Bioinformation by Biomedical Informatics Publishing group *open access*

www.bioinformation.net **Software**

UPIC: Perl scripts to determine the number of SSR markers to run

Renée S. Arias, Linda L. Ballard, Brian E. Scheffler*

USDA/ARS Genomics and Bioinformatics Research Unit, 141 Experiment Station Rd., Stoneville, MS 38776;
Brian E. Scheffler – E-mail: brian.scheffler@ars.usda.gov; *Corresponding author

received March 16, 2009; revised April 6, 2009; accepted April 11, 2009; published April 21, 2009

Abstract:
We introduce here the concept of Unique Pattern Informative Combinations (UPIC), a decision tool for the cost-effective design of DNA fingerprinting/genotyping experiments using simple-sequence/tandem repeat (SSR/STR) markers. After the first screening of SSR-markers tested on a subset of DNA samples, the user can apply UPIC to find marker combinations that maximize the genetic information obtained by a minimum or desirable number of markers. This allows a cost-effective planning of future experiments. We have developed Perl scripts to calculate all possible subset combinations of SSR markers, and determine based on unique patterns or alleles, which combinations can discriminate among all DNA samples included in a test. This makes UPIC an essential tool for optimizing resources when working with microsatellites. An example using real data from eight markers and 12 genotypes shows that UPIC detected groups of as few as three markers sufficient to discriminate all 12-DNA samples. Should markers for future experiments be chosen based only on polymorphism-information content (PIC), the necessary number of markers for discrimination of all samples cannot be determined. We also show that choosing markers using UPIC, an informative combination of four markers can provide similar information as using a combination of six markers (23 *vs*. 25 patterns, respectively), granting a more efficient planning of experiments. Perl scripts with documentation are also included to calculate the percentage of heterozygous loci on the DNA samples tested and to calculate three PIC values depending on the type of fertilization and allele frequency of the organism.

Availability: Perl scripts are freely available for download from http://www.ars.usda.gov/msa/jwdsrc/gbru

Keywords: simple sequence repeats, software, best SSR markers, microsatellites, GeneMapper

Background:
Repetitive DNA sequences, microsatellites or simple sequence/tandem repeats (SSR, STR) are widely spread throughout prokaryotic and eukaryotic genomes [1][2][3], and have a number of applications from marker-assisted breeding in plants [4] to detecting genetic disorders in humans [3]. Given the cost of running SSR markers, primers are usually screened on a subset of DNA samples before designing large scale experiments. Though some useful coefficients exist, such as polymorphism information content (PIC) [5] and Log_{10} of the likelihood ratio (LOD score) [5] to help determine which markers to use, currently, there are no available decision tools for cost-effective planning of fingerprinting or genotyping experiments.

Various PIC formulas are available in the literature, depending on whether the organisms are cross-fertilized [5], cross-fertilized and have equifrequent alleles [6], or are self-fertilized [6] (Formulas S1, 1.1.1, 1.1.2 and 1.1.3). Though software is available for the calculation of PIC values, such as Cervus [7] [**8**] or the on-line PIC calculator [9], no single site calculates the three mentioned PIC values. Other useful information when working with microsatellites is the average heterozygosity per locus (Formulas S1, 1.2.1) as a measure of the genetic variability of the population [10]. Knowing the degree of heterozygosity of the lines tested allows choosing parental lines for further studies, selecting lines with potential environmental fitness [11] or inferring ploidy of the tested DNA samples [12].

It is necessary to make a clear distinction between the polymorphism-information content (**PIC**) value developed by Botstein et al. (1980)[5], and the new approach presented here for choosing the best combinations of SSR markers that we now call UPIC. Whereas PIC values only indicate the information content of individual markers, UPIC calculates all possible subset combinations of markers and finds which combinations are the most informative. We introduce the concept of Unique Pattern Informative Combination (**UPIC**) to provide users of SSR markers with a decision tool that: (**a)** finds the most informative combinations of polymorphic markers based on the presence of unique patterns on the samples tested, and (**b)** allows the user to choose the number of markers to run depending on cost or objectives of the experiment. UPIC calculations do not require prior knowledge of genetic information of the populations to be analyzed such as genome size, ploidy or type of fertilization. In addition to UPIC values, the scripts presented here calculate percentage of heterozygous loci for each DNA sample and three PIC coefficients for self fertilized, cross-fertilized, and cross-fertilized with equifrequent alleles

© 2009 Biomedical Informatics Publishing Group

(Formulas S1) for the user to choose from, thus representing a convenient tool for microsatellite work.

Methodology:
After screening primers for developing SSR markers, a text file containing marker names, DNA samples and amplicon sizes is generated (*i.e.*, by GeneMapper, Applied Biosystems) and used as input for the scripts. The first row in this tab or space delimited text file contains the headers for the columns, please see example in **Table S1**. The scripts calculate: three PIC values [5][6], percent of heterozygous loci for each line, and the UPIC values proposed here.

UPIC calculation
Allele information of eight polymorphic markers that were run on 12 lines (DNA samples) was used in our example to show the mechanics of calculating unique-pattern informative combinations (UPIC). The various allele patterns observed for each marker (fingerprint) were compared as strings of amplicon (peak) sizes (**Table S2a**). In our example, the possible number of combinations of 8 polymorphic markers is 255. If we assign a letter to each pattern observed for a line (**Table S2b)** and then convert the letters to binary values, where "0" is assigned to an allele pattern present more than once across the lines tested, and "1" is assigned to unique patterns (UP), **Table S2c**. Please note that UP differ in at least one allele, therefore, UP values represent unique identifiers for the DNA sample.

