Database, Vol. 2013, Article ID bas054, doi:10.1093/database/bas054

Database tool

PIPEMicroDB: microsatellite database and primer generation tool for pigeonpea genome

Sarika[1], Vasu Arora[1], M. A. Iquebal[2], Anil Rai[1] and Dinesh Kumar[1,*]

[1]Centre for Agricultural Bioinformatics and [2]Division of Biometrics & Statistical Modelling, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012, India

***Corresponding author:** Tel: +91 94161 11753; Fax: +91 11 2584 1564; Email: dineshkumarbhu@gmail.com **or** dineshkumarbhu@iasri.res.in

Submitted 6 July 2012; Revised 16 November 2012; Accepted 20 November 2012

Citation details: Sarika, Vasu Arora, M. A. Iquebal, Anil Rai, and Dinesh Kumar. PIPEMicroDB: microsatellite database and primer generation tool for pigeonpea genome. *Database* (2012) Vol. 2012: article ID bas054; doi:10.1093/database/bas054.

Molecular markers play a significant role for crop improvement in desirable characteristics, such as high yield, resistance to disease and others that will benefit the crop in long term. Pigeonpea (*Cajanus cajan* L.) is the recently sequenced legume by global consortium led by ICRISAT (Hyderabad, India) and been analysed for gene prediction, synteny maps, markers, etc. We present PIgeonPEa Microsatellite DataBase (PIPEMicroDB) with an automated primer designing tool for pigeonpea genome, based on chromosome wise as well as location wise search of primers. Total of 123 387 Short Tandem Repeats (STRs) were extracted from pigeonpea genome, available in public domain using MIcroSAtellite tool (MISA). The database is an online relational database based on 'three-tier architecture' that catalogues information of microsatellites in MySQL and user-friendly interface is developed using PHP. Search for STRs may be customized by limiting their location on chromosome as well as number of markers in that range. This is a novel approach and is not been implemented in any of the existing marker database. This database has been further appended with Primer3 for primer designing of selected markers with left and right flankings of size up to 500 bp. This will enable researchers to select markers of choice at desired interval over the chromosome. Furthermore, one can use individual STRs of a targeted region over chromosome to narrow down location of gene of interest or linked Quantitative Trait Loci (QTLs). Although it is an *in silico* approach, markers' search based on characteristics and location of STRs is expected to be beneficial for researchers.

Database URL: http://cabindb.iasri.res.in/pigeonpea/

Introduction

Pigeonpea or 'tur' is one of the most important pulse crops belonging to family *Fabaceae* and scientifically known as *Cajanus cajan*. It is known to be cultivated in more than 25 countries of the world such as Indian subcontinent, Africa and Central America in purview of the favourable climatic conditions. It is grown on 4.7 million hectares and ranks sixth among grain legumes in production. The world production of this crop figures around 3.69 million tons annually (1). Economic loss due to biotic stress factors has been estimated to be $US 8.48 billion earlier, while abiotic stress may reduce pigeonpea production by 50% (IIPR vision). Pigeonpea is a versatile crop, consumed as a staple food, green vegetable and as a fodder crop besides being a boon for soil. It improves soil fertility through nitrogen fixation as well as from the leaf fall and recycling of the nutrients (2, 3).

Database, Vol. 2013, Article ID bas054, doi:10.1093/database/bas054

Indian economy depends significantly on pulses and India is the largest producer, consumer and importer of pulses. India's pigeonpea production stands at around 2.86 million tons, which is four-fifth share in the world total pigeonpea produced. About 90% of the global pigeonpea area falls in India (1). Pigeonpea is the second largest pulse crop of India accounting for 15.8% of total pulse production (18.09 million tons) during 2010–11 (Source: Ministry of Agriculture, Government of India). The domestic consumption of tur in India is estimated at around 3.4 million tons (4). The production of pigeonpea is not sufficient to satisfy the domestic demand, hence it has to rely on imports of the crop especially from Myanmar and Tanzania. Still the demand per capita is not fulfilled, leading to malnutrition worldwide. To combat this gap between production and consumption, genomics may play a significant role. Researchers in today's world are looking to produce crops that will possess desirable characteristics, such as high yields, resistance to disease and many other characteristics that will benefit the crop in the long term. Genome sequencing has significantly advanced in recent years. Molecular markers can play a significant role in crop improvement. The knowledge of genetic structure based on molecular markers such as microsatellites (5) plays important role in population genetic studies, gene regulation and genome evolution (6–8). These have proved useful in marker-assisted selection of desirable traits to which they are linked, hence are the markers of choice for genome mapping studies. Genome sequencing has contributed greatly to biological science. Till date, many crop genomes have been sequenced, pigeonpea being the recently sequenced genome. There is need to develop platform for mining genic microsatellites to ensure their better utilization as molecular markers, their abundance, distribution, evolution and putative function, if any.

