i^2 p_j^2 \right) \]
   where, \( p_i \) & \( p_j^2 \) = frequency of \( i^{th} \) and \( j^{th} \) allele of a marker

The formula takes the following form in case of markers with null alleles as in the case of RAPDs:

\[ \text{PIC} = 2 p_i (1 - p_i) \]
where, \( p_i \) = frequency of the \( i^{th} \) null allele
Parameters at species level

2. **Percent polymorphic loci:**
   \[ P_s = \frac{\text{Number loci polymorphic}}{\text{total number of loci compared}} \]
   A locus is defined as polymorphic when  \( p_i \) of most common allele is <1.0.

3. **Mean number of alleles per locus:**
   \[ A_s = \text{average of allele frequency over all loci.} \]

4. **Gene diversity:**
   \[ H_{es} = 1 - \sum p_i^2 \]
   Where, \( p_i \) = mean frequency of \( i^{th} \) allele.
   Mean gene diversity, \( H_e \) is then calculated by averaging the \( H_{es} \) values over all loci.
   Ranges from 0 to 1.
   \( H_e \) is also defined as the probability that two alleles randomly chosen from a population will be different.
   \( H_e \) is the expected proportion of heterozygous loci in a randomly chosen individual.
   \( H_e \) is the average proportion of heterozygotes in a random mating population.

5. **Effective number of alleles:**
   \[ A_{es} = \frac{1}{1 - H_{es}} \]
   \( A_{es} \) is equal to the number of alleles when all alleles have the same frequency.

6. **Observed heterozygosity:**
   \[ H_o = \text{Proportion of heterozygous individuals in the sample.} \]
   \( H_o = H_e \) under random mating and no selection.

At population level the parameters are

7. **Percent polymorphic loci:**
   \[ P_p = \text{Proportion of loci polymorphic in each population averaged over all populations.} \]

8. **Number of alleles per locus:**
   \[ A_p = \text{Number of alleles for each population / number of populations} \]

9. **Genetic diversity for each locus and population:**
   \[ H_{ep} = 1 - \sum p_i^2 \]
   Where, \( p_i \) is the frequency of the \( i^{th} \) allele in each population.
   Mean \( H_{ep} \) is obtained for each locus by averaging over all populations and an overall mean (\( H_{ep} \)) is obtained by averaging over all loci.

10. **The effective number of alleles:**
    \[ A_{ep} = \frac{1}{1 - H_{ep}} \]
11. **Shannon-Weaver Information Index:**
   \[ SWI \text{ or } H = \sum p_i \ln p_i \]
   Where, \( p_i \) = Frequency of \( i^{th} \) allele at a locus.
   This parameter is not bound by the 0 to 1 limit.
   More sensitive than \( H_e \)
   No genetic interpretation for this parameter.
   Statistical properties partly known.

12. **Diversity in primary hierarchical groups:**
   \[ H_p = \text{Average of } H \text{ over all loci} \]
   Diversity within a group of collections,
   \[ H = \left(\frac{1}{M}\right) \sum_{i=1}^{m} H_i \]
   Where, \( H_i \) = Diversity at each allele.
   \( M \) = Total number of alleles.

13. **Diversity in secondary hierarchical groups** = \( H_s \)
14. **Diversity in overall collections** = \( H_t \)
15. **Proportion of diversity within first hierarchical level** = \( H_p/H_s \)
16. **Proportion of diversity in second hierarchical level** = \( H_s/H_t \)
17. **Proportion of diversity between groups in first level**
    \[ = \frac{H_s - H_p}{H_s} \]
18. **Proportion of diversity between groups in second level**
    \[ = \frac{H_t - H_s}{H_t} \]
19. **Relative contribution of each group to total diversity**
    \[ = \frac{H_p}{H_t} \text{ and } \frac{H_s}{H_t} \]

Several statistical software are freely available in the Internet, which can be downloaded.
Some of the most often used software and their applications are summarized below:

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Software</th>
<th>URL</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.</td>
<td>GDA</td>
<td><a href="http://biology001.unm.edu/~lewis/gda.htm">http://biology001.unm.edu/~lewis/gda.htm</a></td>
<td>A Windows based program for population genetic analysis</td>
</tr>
<tr>
<td>4.</td>
<td>ARLEQUIN</td>
<td><a href="http://anthropologie.unige.ch/arlequin">http://anthropologie.unige.ch/arlequin</a></td>
<td>A Windows based program for calculating standard diversity indices, molecular indices, genetic distances, neutrality tests, HW equilibrium test, AMOVA from both co-dominant and dominant traits</td>
</tr>
</tbody>
</table>
4.2 Cluster analyses