Since various informative combinations (IC) with different total number of UP can be found, the UPIC script output consists of two columns, one is the total number of UPIC (*i.e.*, 18, **Table S2d**) in the combination, and the other is the marker combination. All UP values of each IC, for the data in our example, are shown in the UPIC plot, **Figure 1**. We have written UPIC version 1.0 which calculates all possible subset combinations of markers, where the range of subsets is selected by the user. The range minimum is combinations of two and the maximum is the number of markers in the input file. Each combination subset is calculated completely before the next larger subset. Details of the calculation of UPIC are provided in Formulas S1.

Details on Input/Output files and Scripts
The input file for UPIC needs to contain four columns, please see example in **Table S1**. The first column (in GeneMapper exported data corresponds to dye and amplicon/peak order) is not used by the scripts. Columns 2, 3 and 4 correspond to marker, DNA sample and amplicon size (peaks) respectively, these are the columns used by the scripts. An example of the output file for the calculation of UPIC values is shown in supplementary **Table S3**. The output shows the number of markers in the group, then the first column corresponds to the number of unique patterns (UP) observed for that combination of markers. An example of the output file for the calculation of percent of heterozygous loci and polymorphic information content (PIC) values is shown in supplementary **Table S4**, where the first column is for the name of the DNA sample (or line), and the second column is the percent of heterozygous loci. In the same output file there are another 5 columns that correspond to name of the marker, square of the allele frequencies, PIC value of self-fertilization, PIC value for equifrequent alleles and PIC values for cross-fertilized organisms. The user needs to select the PIC value that applies for his/her biological system. In order to run the script for UPIC calculation the user must install the Math::Combinatorics and Array::Compare, and Benchmark::Stopwatch Perl modules. The approximate computer time required to run UPIC version 1.0 script for calculating 2 to 8 combinations of 120 polymorphic markers across 6 DNA samples using a Dell Optiplex GX745 2.66 GHz dual-core Intel processor with 3.25 GB of RAM is ca. 5 min. Perl scripts for the calculations of UPIC, PIC and heterozygosity are available from the authors upon request. Each line of Perl script is either clearly self evident as to its function or is preceded by an explanatory comment. The user will receive a self extracting Zip file including test data and a README file with instructions for installation and use. UPIC Perl scripts can be downloaded from http://www.ars.usda.gov/msa/jwdsrc/gbru under *Products and Sevices/Bioinformatics Tools*.

Discussion:
When working with microsatellites, the size of the experiments that can be conducted in terms of number of samples and number of SSR markers to run is often limited by cost. The general recommendation is to run more markers with greater numbers of polymorphism or high PIC values [5]. However, no specific number of markers to run per experiment can be extracted from PIC values. Although PIC value gives a good estimation of the informativeness of a marker, the PIC value only refers to a particular marker, whereas UPIC analyzes all the markers in relation to each other and in the context of all samples evaluated, and provides the user with the most informative marker combinations to choose from. Another useful tool to choose markers is the LOD [5], however, this is used for known pedigrees and known genome sizes, and this information is not always available when working with diverse species and populations.

We have introduced here the concept of UPIC, a decision tool for the cost-effective design of DNA fingerprinting/genotyping experiments using polymorphic simple-sequence/tandem repeat (SSR/STR) markers. UPIC is a set of Perl scripts the user can apply to find the highest number of unique patterns (UP) or alleles on the best informative combination (IC) of polymorphic markers to use in an experiment. UPIC calculations consider the information of all markers and samples used in preliminary screening, and do not require having genetic information of the populations to be analyzed such as genome size, ploidy or type of fertilization. To the best of our knowledge, there is no program available that can assist in choosing the number of polymorphic markers to use as well as determine which combination of markers will provide the maximum

© 2009 Biomedical Informatics Publishing Group

discrimination among the DNA samples for fingerprinting or genotyping.

The UPIC plot in **Figure 1** represents the number of UP obtainable with IC of polymorphic markers for our example of 8 markers and 12 DNA samples. From our example, the benefits of UPIC calculation are: 1) Not all combinations of polymorphic markers are IC, only those that allow discrimination among all samples; in our example, only 72 IC were found out of 255 possible subset combinations of 8 polymorphic markers (histogram, **Figure 1**). 2) UPIC calculations identified a single combination of three markers (**Figure 1A**) that can discriminate all the DNA samples tested. 3) If using an IC of 4 markers, the amount of information (UP value) can vary from 19 to 23 (**Figure 1B**), so the user can choose the most informative one. 4) Running an IC of 4 markers provided almost the same information as running 6 markers (UP = 23,25; **Figure 1B, C**), therefore, the user could maximize information and minimize costs. 5) The scripts presented here also calculate three PIC values (for various fertilization types and allele frequencies) and the percent of heterozygous loci as additional decision tools. The flow diagram for the scripts is shown on **Figure 2**.

