Vol. 23 no. 16 2007, pages 2174–2177
doi:10.1093/bioinformatics/btm296

Databases and Ontologies

# PIP: a database of potential intron polymorphism markers

Long Yang, Gulei Jin, Xiangqian Zhao, Yan Zheng, Zhaohua Xu and Weiren Wu*
Department of Agronomy, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, 310029, China

Received on March 18, 2007; revised on May 12, 2007; accepted on May 25, 2007
Advance Access publication June 1, 2007
Associate Editor: Alfanso Valencia

## ABSTRACT

**Motivation:** With the recent progress made in large-scale plant functional genome sequencing projects, a great amount of EST (express sequence tag) data is becoming available. With the help of complete genomic sequence information of model plants (rice and Arabidopsis), it is possible to predict the joints between adjacent exons after splicing (or termed 'intron positions' for short) in homologous ESTs of other plants. This would allow developing potential intron polymorphism (PIP) markers in these plants by designing primers in exons flanking the target intron.

**Results:** We have extracted a total of 57 658 PIP markers in 59 plant species and created a web-based database platform named PIP to provide detailed information of these PIP markers and homologous relationships among PIP markers from different species. The platform also provides a function of online designing of PIP markers based on cDNA/EST sequences submitted by users. With evaluations performed *in silico*, we have found that the intron position prediction is highly reliable and the polymorphism level of PIP markers is high enough for practical need.

**Availability:** http://ibi.zju.edu.cn/pgl/pip/

**Contact:** wuwr@zju.edu.cn

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Molecular markers are powerful tools for genetic research and breeding. Many types of molecular marker have been developed since 1980, such as restriction fragment length polymorphism (RFLP; Botstein *et al*., 1980), random amplified polymorphic DNA (RAPD; Williams *et al*., 1990), amplified fragment length polymorphism (AFLP; Vos *et al*., 1995), simple sequence repeat polymorphism (SSR; Becker and Heun, 1995), single-nucleotide polymorphism (SNP; Kruglyak, 1997) and intron length polymorphism (ILP; Choi *et al*., 2004).

Introns are noncoding sequences interspersed in genes. In comparison with exons, introns are more variable because in general selective pressure in intronic regions is much less than exonic regions. For example, the average number of SNPs per 1000 bp in introns (12.1) is over three times as high as that in exons (3.6) among eight varieties in rice (*Oryza sativa* L.; Feltus *et al*., 2006). Length polymorphism is the most intuitive variation in introns. So, ILPs have been exploited as molecular markers, which have many desirable properties, including specific, codominant, neutral, convenient and reliable. However, ILP has not been widely utilized because it is newly developed as a kind of molecular marker and the number of ILP markers having been exploited in plants is still very limited. To date, studies of exploiting ILP markers have been restricted to a few species (Choi *et al*., 2004; Wang *et al*., 2005; Wei *et al*., 2005). ILP is detected by PCR with primers designed on exons flanking the target intron. Obviously, the key point of developing ILP markers is to identify suitable introns. A general method for identifying introns is to compare cDNA/EST sequences with the genome sequence. Therefore, introns can be easily identified in many model organisms. However, this method is not applicable to most other organisms because they have only cDNA/EST sequences available. Fortunately, studies have indicated that the exon-intron structures are largely conserved among homologous genes from different species (Batzoglou *et al*., 2000). Therefore, the joints between adjacent exons after splicing (or termed 'intron positions' for short) in a cDNA/EST of an organism can be deduced according to the homologous genes from related model organisms. This provides a way of developing ILP markers in any organisms. Nowadays, many organisms have got a large number of cDNA/EST sequences available in public databases. Therefore, large-scale exploitation of ILP markers in various organisms becomes possible.

Recently, Fredslund (Fredslund *et al*., 2006) developed a web program GeMprospector that allows automatically designing cross-species candidate ILP markers in legumes or grasses. The main point of their work is to develop conserved ILP primers that are usable in different species of legumes or grasses. Such cross-species ILP markers would be useful for the research of comparative genomics. However, the requirement of primer conservation would probably make many ILPs not exploitable as markers. There could be two reasons: (i) the flanking exon sequences of some introns are not conserved enough for designing cross-species primers and (ii) many ESTs do not have homologs available in multiple species at present so that the conserved regions in these ESTs for designing primers cannot be identified.