Similarity and dissimilarity measures
The data obtained with molecular analyses may be classified as ‘qualitative’ or ‘genetic’ depending on (i) whether the analyses have been performed on the individual plants or representative bulked samples; (ii) whether the markers scored are previously mapped or unmapped, and (iii) whether the alternate states of the markers share ‘dominant-recessive’ or ‘co-dominant’ relationships, that is whether the alternate states of the marker/alleles can be distinguished unambiguously or not? Therefore, the data may be in the ‘presence/absence’ (+/-) or ‘gene frequencies’ form. The main purpose of this chapter is to introduce the workers to, arguably, two of the most used softwares in plant diversity analyses, namely, ‘NTSYS’ (Rohlf, 19) and ‘PHYLIP’ (Felsenstein, 19). For details on the genetic diversity parameters and analyses the readers may refer to some of the several treatises available (Nei, 1987; Karp et al., 1998; Weir, 1996; Swofford et al., 1996). The major objective here is to introduce the workers to the basics of molecular data analyses using ‘NTSYS’ and ‘PHYLIP’ that deal with numerical taxonomic analyses and cladistics respectively.

(i) Gel scoring for ‘qualitative’ type of data:

<table>
<thead>
<tr>
<th>Presence/ absence of a band in j&lt;sup&gt;th&lt;/sup&gt; &amp; i&lt;sup&gt;th&lt;/sup&gt; taxa</th>
<th>j</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td>i</td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>a</td>
</tr>
<tr>
<td>-</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>d</td>
</tr>
</tbody>
</table>

Calculate, 

m = a+d (number of matches)

u = b+c (number of unmatches)

n = u+m (total sample size)

(ii) Jaccard’s similarity coefficient (Jaccard, 1908):

\[
J = \frac{a}{n-d}
\]

(iii) Dice coefficient (Dice, 1945):

\[
\text{Dice} = \frac{2a}{2a+b+c}
\]

(iv) Simple matching coefficient:

\[
\text{SM} = \frac{m}{n}
\]

(v) Gel scoring for ‘gene frequency’ type of data:

<table>
<thead>
<tr>
<th>V 1233</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Plant 2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Plant 3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Plant 4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Plant 5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Plant 6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

q = 1 - \text{smallest q}

q = 0.408

p = 1 - q

p = 0.592

705
(vi) **Genetic distance:**

The Nei’s genetic distance is calculated from the data of gene frequencies using the formula (Nei, 1972) –

\[
d_{ij} = -\ln\left[\frac{\sum k |x_{ki}x_{kj}|}{\left(\sum k x_{ki}^2 x_{kj}^2\right)^{1/2}}\right]
\]

### 4.3 Calculation of Genetic Diversity parameters with POPGENE

**Data file format for POPGENE**

**Format for diploid data, co-dominant marker:**

/*Diploid data of 3 populations each with five genotypes & 10 loci*/

Number of populations = 3  
Number of loci = 10  
Locus name: ADH-1 ADH-2 ACP-1 MDH-1 MDH-2 MDH-3 PGD-1 PGD-2 GDH-1 GDH-2

<table>
<thead>
<tr>
<th>Population</th>
<th>ADH-1</th>
<th>ADH-2</th>
<th>ACP-1</th>
<th>MDH-1</th>
<th>MDH-2</th>
<th>MDH-3</th>
<th>PGD-1</th>
<th>PGD-2</th>
<th>GDH-1</th>
<th>GDH-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
</tr>
<tr>
<td>2</td>
<td>AA</td>
<td>AA</td>
<td>AB</td>
<td>BB</td>
<td>A3</td>
<td>AA</td>
<td>AB</td>
<td>BB</td>
<td>AA</td>
<td>AA</td>
</tr>
<tr>
<td>3</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>BC</td>
<td>AC</td>
<td>AA</td>
<td>AB</td>
<td>AB</td>
<td>AB</td>
<td>AB</td>
</tr>
<tr>
<td>4</td>
<td>BA</td>
<td>AA</td>
<td>AA</td>
<td>BB</td>
<td>BC</td>
<td>AA</td>
<td>BB</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
</tr>
<tr>
<td>5</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>BB</td>
<td>AA</td>
<td>AA</td>
<td>AB</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
</tr>
<tr>
<td>6</td>
<td>AA</td>
<td>AB</td>
<td>AB</td>
<td>BB</td>
<td>AA</td>
<td>AA</td>
<td>AB</td>
<td>AB</td>
<td>AB</td>
<td>AB</td>
</tr>
</tbody>
</table>