Information available as Supplementary File (URL: http://www.icrisat.org/gt-bt/iipg/sup_files/Table15.html) is not chromosome wise but it is based on scaffold (9). Unless it is chromosome wise, it cannot be directly used in gene or QTL mapping. Here, we present mining STRs chromosome wise along with user-defined primer designing. Though high numbers of STR markers are available in pigeonpea, their effective use is very limited because the users cannot select primers at equal distance over particular chromosome especially in QTL mapping or fine mapping of the genes. Moreover, since there is no publicly available STR marker database with more user-friendly parameters, the present database with add-on tool having user need-based primer designing options is being made available for global community. Our database is unique and user friendly as it designs primer on loci of choice with amplicon range up to 500 bp, further facilitating the low-cost rapid revalidation in simple agrose gel.

Database development

Database flow

The chromosome-wise pigeonpea genome, available in public domain (http://www.icrisat.org/gt-bt/iipg/genome-data.zip), was downloaded in FASTA format. The tool PIgeonPEa Microsatellite DataBase (PIPEMicroDB) stores catalogue of microsatellites fetched from pigeonpea genome. The chromosome-wise sequences were used to extract microsatellite markers using MIcroSAtellite tool (MISA) (10). The output of MISA was processed using PERL scripts. The data were assembled in proper format in order to create the data file which was further imported to MySQL database. The query for STRs may be made chromosome wise along with the microsatellite characteristics such as motif type, repeat motif and repeat kind. Furthermore, the advance search may be made with the range of chromosomal location, GC content, number of base pairs and copy number. For the graphical user interface, PHP was used. The primers are generated using primer3 standalone tool. For primer generation, only STRs have been considered and not scaffolds. The data flow of tool is illustrated in Figure 1.

Database architecture

The PIPEMicroDB is an online relational database that catalogues information about the microsatellite repeats of the recently sequenced pigeonpea. All the microsatellite markers extracted from pigeonpea genome have been generated using MISA. The database architecture is a 'three-tier architecture' (Figure 2) with a client tier, middle tier and database tier. This user-friendly interface for the database has been developed using PHP (Hypertext Preprocessor), which is an open-source server-side scripting language. The microsatellite marker information is accessed from MySQL database. In first tier of the architecture, the *in silico* mined STRs through MISA were stored in MySQL database. In the middleware, user need-based customized

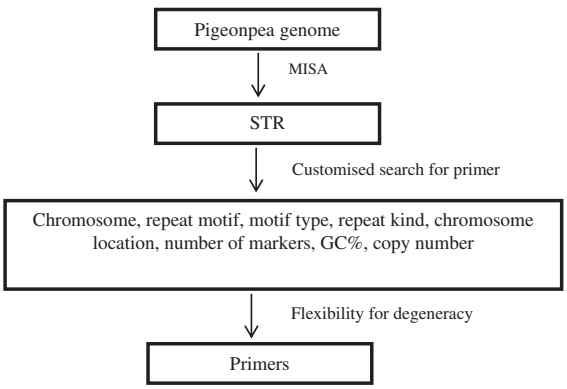

Figure 1. Data exchange flow diagram for PIPEMicroDB.