Aiming at developing specific intron-based markers in individual species instead of emphasizing their cross-species utility, we adopt a different strategy. For any plant, we use a model plant (rice or Arabidopsis) to predict intron positions in its cDNA/EST sequences and then design a pair of primers on both sides of each intron position. These specific primers would

*To whom correspondence should be addressed.

potentially detect ILPs in the target plant. In addition, these specific primers would also allow detecting other potential polymorphisms (e.g. SNP) in introns in the target plant. Therefore, we may call them potential intron polymorphism (PIP) markers. Obviously, our strategy would ensure high specificity of PIP primers and allow developing more PIP markers in the target plant. In addition, although the PIP primers developed in a species are not necessarily usable in other species, correspondence among PIP markers from different plant species can be identified through their common homologous genes of the model plant. Therefore, PIP markers would be also useful for comparative genomic studies. Following this strategy, we have created a web-based database (http://ibi.zju.edu.cn/pgl/pip/), which contains a total of 57 658 PIP markers in 59 plant species and provides a function for online designing of PIP markers based on cDNA/EST sequences submitted by users.

## 2 DATABASE CONTENT

### 2.1 Sources of sequence data

The dicot model plant Arabidopsis and monocot model plant rice were taken as subject species; all other plants with available EST sequence data were taken as query species. The genome, cDNA and CDS (coding sequence) data of rice (*Oryza sativa* L. ssp. japonica cv. Nipponbare) and Arabidopsis (*Arabidopsis thaliana* ecotype Columbia) were downloaded from http://www.tigr.org/ and http://www.arabidopsis.org/, respectively. The EST sequences of 59 plant species (including 47 dicots and 12 monocots; Supplementary Table 1) were downloaded from http://www.plantgdb.org/(PUT-155a edition; Dong *et al.*, 2005). Each of these species had at least 10 000 PUTs (plantGDB-assembled unique transcripts) in the database.

### 2.2 Development of PIP markers

A pipeline in Perl script was developed to develop PIP markers. The procedure consisted of three steps. The first step was to identify intron positions and lengths in each subject species by aligning its CDSs with its genome sequence using program SIM4. A threshold of 100% identity was used in the alignment. The second step (Fig. 1) was to identify possible intron positions in each query species by aligning query EST sequences with subject CDSs using BLASTN (note: monocot plants and dicot plants were compared with rice and Arabidopsis, respectively). A query EST was thought to be homologous to a subject CDS only if there were at least 200 bp overlapping and 80% similarity between them. The third step (Fig. 1) was to design primers for those query ESTs containing possible positions of introns that were expected not to be longer than 400 bp according to the subject species. For each of those query ESTs, a pair of primers was designed using program ePrimer3 (Rozen and Skaletsky, 2000) on a 200 bp sequence cut from the query EST with 100 bp on each side of the target intron. The designed primers were tested by electronic PCR (e-PCR) on the EST sequences of corresponding species. A putative intron was taken as a PIP marker if the e-PCR yielded the unique product as expected. A PIP marker was named with a four-letter abbreviation of the Latin name of the corresponding species (one from the first letter of the genus name and three from the first three letters of the species name, e.g. Acep for *Allium cepa*) followed by 'PIP' and a unique number (e.g. AcepPIP318). In total, we extracted 57 658 PIP markers in the 59 plant species (Supplementary Table 1), with 22 273 in monocots and 35 385 in dicots. The number of PIP markers in one species varied from 81 (silverleaf sunflower) to 4314 (sorghum), with an average of 977. Detailed information of those PIP markers can be found from http://ibi.zju.edu.cn/pgl/pip.

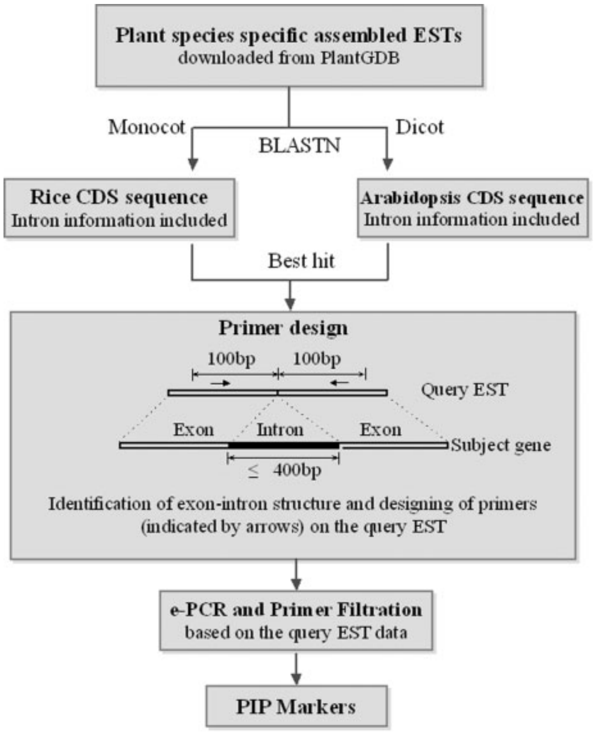
Plant species specific assembled ESTs
downloaded from PlantGDB
Monocot
BLASTN
Dicot
Rice CDS sequence
Intron information included
Arabidopsis CDS sequence
Intron information included
Best hit
Primer design
100bp
100bp
Query EST
Exon
Intron
Exon
Subject gene
≤ 400bp
Identification of exon-intron structure and designing of primers
(indicated by arrows) on the query EST
e-PCR and Primer Filtration
based on the query EST data
PIP Markers