**Format for diploid data, dominant marker:**

/* Diploid RAPD Data Set */

Number of populations = 4  
Number of loci = 28  
Locus name: OPA01-1 OPA01-2 OPA01-3 OPA01-4 OPA01-5  
OPA02-1 OPA02-2 OPA02-3 OPA02-4 OPA02-5 OPA02-6  
OPA03-1 OPA03-2 OPA03-3 OPA03-4 OPA03-5 OPA03-6  
OPA04-1 OPA04-2 OPA04-3 OPA04-4 OPA04-5 OPA04-6 OPA04-7  
OPA05-1 OPA05-2 OPA05-3 OPA05-4 OPA05-5 OPA05-6  
OPA06-1 OPA06-2 OPA06-3 OPA06-4 OPA06-5 OPA06-6  
OPA07-1 OPA07-2 OPA07-3 OPA07-4 OPA07-5 OPA07-6  
OPA08-1 OPA08-2 OPA08-3 OPA08-4 OPA08-5 OPA08-6  
OPA09-1 OPA09-2 OPA09-3 OPA09-4 OPA09-5 OPA09-6  
OPA10-1 OPA10-2 OPA10-3 OPA10-4 OPA10-5 OPA10-6  
OPA11-1 OPA11-2 OPA11-3 OPA11-4
DNA Fingerprinting Techniques and Cultivar Identification

name = Slave Lake
11101 100100 0011010 001100 0001
11111 100100 0011010 001100 0001
11101 100101 0011010 110110 1011
11101 111000 0011011 110100 0001
11000 100000 0011010 001000 0111

name = Little Smoky
11101 111101 0111010 111000 1111
11101 111100 0011010 001000 1111
11110 100001 1011010 101010 1011
11101 111000 0011010 001011 0001
10101 111111 0011010 111000 0101

name = Whitecourt
10001 101100 0000100 001001 0001
11111 100101 0010010 001001 1011
11111 100100 1101010 001000 1011
11011 111000 0010010 001100 0001
11001 100000 0011010 001000 0101

name = Devon
11111 100111 0001010 100010 0011
10111 101111 0000010 100010 1101
10101 111110 1000010 010010 0001
10111 101111 0111010 010001 0101
10100 101100 0001010 010101 0110

OUTPUT FROM POPGENE

Single-Population Descriptive Statistics
Population ID :      1
Population name :      Slave Lake

Gene Frequency :  

<table>
<thead>
<tr>
<th>All</th>
<th>OPA01-1</th>
<th>OPA01-2</th>
<th>OPA01-3</th>
<th>OPA01-4</th>
<th>OPA01-5</th>
<th>OPA03-1</th>
<th>OPA03-2</th>
<th>OPA03-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele 0</td>
<td>0.4472</td>
<td>0.8944</td>
<td>0.4472</td>
<td>0.8944</td>
<td>0.8944</td>
<td>0.8944</td>
<td>0.8944</td>
<td>0.8944</td>
</tr>
<tr>
<td>Allele 1</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.5528</td>
<td>0.1056</td>
<td>0.5528</td>
<td>1.0000</td>
<td>0.1056</td>
<td>0.1056</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>All</th>
<th>OPA03-4</th>
<th>OPA03-5</th>
<th>OPA03-6</th>
<th>OPA04-1</th>
<th>OPA04-2</th>
<th>OPA04-3</th>
<th>OPA04-4</th>
<th>OPA04-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele 0</td>
<td>0.6325</td>
<td>0.1056</td>
<td>0.8944</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele 1</td>
<td>0.3675</td>
<td>0.1056</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### DNA Fingerprinting Techniques and Cultivar Identification