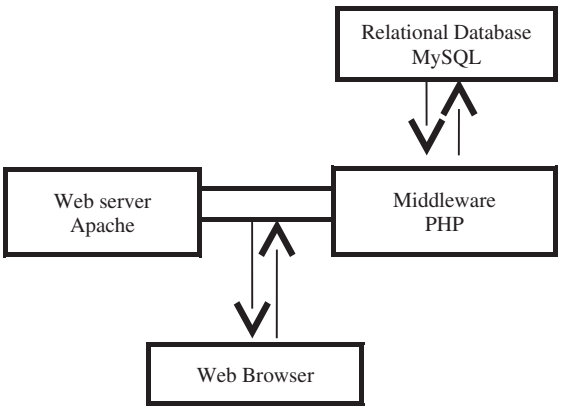

Figure 2. Three-tier architecture of PIPEMicroDB.

query provisions have been made. For primer designing, Primer3 standalone code computes primers on user request. The information generated at the client end, i.e. third tier of the architecture are list of multiple primers along with their respective melting temperature, GC content, start position and product size (amplicon size).

PIPEMicroDB has eight tabs (Home, About, Database, Analysis, Tutorial, Links, Contact and Team). General information of the developed microsatellite database, information about pigeonpea, microsatellite markers and analysis of the pigeonpea genome has been discussed. The tutorial of this database contains the guidelines for users and terminologies used in the database contents. PIPEMicroDB is appended with other useful links, the team and contact persons.

Accessing database

PIPEMicroDB can be accessed to extract microsatellites based on motif type (mono, di, tri, tetra, penta and hexa), repeat motif and repeat kind (perfect and compound). These microsatellites can be mined based on the choice of chromosome where more than one chromosome may also be selected. Also, the flexibility has been given for limiting the search based on chromosomal location as well as the number of markers in that range. This is a novel approach and is not been implemented in any of the existing marker database, which may be useful for the breeders and biotechnologists. The flexibilities provided will enable researcher to select markers of choice at desired interval over the chromosome which may be coding or non-coding. However, if user wishes to know whether they are from coding or non-coding region, it can be traced/verified by our numeric counter showing start and end position of each amplicons over a particular chromosome. Furthermore, one can use each individual STRs of a targeted region over chromosome to narrow down location of gene of interest or linked QTL. PIPEMicroDB also fulfils this customized search according to the requirement of researcher, based on ranges of GC content (%), base pair and copy number.

PIPEMicroDB is further appended with Primer3. After selecting the STR based on customized search with the help of radiobutton, the marker of interest may be further carried for primer designing. The user may go for primer designing as a default setting or may go for modified degenerate bases in the sequence, if present. The selected markers may be designed with left and right flankings of size up to 500 bp. The output generated gives five primers along with the melting temperature, product size and GC content. The novel add-on for degenerate bases incorporated in this database search gives the users flexibility to replace degenerate bases with any of the alternative bases (A, T, G, C). This feature has been added to resolve the issue of some of the degenerate bases present in current pigeonpea genome assembly making the primer designing very difficult otherwise. The gap present in the region over chromosome has been deliberately avoided while designing the primer in flanking region of STR so that there is no inaccuracy in the predicted amplicon size for user's ease in genotyping.

Discussion and conclusion

Utility of the database

PIPEMicroDB is of great use to pigeonpea breeders. The customization of this tool for search based on chromosome may be used for mapping of gene and QTL markers for crop improvement purpose. It is likely to be accessed by biologists engaged in research with diverse objectives in the crop primarily to develop molecular markers and also to understand the functional significance of microsatellites in regulating gene expression and genome evolution. The comprehensive options to search for simple and compound microsatellites repeats in the genic regions allow users to explore new avenues of investigations on these repeats. The primer designing for PCR amplification of desired motifs will facilitate studies on mutability, microsatellite abundance, etc. Association of microsatellites with a particular disease or phenotype may also be explored. Microsatellite data can also be used to investigate various anomalies using candidate gene approach. The database may be used for various research programmes, especially for genome mapping and gene tagging. This microsatellite database will serve as an important application for extracting information in order to design experiments in new directions elucidating novel roles and functions of microsatellites.

Analysis of pigeonpea genome

The whole genome was analysed for getting an overview of the pigeonpea genome. It was observed that 87% and

Database, Vol. 2013, Article ID bas054, doi:10.1093/database/bas054

13% of the STR markers were of simple and compound (includes both simple and interrupted) type, respectively (Figure 3). Also, the 'single' repeat type was found to occur dominantly followed by 'di' type (Figure 4). Figure 5 shows the distribution of mono, di, tri, tetra, penta and hexa STRs in the whole genome of pigeonpea. The GC content between the range 0% and 25% occurs maximum as shown in Figure 6. Table 1 shows the distribution of simple and compound STRs along with repeat types (mono, di, tri, tetra, penta and hexa) on each chromosome.