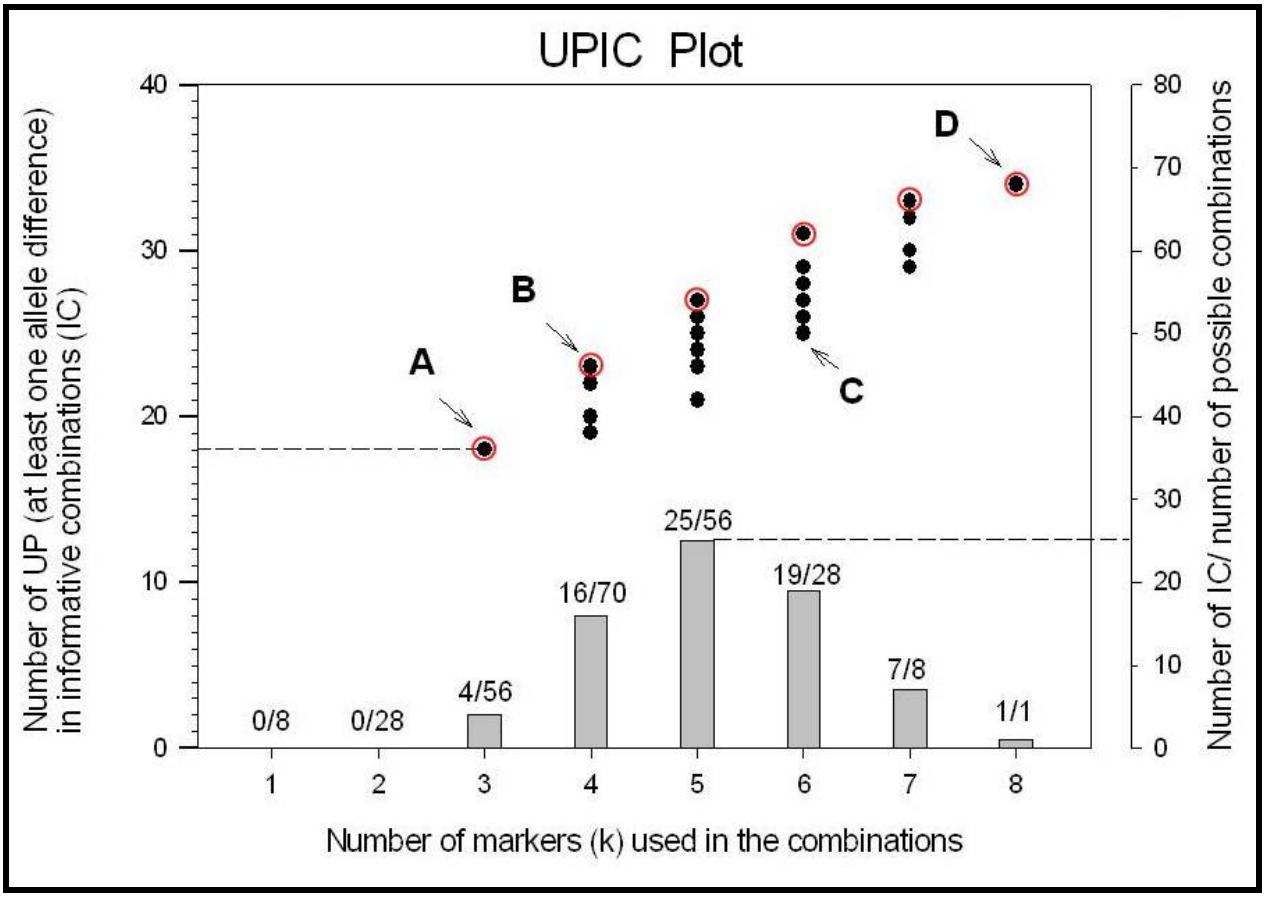

Figure 1: Graphic representation of UPIC values for the 8 markers and 12 DNA samples in our example. •: unique patterns (UP)(*y-axis*,*left*) that allow discrimination of the 12 DNA samples tested, corresponding to informative combinations (IC) of variable number of polymorphic markers (*x-axis*). **O**: optimum UPIC values for different number of markers in the combination. **A**: minimum number of markers (3) in an IC that can discriminate the 12 DNA samples, the 3 markers can detect up to 18 unique patterns (UP) or alleles; **B** and **C**: point to IC of 4 and 6 markers (**B, C**) respectively, both providing similar amount of information in UP values; **D**: shows the maximum number of UP (34) detectable by all 8 markers. Numbers on top of the histogram are the actual number of **IC** for **K** number of markers used in the combinations, *i.e.*, for combinations of 5 markers, there are 25 IC out of 70 possible combinations.