<table>
<thead>
<tr>
<th>Allele 0</th>
<th>0.8944</th>
<th>0.7746</th>
<th>0.7746</th>
<th>0.6325</th>
<th>0.4472</th>
<th>0.8944</th>
<th>1.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele 1</td>
<td>1.000</td>
<td>0.1056</td>
<td>0.2254</td>
<td>0.2254</td>
<td>0.3675</td>
<td>0.5528</td>
<td>0.1056</td>
</tr>
</tbody>
</table>

### Summary Statistics

**Summary of Genic Variation Statistics for All Loci**

[See Nei (1987) Molecular Evolutionary Genetics (p. 176-187)]

<table>
<thead>
<tr>
<th>Locus</th>
<th>Sample Size</th>
<th>na*</th>
<th>ne*</th>
<th>h*</th>
<th>I*</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPA01-2</td>
<td>5</td>
<td>1.000</td>
<td>1.000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>OPA01-3</td>
<td>5</td>
<td>1.000</td>
<td>2.0000</td>
<td>1.9780</td>
<td>0.4944</td>
</tr>
<tr>
<td>OPA01-4</td>
<td>5</td>
<td>2.0000</td>
<td>1.2328</td>
<td>0.1889</td>
<td>0.3372</td>
</tr>
<tr>
<td>OPA01-5</td>
<td>5</td>
<td>2.0000</td>
<td>1.9780</td>
<td>0.4944</td>
<td>0.6876</td>
</tr>
<tr>
<td>OPA11-4</td>
<td>5</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Mean: 1.5714  St. Dev: 0.3500  0.1860  0.2682

* na = Observed number of alleles
* ne = Effective number of alleles [Kimura and Crow (1964)]
* h = Nei's (1973) gene diversity
* I = Shannon's Information index [Lewontin (1972)]

The number of polymorphic loci is: 16
The percentage of polymorphic loci is: 57.14 %
Grouped Populations Descriptive Statistics
Group 1:
Population: 1 2
Group Gene Frequency:
Nei's Analysis of Gene Diversity in Subdivided Populations
[See Nei (1987) Molecular Evolutionary Genetics (p. 187-192)]

<table>
<thead>
<tr>
<th>Locus</th>
<th>Sample Size</th>
<th>Ht</th>
<th>Hs</th>
<th>Gst</th>
<th>Nm*</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPA01-1</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>****</td>
<td>****</td>
</tr>
<tr>
<td>OPA01-2</td>
<td>10</td>
<td>0.3472</td>
<td>0.2472</td>
<td>0.2880</td>
<td>1.2361</td>
</tr>
<tr>
<td>OPA01-3</td>
<td>10</td>
<td>0.3472</td>
<td>0.2472</td>
<td>0.2880</td>
<td>1.2361</td>
</tr>
<tr>
<td>OPA01-4</td>
<td>10</td>
<td>0.1889</td>
<td>0.1889</td>
<td>0.0000</td>
<td>2000.0000</td>
</tr>
<tr>
<td>OPA01-5</td>
<td>10</td>
<td>0.4944</td>
<td>0.4944</td>
<td>0.0000</td>
<td>2000.0000</td>
</tr>
<tr>
<td>OPA03-1</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>****</td>
<td>****</td>
</tr>
<tr>
<td>OPA03-2</td>
<td>10</td>
<td>0.4416</td>
<td>0.3416</td>
<td>0.2264</td>
<td>1.7082</td>
</tr>
<tr>
<td>OPA03-3</td>
<td>10</td>
<td>0.4416</td>
<td>0.3416</td>
<td>0.2264</td>
<td>1.7082</td>
</tr>
<tr>
<td>OPA03-4</td>
<td>10</td>
<td>0.4649</td>
<td>0.4649</td>
<td>0.0000</td>
<td>2000.0000</td>
</tr>
<tr>
<td>OPA03-5</td>
<td>10</td>
<td>0.1000</td>
<td>0.0944</td>
<td>0.0557</td>
<td>8.4721</td>
</tr>
<tr>
<td>OPA03-6</td>
<td>10</td>
<td>0.3612</td>
<td>0.3269</td>
<td>0.0950</td>
<td>4.7630</td>
</tr>
<tr>
<td>OPA04-1</td>
<td>10</td>
<td>0.1000</td>
<td>0.0944</td>
<td>0.0557</td>
<td>8.4721</td>
</tr>
<tr>
<td>OPA04-2</td>
<td>10</td>
<td>0.1000</td>
<td>0.0944</td>
<td>0.0557</td>
<td>8.4721</td>
</tr>
<tr>
<td>OPA11-2</td>
<td>10</td>
<td>0.3612</td>
<td>0.3269</td>
<td>0.0950</td>
<td>4.7630</td>
</tr>
<tr>
<td>OPA11-3</td>
<td>10</td>
<td>0.4172</td>
<td>0.4071</td>
<td>0.0242</td>
<td>20.1470</td>
</tr>
<tr>
<td>OPA11-4</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>****</td>
<td>****</td>
</tr>
<tr>
<td>Mean</td>
<td>10</td>
<td>0.2358</td>
<td>0.2032</td>
<td>0.1380</td>
<td>3.1234</td>
</tr>
</tbody>
</table>