To make best use of this web-based tool, the wet lab validation of these STRs warranted in order to reap the benefit of genomics in endeavour of improvement in pigeonpea productivity.

STR validations

In silico validation of *in vitro* validated primers was done. The set of STR markers (11) were evaluated in the database using PERL script and the results are presented in Table 2.

Limitations

The total numbers of STR markers reported are 123 387, which needs wet lab validation. Furthermore, the primers designed here can only be used in simple singleplex PCR, not for multiplex PCR used for throughput analysis. However, the loci searched can be of immense use as it has location and chromosome, which is required input for multiplex designing. Relatively higher abundance of mono- over di-nucleotide repeat type might be due to inherent limitation of the 454 pyrosequencing technology which adds more mono-nucleotide causing sequencing error (12).

Availability

PIPEMicroDB can be accessed freely at http://cabindb.iasri.res.in/pigeonpea/.

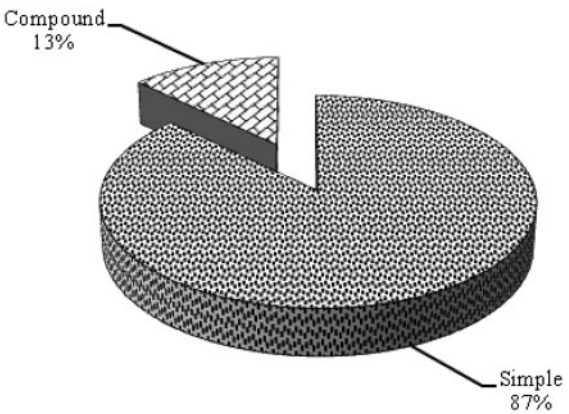

Figure 3. Distribution of STR marker types.

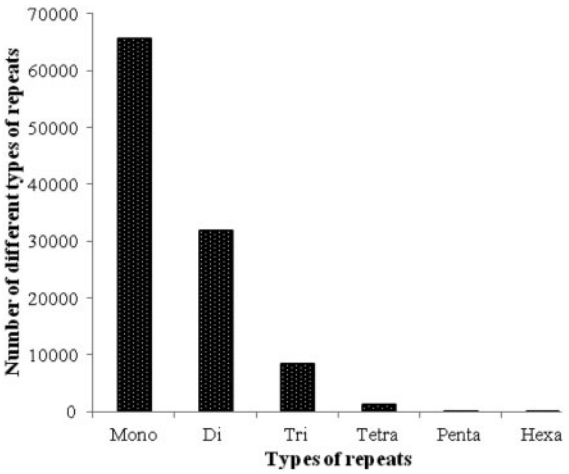

Figure 4. Distribution of repeat types.

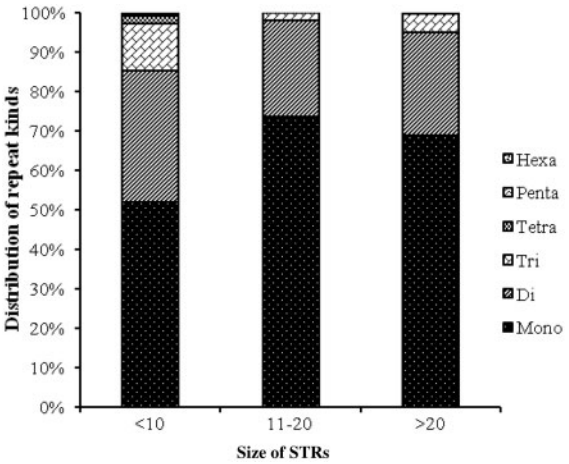

Figure 5. Distribution of repeat kinds.