© 2009 Biomedical Informatics Publishing Group

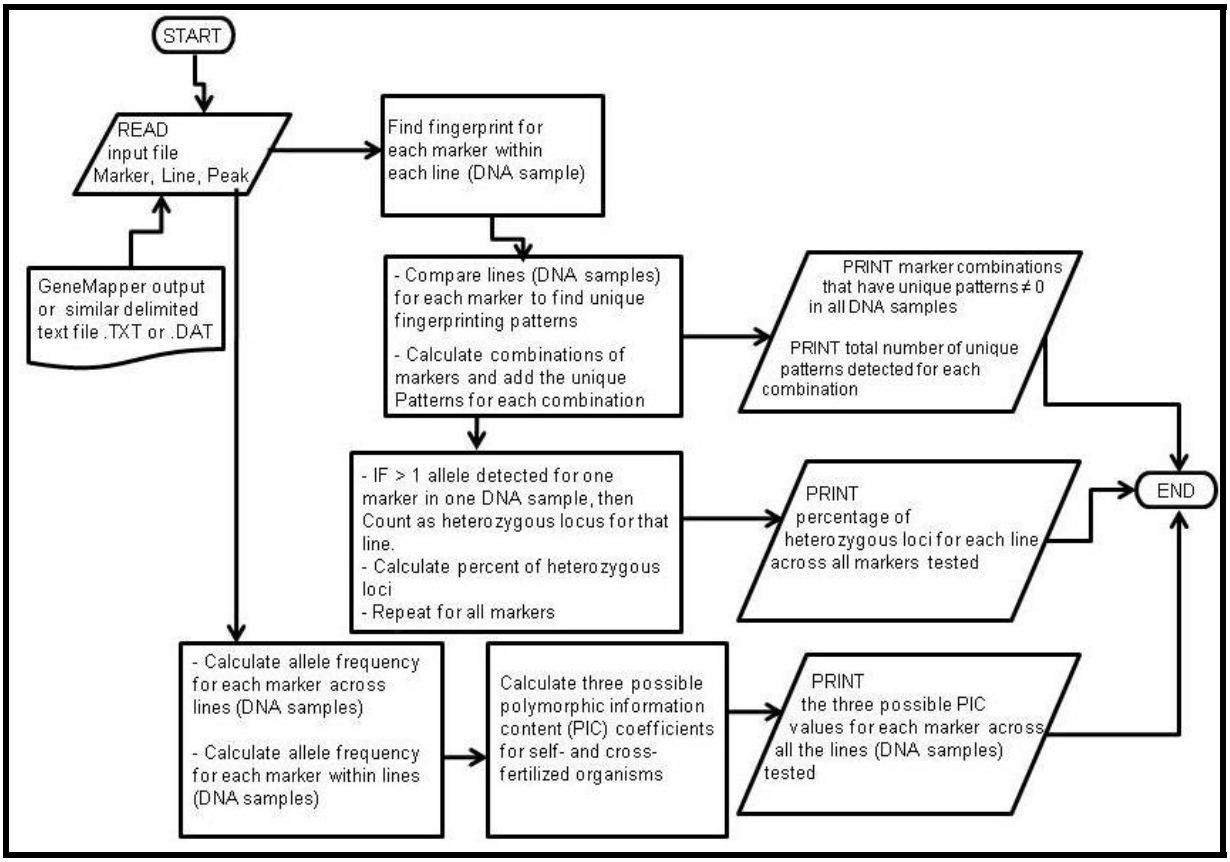

Figure 2: Flow diagram of Perl scripts for calculation of three PIC values, percentage of heterozygous loci and UPIC values.

Conclusion:

We believe that UPIC values will become a very useful tool for planning cost-effective studies using SSR markers. UPIC will minimize the cost of experiments while maximizing the information obtained by polymorphic SSR markers. The users will also be able to choose the number of markers to run based on the obtainable information. In addition to UPIC values, the scripts presented here calculate the percent of heterozygosity of the samples and PIC values for various types of fertilization in populations. Having this information available at a single location in a user-friendly format will also facilitate research with microsatellites.

References:

[1] J. Mrázek *et al*., *Proc Natl Acad Sci*, (2007) 104(20):8472-8477 [PMID: 17485665]
[2] T. Anwar & A.U. Khan, *Bioinformation*, (2005) 1(1):64-68 [PMID: 17597856]
[3] G.F. Richard *et al*., *Microbiol Mol Biol Rev*, (2008) 72(4):686-727 [PMID: 19052325]
[4] R. K. Varshney *et al*., *Trends Plant Sci*, (2005) 10(120):621-630 [PMID: 16290213]
[5] D. Botstein *et al*., *Am J Hum Genet*, (1980) 32:314-331 [PMID: 6247908]
[6] S. Shete *et al*., *Theor Popul Biol*, (2000) 57:265-271 [PMID: 10828218]
[7] Cervus, http://www.fieldgenetics.com/pages/aboutCervus_Functions.jsp
[8] T. C. Marshall *et al*., *Mol Ecol*, (1998) **7**: 639-655 [PMID: 9633105]
[9] PIC calculator, http://www.liv.ac.uk/~kempsj/pic.html
[10] M. Nei & A. K. Roychoudhury, *Genetics* (1974) 76:379-390 [PMID: 4822472]
[11] B. Hansson & L. Westerberg, *Mol Ecol*, (2002) 11:2467-2474 [PMID: 12453232]
[12] R. Bruvo *et al*., *Mol Ecol* (2004) 13:2101-2106 [PMID: 15189230]

Edited by P. Kangueane
Citation: Arias et al. Bioinformation 3(8): 352-360 (2009)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

© 2009 Biomedical Informatics Publishing Group

Supplementary material:

Formulas S1: Formulas used to calculate three polymorphic information content values, percentage of heterozygous loci and optimum combination of polymorphic markers, UPIC values.