**Fig. 1.** Flowchart of developing PIP markers in a query species.

### 2.3 Query interface

A query interface is provided for searching PIP markers on the Search web page. Users can enter any of the following information in corresponding boxes to condition the search: (i) species name (if only this information is entered, all the PIP markers of the selected species will be presented); (ii) marker ID (either a full or a partial ID of a marker, e.g. AcepPIP318, acep or 318); (iii) gene name in subject species; (iv) PlantGDB PUT ID; (v) intron length range in subject species; (vi) Gene Ontology (GO) number in subject species and (vii) gene description in subject species. The result page(s) will present the list of searched markers and some important information for each marker including the corresponding gene in Arabidopsis/rice, forward and reverse primers and their melting temperatures, predicted size of PCR product without intron and so on. More detailed information of each PIP marker can be found by clicking on the marker's name. Users can also download all the PIP makers of one or all species from the Download web page.

To facilitate comparative genomic research, corresponding PIP markers with the same or similar primers in other species are also presented in the search result of each PIP marker (see Section 3.3). In addition, a function is provided on the Compare web page for comparing PIP markers between any two monocot or dicot species, from which markers in the two compared species corresponding to the same genes of the model plant will be presented.

### 2.4 Developing PIP markers online

Online development of PIP markers can be performed on the Develop web page. Users can submit query EST sequences in FASTA format and select the type of the query plant (monocot or dicot), and then click on the button 'Submit'. The result will provide the following information: query EST sequences; possible intron positions in the query sequences; intron lengths in the subject species; primer pairs bracketing single introns; primer positions in query sequences and sizes of PCR products without introns in query species.

## 3 EVALUATION OF PIP MARKERS

### 3.1 Conservation of intron position

In our strategy, intron positions in the EST sequences of query plants are predicted according to the homologous genes of model plants. Therefore, the conservation of intron position is a key factor determining the efficiency and reliability of PIP marker exploitation. To evaluate the conservation of intron position in plant, we performed e-PCR both on the genome sequence (downloaded from http://genome.jgi-psf.org/Poptr1/Poptr1.download.html) and on the EST sequences of *Populus trichocarpa* with its PIP primers. The existence of intron in a PIP marker was deduced if the e-PCR product from the genome sequence was larger than that from the EST sequences and then was further confirmed by checking the genome annotation. Under a constraint of 3 kb for the maximum size of e-PCR product, 836 out of 959 primer pairs could obtain e-PCR products from the genome sequence, of which 831 (99.40%) appeared to contain introns as expected. The result indicates that intron position is highly conserved across plant species.

Based on the genome sequence data of *Medicago truncatula* downloaded from http://www.medicago.org/genome/, we also examined the conservation level of intron positions in this species. Because the genome sequence was not complete, only 389 out of 1120 primer pairs yielded e-PCR products from the genome sequence. Nevertheless, among these primer pairs, 384 (98.71%) detected introns. This percentage is very similar to that found in *Populus trichocarpa*. This result further indicates that intron position is highly conserved in plants and therefore using model plants to predict intron positions in other plants is feasible.

### 3.2 Polymorphism level of PIP markers

To evaluate the polymorphism level of PIP markers, we also developed PIP markers in rice and Arabidopsis by taking the two subject species themselves as query species. We examined ILPs by performing e-PCR with the rice and Arabidopsis PIP primers on the genome sequences of two rice cultivars (representing two different subspecies), Nipponbare (japonica) and 93-11 (indica) and of two Arabidopsis ecotypes, Columbia and Landsberg, respectively. For those PIP markers that did not show ILP, we further examined SNP in their introns (or termed intron single nucleotide polymorphism, ISNP) between the two rice cultivars or between the two Arabidopsis ecotypes. The percentages of ILP and ISNP were found to be 17.98% (835/4645) and 51.22% (2379/4645) between the two rice cultivars and 18.61% (2653/14258) and 53.18% (7563/14258) between the two Arabidopsis ecotypes, respectively (Table 1). Interestingly, both the percentage of ILP and the percentage of ISNP are very similar in the two species. The results suggest that the polymorphism level of PIP markers is quite high in plant, which could meet the need of genetic research and plant breeding.