St. Dev 0.0335 0.0260

*Nm= estimate of gene flow from Gst or Gcs. E.g., Nm=0.5(1-Gst)/Gst;
The number of polymorphic loci is: 21
The percentage of polymorphic loci is: 75.00

Nei's Original Measures of Genetic Identity and Genetic distance

<table>
<thead>
<tr>
<th>pop ID</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>****</td>
<td>0.9188</td>
</tr>
<tr>
<td>2</td>
<td>0.0847</td>
<td>****</td>
</tr>
</tbody>
</table>

Nei's genetic identity (above diagonal) and genetic distance (below diagonal).
DNA Fingerprinting Techniques and Cultivar Identification

Dendrogram Based Nei's (1972) Genetic distance: Method = UPGMA
--Modified from NEIGHBOR procedure of PHYLIP Version 3.5

+---------------------------------------------------+---+---+---+---+---+---+---+---+---+
| pop1      | 1 | pop2 | 1 |
+---------------------------------------------------+---+---+---+---+---+---+---+---+---+
Between And Length
1 pop1 4.23382
1 pop2 4.23382

Nei's Unbiased Measures of Genetic Identity and Genetic distance
[See Nei (1978) Genetics 89:583-590]

| pop ID | 1 | 2 |
+-------+---+---+
1 **** 0.9458
2 0.0558 ****

Nei's genetic identity (above diagonal) and genetic distance (below diagonal).

Dendrogram Based Nei's (1978) Genetic distance: Method = UPGMA
--Modified from NEIGHBOR procedure of PHYLIP Version 3.5

+-------------------------------------------+---+---+
| pop1      | 1 | pop2 | 1 |
+-------------------------------------------+---+---+
Between And Length
------- --- -------
1 pop1 2.78823
1 pop2 2.78823

Nei's Analysis of Gene Diversity in Subdivided Populations
[See Nei (1987) Molecular Evolutionary Genetics (p. 187-192)]

<table>
<thead>
<tr>
<th>Locus</th>
<th>Sample Size</th>
<th>Ht</th>
<th>Hc</th>
<th>Hs</th>
<th>Gst</th>
<th>Gcs</th>
<th>Nm(Gst)*</th>
<th>Nm(Gcs)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPA01-2</td>
<td>20</td>
<td>0.4944</td>
<td>0.2944</td>
<td>0.3944</td>
<td>0.2023</td>
<td>0.2535</td>
<td>1.9721</td>
<td>1.4721</td>
</tr>
<tr>
<td>OPA01-3</td>
<td>20</td>
<td>0.4243</td>
<td>0.2109</td>
<td>0.4109</td>
<td>0.0316</td>
<td>0.4867</td>
<td>15.3352</td>
<td>0.5273</td>
</tr>
<tr>
<td>OPA01-4</td>
<td>20</td>
<td>0.3612</td>
<td>0.3269</td>
<td>0.3269</td>
<td>0.0950</td>
<td>0.0000</td>
<td>4.7630</td>
<td>2000.00</td>
</tr>
<tr>
<td>OPA01-5</td>
<td>20</td>
<td>0.4458</td>
<td>0.3708</td>
<td>0.4208</td>
<td>0.0561</td>
<td>0.1188</td>
<td>8.4164</td>
<td>3.7082</td>
</tr>
<tr>
<td>OPA03-2</td>
<td>20</td>
<td>0.3402</td>
<td>0.2652</td>
<td>0.3152</td>
<td>0.0735</td>
<td>0.1586</td>
<td>6.3050</td>
<td>2.6525</td>
</tr>
<tr>
<td>OPA03-3</td>
<td>20</td>
<td>0.4603</td>
<td>0.3817</td>
<td>0.4585</td>
<td>0.0039</td>
<td>0.1675</td>
<td>127.727</td>
<td>2.4854</td>
</tr>
<tr>
<td>Mean</td>
<td>20</td>
<td>0.3007</td>
<td>0.2171</td>
<td>0.2672</td>
<td>0.1114</td>
<td>0.1874</td>
<td>3.9867</td>
<td>2.1685</td>
</tr>
<tr>
<td>St. Dev</td>
<td>0.0232</td>
<td>0.0131</td>
<td>0.0190</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DNA Fingerprinting Techniques and Cultivar Identification