- Polymorphic information content [5], where p_i is the frequency of the i^{th} allele, j is the j^{th} line (DNA sample or taxonomic unit) and n is the number of alleles for the marker.

$$PIC=1-\sum_{i=1}^{n} p_i^2-\sum_{i=1}^{n}\sum_{j=i+1}^{n} 2p_i^2p_j^2 \qquad \text{Formula 1.1.1}$$

- For the particular case of cross fertilized organisms that have equifrequent alleles, the formula can be simplified [6]. Same variable notation as in formula 1.1.1.

$$PIC=1-\sum_{i=1}^{n} p_i^2-\sum_{i=1}^{n} p_i^4 \qquad \text{Formula 1.1.2}$$

- In case of self fertilized organisms, or x-chromosome-linked markers in humans, this third term becomes "zero", and the PIC value is identical to the heterozygosity of the marker [6]. Same variable notation as in formula 1.1.1.

$$PIC=1-\sum_{i=1}^{n} p_i^2 \qquad \text{Formula 1.1.3}$$

- Percentage of heterozygous loci, where H_j is the number of heterozygous loci for the *jth* line (DNA sample), m is the total number of loci or markers, and M_j is the number of loci for which no alleles were detected (missing data) in the *jth* line.

$$\%\ \text{Heterozygous loci}j=[H_j/m-M_j]\text{x}100 \qquad \text{Formula 1.2.1}$$

UPIC calculation

Allele information of eight polymorphic markers ($m_1,\ldots,m_8$) that were run on 12 lines (DNA samples) ($j_1,\ldots,j_{12}$) was used in our example to show the mechanics of calculating unique-pattern informative combinations (UPIC). If we assign a letter to each pattern observed for a line (**Table S2b)** and then convert the letters to binary values, where "0" is assigned to an allele pattern present more than once across the lines tested, and "1" is assigned to unique patterns (UP) we obtain **Table S2c**. Let UP_{wj} represent the unique pattern indicator for the wth marker and the jth line shown in **Table S2c**. The number of combinations of markers that can be formed is calculated by the formula below, where m is the number of polymorphic markers taken in groups of k, and $k!$ is k factorial

$${}^{m}C_k = \frac{m(m-1)(m-2)\ldots\text{to } k \text{ factors}}{k!}$$

Our Perl script that calculate UPIC values only includes possible combinations of m polymorphic markers taken in groups of k, for $k = 2$ to $k = m/2$, as $k = m$ for large values of m will be computationally intensive.

$$\sum_{k=2\rightarrow m/2} {}^{m}C_k = {}^{m}C_2+{}^{m}C_3+{}^{m}C_4+\ldots{}^{m}C_{m/2}$$

Let N=Number of possible combinations as defined above:

for m=8:

$$= {}^{8}C_2+{}^{8}C_3+{}^{8}C_4=\frac{8.7}{2!}+\frac{8.7.6}{3!}+\frac{8.7.6.5}{4!}=28+56+70=154 \text{ possible combinations } {}^{m}C_k$$

For each combination of markers, **Table S2d** shows:

- Combined unique patterns ($CUP_{i,j}$) that characterize each line calculated as follows:

$$CUP_{i,j} = \sum_{W^*} C_{w,j}$$

Where i= ith set of markers, j=jth line, W* indicates sum over $C_{w,j}$ (**Table S2c**) values for all marker in ith set of markers.

- Number of Combined Unique Patterns NUP_i

$$NUP_i = \sum_{j=1}^{l} CUP_{i,j}$$

- Informative combinations (IC_i),

IC_i = True if $CUP_{i,j} \neq 0$ for all lines(j), meaning those marker combinations that allow unique identification of each of the lines tested.

Since various IC with different total number of UP can be found, our UPIC script output consists of two columns, one is the total number of UPIC (*i.e.*, 18, **Table S2d**) in the combination, and the other is the marker combination (*i.e.*, m_3, m_4, m_8). All UP values of each IC, UPIC plot, for the data in our example are plotted in **Figure 1**.

Bioinformation, an open access forum
© 2009 Biomedical Informatics Publishing Group

Table S1: INPUT FILE for both scripts, UPIC and PIC/Heterozygous loci
I = Dye/Peak-order; II – Marker; III = Line or DNA sample; IV = amplicon size