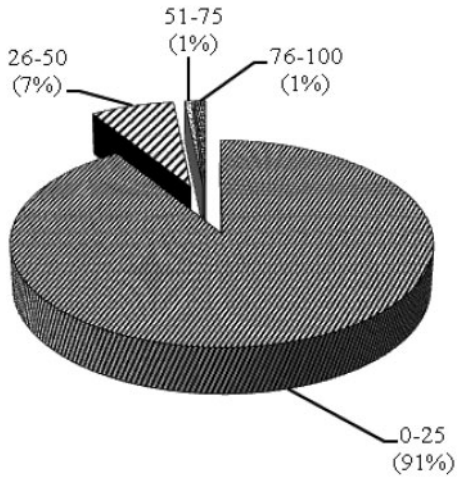

Figure 6. Distribution based on GC content.

Table 1. Chromosome wise distribution of STRs

Chromosomes	Simple						Compound
	Mono	Di	Tri	Tetra	Penta	Hexa	
Chromosome 1	4771	2255	628	91	21	21	1142
Chromosome 2	9827	4918	1339	173	39	55	2492
Chromosome 3	8183	3841	1019	124	29	37	1941
Chromosome 4	3581	1620	440	69	11	19	808
Chromosome 5	1429	668	173	30	4	9	337
Chromosome 6	6757	3231	834	115	29	32	1569
Chromosome 7	5145	2568	712	113	14	15	1261
Chromosome 8	5421	2644	720	103	19	25	1286
Chromosome 9	2788	1318	330	43	7	10	630
Chromosome 10	5469	2685	724	97	15	20	1340
Chromosome 11	12266	6018	1581	218	50	62	2959

Table 2. STR validation result of primers from pigeonpea

	Kumawat *et al.* (11)
Total no. of primers reported	28
No. of positive primers (forward)	11
No. of positive primers (reverse)	12
No. of positive primers (common to both forward and reverse)	11

Acknowledgements

The technical assistance of Jai Bhagwan in maintaining the web server and A.R. Paul in designing the logo of PIPE MicroDB is thankfully acknowledged. The authors acknowledge the critical input of all the four anonymous reviewers and associate editor in improvement of the manuscript.

Funding

National Agricultural Innovation Project, Indian Council of Agricultural Research, New Delhi, Ministry of Agriculture, Govt. of India.

Conflict of interest. None declared.

References

1. http://faostat.fao.org (15 June 2012, date last accessed).
2. Snapp,S.S., Rohrbach,D.D., Simtowe,F. *et al*. (2002) Sustainable soil-management options for Malawi: can smallholder farmers grow more legumes? *Agri. Ecosys. Environ.*, **91**, 159–174.
3. Mapfumes,P., Mpepereki,S. and Mafongoya,P. (1998) Pigeonpea in Zimbabwe: A new crop with potential. In: Waddington,S.R., Murwira,H.K., Kumwenda,J.D.T. *et al*. (eds), *Proceedings of the Soil Fertility Network Results and Planning Workshop*, Soil Fert Net/CIMMYT, ISBN 970-648-006-4, pp. 93–98.
4. http://www.crnindia.com/commodity/tur.html (15 June 2012, date last accessed).
5. Gonçalves,E.C., Silva,A., Barbosa,M.S.R. *et al*. (2004) Isolation and characterization of microsatellite loci in Amazonian red-handed howlers *Alouatta belzebul* (Primates, Plathyrrini). *Mol. Ecol.*, **4**, 406–408.
6. Kashi,Y. and King,D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.*, **22**, 253–259.
7. Li,Y.C., Korol,A.B., Fahima,T. *et al*. (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.*, **11**, 2453–2465.
8. Li,Y.C., Korol,A.B., Fahima,T. *et al*. (2004) Microsatellites within genes: structure, function and evolution. *Mol. Biol. Evol.*, **21**, 991–1007.
9. Varshney,R.K., Chen,W., Li,Y. *et al*. (2011) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.*, **30**, 83–89.
10. http://pgrc.ipk-gatersleben.de/misa/ (15 June 2012, date last accessed).
11. Kumawat,G., Raje,R.S., Bhutani,S. *et al*. (2012) Molecular mapping of QTLs for plant type and earliness traits in pigeonpea (Cajanus cajan L. Millsp.). *BMC Genet.*, **13**, 84–100.
12. Haseneyer,G., Schmutzer,T., Seidel,M. *et al*. (2011) From RNA-seq to large-scale genotyping-genomics resources for rye (*Secale cereal* L.). *BMC Plant Biol.*, **11**, 131–143.