*Nm* = estimate of gene flow from *Gst* or *Gcs*. E.g., *Nm* = 0.5(1 - *Gst*)/Gst;
The number of polymorphic loci is: 26
The percentage of polymorphic loci is: 92.86

**Nei's Original Measures of Genetic Identity and Genetic distance**

<table>
<thead>
<tr>
<th>pop ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>****</td>
<td>0.9188</td>
<td>0.8846</td>
<td>0.8178</td>
</tr>
<tr>
<td>2</td>
<td>0.0847</td>
<td>****</td>
<td>0.8705</td>
<td>0.8285</td>
</tr>
<tr>
<td>3</td>
<td>0.1226</td>
<td>0.1387</td>
<td>****</td>
<td>0.8246</td>
</tr>
<tr>
<td>4</td>
<td>0.2011</td>
<td>0.1882</td>
<td>0.1929</td>
<td>****</td>
</tr>
</tbody>
</table>

Nei's genetic identity (above diagonal) and genetic distance (below diagonal).

**Dendrogram Based Nei's (1972) Genetic distance: Method = UPGMA**
--Modified from NEIGHBOR procedure of PHYLIP Version 3.5

```
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>****</td>
<td>0.9188</td>
<td>0.8846</td>
<td>0.8178</td>
</tr>
<tr>
<td>2</td>
<td>0.0847</td>
<td>****</td>
<td>0.8705</td>
<td>0.8285</td>
</tr>
<tr>
<td>3</td>
<td>0.1226</td>
<td>0.1387</td>
<td>****</td>
<td>0.8246</td>
</tr>
<tr>
<td>4</td>
<td>0.2011</td>
<td>0.1882</td>
<td>0.1929</td>
<td>****</td>
</tr>
</tbody>
</table>
```

Nei's genetic identity (above diagonal) and genetic distance (below diagonal).

- **Nei's Unbiased Measures of Genetic Identity and Genetic distance**
  [See Nei (1978) Genetics 89:583-590]

<table>
<thead>
<tr>
<th>pop ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>****</td>
<td>0.9458</td>
<td>0.9087</td>
<td>0.8440</td>
</tr>
<tr>
<td>2</td>
<td>0.0558</td>
<td>****</td>
<td>0.8983</td>
<td>0.8590</td>
</tr>
<tr>
<td>3</td>
<td>0.0957</td>
<td>0.1072</td>
<td>****</td>
<td>0.8532</td>
</tr>
<tr>
<td>4</td>
<td>0.1696</td>
<td>0.1520</td>
<td>0.1588</td>
<td>****</td>
</tr>
</tbody>
</table>

Nei's genetic identity (above diagonal) and genetic distance (below diagonal).
Dendrogram Based Nei's (1978) Genetic distance: Method = UPGMA
--Modified from NEIGHBOR procedure of PHYLP Version 3.5

```
+------------------pop1
+--------------1
+-------------------2              +------------------pop2
!                   !
--3                   +---------------------------------pop3
!
+------------------------------------------pop4
```

Between    And         Length
3             2           2.93227
2             1           2.28566
1          pop1         2.78823
1          pop2         2.78823
2          pop3         5.07389
3          pop4         8.00616

USING ‘NTSYS’
The NTSYS-pc package from M/S Exeter Software, 100 North Country Road, Setauket, New York 11733, USA (price $250) is available in both MSDOS as well as Windows versions. It consists of programs for performing cluster analysis; several ordination measures and comes with useful programs to produce high quality graphics.