I	II	III	IV	I	II	III	IV	I	II	III	IV
B,19	0094_a	Chn1	176	B,18	0015_a	Chn10	206	B,23	0284_a	Chn11	124
B,8	0094_a	Chn10	161	B,30	0015_a	Chn11	206	B,26	0284_a	Chn11	131
B,21	0094_a	Chn11	161	B,19	0015_a	Chn12	206	B,28	0284_a	Chn11	135
B,9	0094_a	Chn12	161	B,18	0015_a	Chn2	201	B,11	0284_a	Chn12	131
B,9	0094_a	Chn2	176	B,19	0015_a	Chn2	206	B,12	0284_a	Chn12	135
B,10	0094_a	Chn3	176	B,17	0015_a	Chn3	201	B,9	0284_a	Chn12	123
B,10	0094_a	Chn4	188	B,18	0015_a	Chn3	206	B,22	0284_a	Chn2	124
B,15	0094_a	Chn5	176	B,15	0015_a	Chn4	201	B,24	0284_a	Chn2	131
B,11	0094_a	Chn6	183	B,16	0015_a	Chn4	207	B,25	0284_a	Chn2	136
B,19	0094_a	Chn8	161	B,17	0015_a	Chn4	212	B,10	0284_a	Chn3	124
B,10	0094_a	Chn9	159	B,25	0015_a	Chn5	201	B,12	0284_a	Chn3	131
B,11	0094_a	Chn9	161	B,26	0015_a	Chn5	206	B,14	0284_a	Chn3	136
B,15	0094_a	Osn7	151	B,14	0015_a	Chn6	201	B,10	0284_a	Chn4	124
B,17	0094_a	Osn7	155	B,15	0015_a	Chn6	207	B,12	0284_a	Chn4	131
B,20	0094_a	Osn7	163	B,24	0015_a	Chn8	123	B,14	0284_a	Chn4	135
B,21	0094_a	Osn7	165	B,33	0015_a	Chn8	206	B,25	0284_a	Chn5	124
B,23	0094_a	Osn7	169	B,17	0015_a	Chn9	206	B,27	0284_a	Chn5	131
B,17	0114_a	Chn1	160	B,20	0015_a	Osn7	114	B,28	0284_a	Chn5	136
B,12	0114_a	Chn10	164	B,25	0015_a	Osn7	166	B,12	0284_a	Chn6	124
B,17	0114_a	Chn11	165	B,32	0015_a	Osn7	207	B,14	0284_a	Chn6	131
B,9	0114_a	Chn12	165	B,23	0076_a	Chn1	118	B,16	0284_a	Chn6	135
B,11	0114_a	Chn2	160	B,28	0076_a	Chn1	133	B,25	0284_a	Chn8	124
B,9	0114_a	Chn3	160	B,14	0076_a	Chn10	112	B,27	0284_a	Chn8	131
B,9	0114_a	Chn4	160	B,33	0076_a	Chn11	112	B,28	0284_a	Chn8	134
B,11	0114_a	Chn5	160	B,14	0076_a	Chn12	112	B,29	0284_a	Chn8	136
B,9	0114_a	Chn6	165	B,13	0076_a	Chn2	124	B,10	0284_a	Chn9	135
B,25	0114_a	Chn8	164	B,14	0076_a	Chn3	127	B,8	0284_a	Chn9	124
B,11	0114_a	Chn9	165	B,16	0076_a	Chn3	133	B,9	0284_a	Chn9	131
B,23	0114_a	Osn7	164	B,16	0076_a	Chn4	136	B,33	0284_a	Osn7	69
B,24	0114_a	Osn7	166	B,19	0076_a	Chn5	124	B,35	0284_a	Osn7	225
B,20	0124_a	Chn1	132	B,16	0076_a	Chn6	162	B,33	0350_a	Chn1	173
B,10	0124_a	Chn10	110	B,17	0076_a	Chn6	165	B,17	0350_a	Chn10	194
B,16	0124_a	Chn10	125	B,21	0076_a	Chn8	112	B,18	0350_a	Chn10	196
B,17	0124_a	Chn10	132	B,22	0076_a	Chn8	115	B,15	0350_a	Chn11	172
B,17	0124_a	Chn11	110	B,14	0076_a	Chn9	112	B,16	0350_a	Chn11	181
B,23	0124_a	Chn11	125	B,15	0076_a	Chn9	115	B,12	0350_a	Chn12	172
B,10	0124_a	Chn12	110	B,18	0076_a	Osn7	105	B,10	0350_a	Chn2	177
B,15	0124_a	Chn12	125	B,34	0194_a	Chn1	167	B,12	0350_a	Chn3	173
B,12	0124_a	Chn2	111	B,12	0194_a	Chn10	158	B,14	0350_a	Chn3	177
B,17	0124_a	Chn2	132	B,34	0194_a	Chn11	158	B,11	0350_a	Chn4	173
B,10	0124_a	Chn3	132	B,16	0194_a	Chn12	158	B,14	0350_a	Chn4	181
B,14	0124_a	Chn4	132	B,12	0194_a	Chn2	167	B,18	0350_a	Chn5	171
B,9	0124_a	Chn4	110	B,23	0194_a	Chn3	167	B,19	0350_a	Chn5	177
B,15	0124_a	Chn5	110	B,15	0194_a	Chn4	167	B,13	0350_a	Chn6	175
B,19	0124_a	Chn5	132	B,16	0194_a	Chn4	172	B,29	0350_a	Chn8	196
B,14	0124_a	Chn6	132	B,14	0194_a	Chn5	167	B,17	0350_a	Chn9	196
B,9	0124_a	Chn6	110	B,35	0194_a	Chn6	167	B,13	0350_a	Osn7	121
B,19	0124_a	Chn8	110	B,32	0194_a	Chn8	158	B,9	0284_a	Chn12	123
B,22	0124_a	Chn8	125	B,13	0194_a	Chn9	158	B,22	0284_a	Chn2	124
B,10	0124_a	Chn9	110	B,21	0194_a	Osn7	168	B,24	0284_a	Chn2	131
B,15	0124_a	Chn9	125	B,22	0194_a	Osn7	175	B,25	0284_a	Chn2	136
B,17	0124_a	Chn9	134	B,23	0284_a	Chn1	124	B,10	0284_a	Chn3	124
B,24	0124_a	Osn7	110	B,25	0284_a	Chn1	131	B,12	0284_a	Chn3	131
B,25	0124_a	Osn7	119	B,26	0284_a	Chn1	136	B,14	0284_a	Chn3	136
B,30	0124_a	Osn7	132	B,10	0284_a	Chn10	131	B,10	0284_a	Chn4	124
B,27	0015_a	Chn1	206	B,11	0284_a	Chn10	135	B,12	0284_a	Chn4	131

© 2009 Biomedical Informatics Publishing Group

Bioinformation by Biomedical Informatics Publishing group *open access*
www.bioinformation.net **Software**

Table S2: UPIC: Calculation of Unique Pattern Informative Combination of polymorphic markers.