**Table 1.** Polymorphism levels of PIP markers in Arabidopsis and rice

| Species | PIP number | ILP number | ILP % | ISNP number | ISNP% | Total% |
|---|---|---|---|---|---|---|
| Arabidopsis | 14258 | 2653 | 18.61 | 7563 | 53.18 | 71.79 |
| Rice | 4645 | 835 | 17.98 | 2379 | 51.22 | 69.2 |

### 3.3 Utility of PIP markers for comparative genomics

Unlike the method of Fredslund that emphasizes primer conservation across species; our method does not care about primer conservation but only stresses primer specificity within individual species. This determines that our method could develop much more intron-based markers in individual species. For example, Fredslund only developed 312 markers in maize, while we developed 3123 (~9 times more) markers based on similar EST data sets. This should be an advantage of our method.

Although PIP markers are not designed for cross-species applications, they are also useful for comparative genomic studies because, as we have pointed out above, the correspondence between PIP markers from different species can be determined via their common corresponding genes of Arabidopsis or rice. In addition, some PIP markers also possess conserved primers and therefore could serve as cross-species markers. By examining all the 57 658 PIP markers developed in this study, we found that 7462 (12.94%) markers have got completely conserved primers across at least two species each; in other words, there are 3516 primer pairs shared by at least two PIP markers each in different species.

Actually, the proportion of cross-species PIP markers could be far greater than that because practical PCR does not necessarily require primers exactly matching templates. Experiments have shown that rice ILP primers are applicable to other plants (including monocots and dicots) under less stringent PCR conditions (Wang *et al.*, 2005). For this reason, we performed e-PCR with PIP primers from one species on the EST sequences of all other species under a condition of allowing at most two mismatches between a primer and a template. We found that 29 658 (48.50%) PIP markers can obtain e-PCR products in other species. This proportion is quite high. We can expect that the proportion could be even higher if complete EST data sets covering all genes in other species were used as templates for the e-PCR.

## 4 FUTURE WORK

### 4.1 Updating the database

EST sequencing has proven to be an economically feasible alternative for gene discovery in species lacking a draft genome sequence (Masoudi-Nejad *et al*., 2006). With many large-scale EST sequencing projects ongoing, we shall keep updating the database. Apart from adding in new PIP markers of already existed species, we shall also add in the information of new species when their PUT numbers have reached 10 000 or more.

### 4.2 Adding in experimental results

Users can conduct designing experiments (e.g. construction of genetic maps, mapping of genes or quantitative trait loci, gene cloning) using the PIP markers and are encouraged to submit to us any useful information (such as, PCR conditions, polymorphisms and map positions) of the PIP markers they have analyzed. We shall augment the PIP database with submitted information to make it more valuable. We expect that PIP markers will play important roles in the genetic studies and breeding of non-model plants.

## ACKNOWLEDGEMENTS

This work was financially supported by the Hi-Tech Research and Development (863) Program of China (grant no.: 2006AA10Z1E2).

*Conflict of Interest*: none declared.

## REFERENCES

Batzoglou,S. *et al*. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res*., **10**, 950–958.

Becker,J. and Heun,M. (1995) Barley microsatellites: allele variation and mapping. *Plant Mol. Biol*., **27**, 835–845.

Botstein,D. *et al*. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet*., **32**, 314–331.

Choi,H.K. *et al*. (2004) A sequence-based genetic map of Medicago truncatula and comparison of marker colinearity with M. sativa. *Genetics*., **166**, 1463–1502.

Dong,Q. *et al*. (2005) Comparative plant genomics resources at PlantGDB. *Plant Physiol*., **139**, 610–618.

Feltus,F.A. *et al*. (2006) A comparative genomics strategy for targeted discovery of single-nucleotide polymorphisms and conserved-noncoding sequences in orphan crops. *Plant Physiol*., **140**, 1183–1191.

Fredslund,J. *et al*. (2006) GeMprospector–online design of cross-species genetic marker candidates in legumes and grasses. *Nucleic Acids Res*., **34**, W670–W675.

Kruglyak,L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nat. Genet*., **17**, 21–24.

Masoudi-Nejad,A. *et al*. (2006) EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. *Nucleic Acids Res*., **34**, W459–W462.

Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.

Vos,P. *et al*. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res*., **23**, 4407–4414.

Wang,X. *et al*. (2005) Genome-wide investigation of intron length polymorphisms and their potential as molecular markers in rice (Oryza sativa L.). *DNA Res*., **12**, 417–427.

Wei,H. *et al*. (2005) Intron-flanking EST-PCR markers: from genetic marker development to gene structure analysis in Rhododendron. *Theor. Appl. Genet*., **111**, 1347–1356.

Williams,J.G. *et al*. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res*., **18**, 6531–6535.