Data file format for NTSYS-pc:
```
1 6B 15 19
V1233   0 1 1 1 0 1 1 1 1 0 1 1 0 1 0 1 1 0 9 1
PUSA605 0 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1
UPAS120 0 1 1 1 1 1 1 1 1 0 1 1 0 1 1
892212   0 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 1
PUSA33   0 1 1 1 1 1 1 1 1 1 1 1 1 1 1
PUSA982  0 1 1 1 1 1 1 1 1 0 1 1 0 1 1
```

Alternate data format:
```
1 6L 15 0
V1233 PUSA605 UPAS120 892212 PUSA33 PUSA982
0 1 1 1 0 1 1 1 1 0 1 1 0 1 1
0 1 1 1 1 1 1 1 1 0 1 1 0 1 1
0 1 1 1 1 1 1 1 0 1 1 0 1 1 1
0 1 1 1 1 1 1 1 1 0 1 1 1 1 1
0 1 1 1 1 1 1 1 0 1 1 0 1 1 1
```

The first line is the header. ‘1’ indicates the type of file, i.e. raw data; ‘6B’ indicates number of rows, ‘B’ is included if row labels are provide in the beginning of the rows;
‘15’ indicates the number of data points in the columns; the last two places refer to the presence of missing data in the file, ‘1’ indicates presence of missing data or ‘0’ if no missing data is included; ‘9’ is the code denoting the missing data point.

The data starts with the second line in the first format. The first 16 places are for the row labels. In case of the alternate format, the second line is reserved for the row labels; each label can occupy up to 16 places. The third line onwards is reserved only for the data devoid of labels.

Output from “Simquality” program

Table of Jaccard’s similarity coefficients:
" SIMQUAL: input=C:\MYPROGRA\PPEA\PEARAPD\TEST.DTA, coeff=J, By Rows, += 1.00000, -= 0.00000
3 6 L 6 0
V123 PUSA605 UPAS120 892212 PUSA33 PUSA982
1.0000000
0.9090909 1.0000000
0.9090909 1.0000000 1.0000000
0.8333333 0.9230769 0.9230769 1.0000000
0.7692308 0.8571429 0.8571429 0.9285714 1.0000000
0.9090909 1.0000000 1.0000000 0.9285714 0.8571429 1.0000000

The UPGMA dendrogram:

---

Flow chart indicating the sequence of data analyses in ‘NTSYS’
References


Table 1: Comparison of molecular techniques and their applicability.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Proteins/Isozymes</th>
<th>RFLPs</th>
<th>RAPDs</th>
<th>SSRs</th>
<th>AFLPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development costs (US$/probe)</td>
<td>Low (None)</td>
<td>Medium (87)</td>
<td>Low (None)</td>
<td>High (437)</td>
<td>Low (None)</td>
</tr>
<tr>
<td>Level of polymorphism</td>
<td>Low</td>
<td>Medium</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Automation</td>
<td>No</td>
<td>No</td>
<td>Yes/No</td>
<td>Yes/No</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Cost of automation</td>
<td>High</td>
<td>High</td>
<td>Medium</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Reliability</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Level of skill required</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low/ Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Cost (US$/sample)</td>
<td>Low (0.3)</td>
<td>High (5.5)</td>
<td>Low (3.4)</td>
<td>Low (3.8)</td>
<td>Medium (19.7)</td>
</tr>
<tr>
<td>Radioactivity</td>
<td>No</td>
<td>Yes/ No</td>
<td>No</td>
<td>Yes/ No</td>
<td>Yes/ No</td>
</tr>
<tr>
<td>Samples/day (research)</td>
<td>100</td>
<td>20</td>
<td>50</td>
<td>50</td>
<td>48</td>
</tr>
<tr>
<td>Genetic nature of the marker</td>
<td>Co-dominant</td>
<td>Co-dominant</td>
<td>Dominant</td>
<td>Co-dominant</td>
<td>Dominant</td>
</tr>
<tr>
<td>Genetic control</td>
<td>Single locus</td>
<td>Single/multi locus</td>
<td>Multilocus</td>
<td>Single locus</td>
<td>Multilocus</td>
</tr>
<tr>
<td>Genome coverage</td>
<td>Poor</td>
<td>Medium</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Genome localization</td>
<td>Expressed loci, clustered</td>
<td>Clustered</td>
<td>Ubiquitous</td>
<td>Ubiquitous</td>
<td>Ubiquitous</td>
</tr>
</tbody>
</table>