	Amplicon	Line 1	Line 2	Line 3	Line 4	Line 5	Line 6	Line 7	Line 8	Line 9	Line 10	Line 11	Line 12
Marker 15	114												0.333
a	123							0.5					
	166												0.333
	201		0.5	0.5	0.333	0.5	0.5						
	206	1	0.5	0.5		0.5		0.5	1	1	1	1	
	207				0.333		0.5						0.333
	212				0.333								

	MARKERS	Marker order	PATTERNS: identical patterns within each marker are identified by the same letter.											
b	94	1	A	A	A	B	A	C	D	E	D	D	D	F
	114	2	A	A	A	A	A	B	C	B	C	B	B	D
	124	3	A	B	A	C	C	C	D	E	F	D	D	G
	15	4	A	B	B	C	B	D	E	A	A	A	A	F
	76	5	A	B	C	D	B	E	F	F	G	G	G	H
	194	6	A	A	A	B	C	C	D	D	D	D	D	E
	284	7	A	A	A	B	A	B	C	B	D	B	E	F
	350	8	A	B	C	D	E	F	G	G	H	I	J	K

	MARKERS	Marker order	Only unique patterns are replaced by # "1", others by "zero"											
c	94	1	0	0	0	1	0	1	0	1	0	0	0	1
	114	2	0	0	0	0	0	0	0	0	0	0	0	1
	124	3	0	1	0	0	0	0	0	1	1	0	0	1
	15	4	0	0	0	1	0	1	1	0	0	0	0	1
	76	5	1	0	1	1	0	1	0	0	0	0	0	1
	194	6	0	0	0	1	0	0	0	0	0	0	0	1
	284	7	0	0	0	0	0	0	1	0	1	0	1	1
	350	8	1	1	1	1	1	1	0	0	1	1	1	1

	Number of UPIC	Marker combination	Examples of combinatorial addition of patterns. Non-zero additions are IC											
d	-	1+2	0	0	0	1	0	1	0	1	0	0	0	2
	-	1+7	0	0	0	1	0	1	1	1	1	0	1	2
	-	2+3	0	1	0	0	0	0	0	1	1	0	0	2
	-	2+6	0	0	0	1	0	0	0	0	0	0	0	2
	18	3+4+8	1	2	1	2	1	2	1	1	2	1	1	3
	23	3+4+5+8	2	2	2	3	1	3	1	1	2	1	1	4
	27	1+3+4+5+8	2	2	2	4	1	4	1	2	2	1	1	5

© 2009 Biomedical Informatics Publishing Group

Table S3: UPIC values, output of UPIC version 1.0. The first column corresponds to Unique Patterns (UP) or alleles detected by the marker combination. Combinations of Markers are listed ONLY if they can discriminate ALL DNA samples, we define these combinations as "Informative Combinations" (IC), and their corresponding value UP is the number of patters or alleles that the combination detects. The user can choose HOW MANY markers to run according to his/her budget, and then finds the combination with the highest UP value. Example of 8 polymorphic markers runs for 12 DNA samples. Note: the combination of markers 350, 15, 76 and 124 can discriminate the 12 DNAs showing 23 different alleles or patterns.

UPIC		M1 M2 M3 M4 M5 M6 M7 M8
2_Marker_Combinations	18	0015_a 0124_a 0350_a
	18	0094_a 0015_a 0350_a
	18	0284_a 0124_a 0350_a
	18	0284_a 0094_a 0350_a
3_Marker_Combinations	23	0350_a 0015_a 0076_a 0124_a
	23	0094_a 0350_a 0284_a 0076_a
	23	0350_a 0284_a 0076_a 0124_a
	23	0094_a 0350_a 0015_a 0076_a
	22	0094_a 0350_a 0015_a 0124_a
	22	0350_a 0284_a 0015_a 0124_a
	22	0094_a 0350_a 0284_a 0124_a
	22	0094_a 0350_a 0284_a 0015_a
	20	0194_a 0094_a 0350_a 0015_a
	20	0194_a 0094_a 0350_a 0284_a
	20	0194_a 0350_a 0015_a 0124_a
	20	0194_a 0350_a 0284_a 0124_a
	19	0350_a 0114_a 0015_a 0124_a
	19	0094_a 0350_a 0114_a 0015_a
	19	0350_a 0114_a 0284_a 0124_a
	19	0094_a 0350_a 0114_a 0284_a
4_Marker_Combinations		
5_Marker_Combinations	27	0094_a 0015_a 0350_a 0076_a 0124_a
	27	0094_a 0284_a 0350_a 0076_a 0124_a
	27	0094_a 0284_a 0015_a 0350_a 0076_a
	27	0284_a 0015_a 0350_a 0076_a 0124_a
	26	0094_a 0284_a 0015_a 0350_a 0124_a
	25	0194_a 0015_a 0350_a 0076_a 0124_a
	25	0194_a 0284_a 0350_a 0076_a 0124_a
	25	0194_a 0094_a 0284_a 0350_a 0076_a
	25	0194_a 0094_a 0015_a 0350_a 0076_a
	24	0015_a 0350_a 0114_a 0076_a 0124_a
	24	0094_a 0284_a 0350_a 0114_a 0076_a
	24	0194_a 0094_a 0284_a 0015_a 0350_a
	24	0284_a 0350_a 0114_a 0076_a 0124_a
	24	0194_a 0094_a 0284_a 0350_a 0124_a
	24	0194_a 0284_a 0015_a 0350_a 0124_a
	24	0194_a 0094_a 0015_a 0350_a 0124_a
	24	0094_a 0015_a 0350_a 0114_a 0076_a
	23	0094_a 0284_a 0350_a 0114_a 0124_a
	23	0094_a 0015_a 0350_a 0114_a 0124_a
	23	0284_a 0015_a 0350_a 0114_a 0124_a
	23	0094_a 0284_a 0015_a 0350_a 0114_a
	21	0194_a 0284_a 0350_a 0114_a 0124_a
	21	0194_a 0094_a 0284_a 0350_a 0114_a
	21	0194_a 0094_a 0015_a 0350_a 0114_a
	21	0194_a 0015_a 0350_a 0114_a 0124_a
6_Marker_Combinations	31	0350_a 0284_a 0015_a 0094_a 0076_a 0124_a
	29	0350_a 0284_a 0015_a 0194_a 0076_a 0124_a
	29	0350_a 0015_a 0194_a 0094_a 0076_a 0124_a
	29	0350_a 0284_a 0194_a 0094_a 0076_a 0124_a
	29	0350_a 0284_a 0015_a 0194_a 0094_a 0076_a
	28	0350_a 0015_a 0094_a 0114_a 0076_a 0124_a
	28	0350_a 0284_a 0015_a 0114_a 0076_a 0124_a
	28	0350_a 0284_a 0015_a 0194_a 0094_a 0124_a
	28	0350_a 0284_a 0015_a 0094_a 0114_a 0076_a
	28	0350_a 0284_a 0094_a 0114_a 0076_a 0124_a
	27	0350_a 0284_a 0015_a 0094_a 0114_a 0124_a
	26	0350_a 0284_a 0194_a 0114_a 0076_a 0124_a
	26	0350_a 0015_a 0194_a 0114_a 0076_a 0124_a
	26	0350_a 0015_a 0194_a 0094_a 0114_a 0076_a
	26	0350_a 0284_a 0194_a 0094_a 0114_a 0076_a
	25	0350_a 0015_a 0194_a 0094_a 0114_a 0124_a
	25	0350_a 0284_a 0015_a 0194_a 0114_a 0124_a
	25	0350_a 0284_a 0194_a 0094_a 0114_a 0124_a
	25	0350_a 0284_a 0015_a 0194_a 0094_a 0114_a
7_Marker_Combinations	33	0094_a 0350_a 0015_a 0194_a 0076_a 0124_a 0284_a
	32	0094_a 0350_a 0015_a 0114_a 0076_a 0124_a 0284_a
	30	0350_a 0015_a 0194_a 0114_a 0076_a 0124_a 0284_a
	30	0094_a 0350_a 0015_a 0194_a 0114_a 0076_a 0284_a
	30	0094_a 0350_a 0015_a 0194_a 0114_a 0076_a 0124_a
	30	0094_a 0350_a 0194_a 0114_a 0076_a 0124_a 0284_a
	29	0094_a 0350_a 0015_a 0194_a 0114_a 0124_a 0284_a
8_Marker_Combinations	34	0015_a 0194_a 0114_a 0284_a 0350_a 0094_a 0076_a 0124_a

Bioinformation, an open access forum
© 2009 Biomedical Informatics Publishing Group

Table S4: Output for PIC values and Percentage of Heterozygous Loci. The five columns on the bottom correspond to: marker, square of the allele frequency, PIC value for self fertilized organisms, PIC value of equifrequent alleles, and PIC value for cross fertilized organisms with variable allele number.

DNA Sample	% Heterozygous Loci
1	0.25
2	0.375
3	0.5
4	0.625
5	0.5
6	0.5
7	0.78
8	0.5
9	0.5
10	0.375
11	0.375
12	0.25

marker	sum_Fsquare	picval-self_fert	picval_equifreq	picval_cross_fertil.
0015_a	0.3191589	0.6808411	0.578978697	0.568978697
0076_a	0.16666666	0.83333334	0.805555564	0.795555564
0094_a	0.26847222	0.73152778	0.659450447	0.649450447
0114_a	0.32986111	0.67013889	0.561330538	0.551330538
0124_a	0.2982253	0.7017747	0.61283637	0.60283637
0194_a	0.3888888	0.6111112	0.459876701	0.449876701
0284_a	0.20476466	0.79523534	0.753306774	0.743306774
0350_a	0.1388888	0.8611112	0.841821101	0.831821101

© 2009 Biomedical Informatics Publishing Group