APMicroDB: A microsatellite database of *Acrithosiphon pism* 

Ritika Bishnoia, Deepak Singlab

**A b s t r a c t**

Pea aphids represent a complex genetic system that could be used for QTL analysis, genetic diversity and population genetics studies. Here, we described the development of first microsatellite repeat database of the pea aphid (APMicroDB), accessible at "http://deepaklab.com/aphidmicrodb". We identified 3,40,233 SSRs using MIcroSAtellite (MISA) tool that was distributed in 14,067 (out of 23,924) scaffold of the pea aphid. We observed 89.53% simple repeats of which 73.41% were mono-nucleotide, followed by di-nucleotide repeats. This database stored information about the repeats kind, GC content, motif type (mono - hexa), genomic location etc. We have also incorporated the primer information derived from Primer3 software of the 2504 bp flanking region of the identified marker. Blast tool is also provided for searching the user query sequence for identified marker and their primers. This work has an immense use for scientific community working in the field of agricultural pest management, QTL mapping, and host-pathogen interaction analysis.

**1. Introduction**

Simple Sequence Repeats (SSRs) also known as Microsatellites, are the extensively dispersed short tandem repeat units harbor substantial length variation [1,2]. A major proportion of eukaryotic genomes (up to 4%) are composed of these markers. Despite their presence in both coding and non-coding region, high abundance was only observed in the non-coding region of the genome [3,4]. Previous studies suggested that short tandem repeats (STRs) are under the selective pressure that played an important role in genome structure and evolution [5–7]. SSRs offers several advantages such as their distribution, specificity, and reproducibility, therefore, they were extensively employed in population genetics [8,9], genetic diversity [10–13] and evolution [14,15]. Based on the origin, SSRs has been classified into two types: 1) genomic SSRs (that derived from genome), and 2) EST-SSRs (that comes from expressed sequence tags) [10,14]. EST-based SSRs were originated from transcribed region which is more conserved as compared to genomic SSRs [16,17]. Therefore, genomic SSRs are highly polymorphic and fitted for genetic diversity studies within a particular species.

The present study is focused on the identification of SSRs from the genome of *A. pism*. Pea aphids (*Acrithosiphon pism*) are the phloem-feeding insects having several advantages over other aphid species [18]. Association of pea aphid with more than 20 legume genera represents their host race specific evolution. Each race is more or less specialized and genetically differentiated from other host races [19,20]. To reveals the host-pathogen relationship, it is important to understand the genomic architecture of aphid genome. Hence, the international aphid genome consortium first time reported the draft genome of the pea aphid of size 464 Mb. Initially, ~3.13 million reads were assembled into 72,844 contigs using Atlas assembly pipeline. However, in the second version, the number of contigs was reduced to 60,596 with the N50 length of around 28 kb. Previously, only few studies have been reported to experimentally characterize the microsatellite markers in pea aphid [21–23]. However, the wet-lab characterization is very tedious and time-consuming job. Therefore, researchers paved the attention for in silico identification of SSRs in the aphid genome [2,24]. For e.g. Behura et al. reported 1,69,601 and 4283 microsatellite repeats in whole genome and coding region of *A. pism* respectively. Based on the identified SSRs, few insect specific databases such as InSatDb, EuMicrosatdb etc. has been developed in the past [25,26]. Best of the author knowledge, no publicly accessible database of SSRs has been reported for the pea aphid. Owing to the importance of microsatellite, and pea aphid as model insect species, the foremost purpose of this manuscript is to discover the abundance and distribution of SSRs in the pea aphid genome.

**2. Database development**

2.1. Database construction and architecture

We have downloaded the pea aphid genome v2.0 from the NCBI database in FASTA format [27]. The complete genome was scaffold-wise scan for the occurrence of microsatellite repeats using MicroSatellite (MISA) tool (http://pgcr.ipk-gatersleben.de/misa/). We used the PRIMEER3 software to predict the primer of the identified microsatellite markers [28]. For this, we extracted a flanking region of 250 bp of the
repeats on both sides using bedtools [29]. The custom PERL scripts were used to process the MISA output in CSV format. Finally, the file was uploaded into MySQL database. The front-end of the database was developed using HTML, PHP language, and JAVA scripts.

2.2. Genome analysis

We analyzed the distribution of STRs across the scaffold and observed that simple microsatellite repeats represents 89.53% of the total STRs (Table 1). We also plotted the different motif repeats from mono- hexa to show their relative abundance in pea aphid genome. As evident from Fig. 1 and Table-S1, Mononucleotide type repeats (73.41%) was most abundant as compared to other types [30,31]. However, hexanucleotide repeats (0.03%) was the least ones (suppl-1.docx, Table-S1). Our analysis also supported the Katti et al. analysis that trinucleotide repeats have a maximum length 441 bp followed by dinucleotides (suppl-1.docx, Table-S1) [32]. We also observed that STRs of length up to 15 bp represents the major proportion in the genome followed by length 16–20 (Fig. 2). However, the motif of length 46–50 bp was represented by only 0.13% (Fig. 2, Table-S2).

2.3. STR validation

Previously, Kurokawa et al. reported six microsatellite markers in pea aphid using experimental approach [21]. In the same year, Caillaud et al. reported fifteen markers from pea aphids [22]. In order to validate this, we used the FASTA sequence of reported marker and search in our database using blast tool. We observed that 76% of the markers were partially or completely matched with our database (Table 2). Out of the 15 markers, we found six were exactly matched, and seven markers matched with repeat kind but their copy number has been changed. This might be because the assembly of pea aphid genome is only available at preliminary scaffold level but not at the chromosome level.

3. Utility

3.1. Search

We provided the scaffold wise search option for STRs along with the marker properties such as the type of motif, repeat kind etc. Furthermore, we have also given the advanced search option to filter the results based on the scaffold region, copy number of the marker, and GC content. This will be helpful to the user interested in locating the marker in the given genomic region of the genome, which may be coding or non-coding. The search result is shown in a well-organized tabular format with an additional button for extracting primer information of a particular SSR (Fig. 3). On clicking the show primer button, users will

Table 1
Overall distribution of SSRs and their percentage in pea aphid genome.

| SSR type | Count   | Percentage |
|----------|---------|------------|
| Simple   | 3,04,595| 89.53%     |
| Compound | 34,196  | 10.05%     |
| Complex  | 1442    | 0.42%      |
| Total    | 3,40,233| 100%       |

Fig. 1. Histogram plot of SSRs with the type of repeats in the x-axis and their percentage in the y-axis.

Fig. 2. Pie chart showing the percent distribution of microsatellite repeats within different length ranges.
get the information about the primers (250 bp flanking region of marker) and their properties.

3.2. Web tool

A customized BLAST tool is implemented in this database for similarity search. The user input query sequence will be searched against the database of repeats containing flanking region. A user-friendly search option for e-value cut off, query coverage and a number of hits to be displayed is provided in the blast search. The identified hit is further linked with the primer information of the identified hits (Fig. 4).

| GenBank Scaffold no. | Reported [21,22] | APMicroDB |
|----------------------|------------------|-----------|
| AY528722 NW_003383558 | (TC)12 (TC)10 | (TC)10 |
| AY528723 NW_003383960 | (CT)12 (CT)10 (CT)7 | (TC)10 |
| AY528724 NW_003383777 | (GA)11 (GA)10 | (GA)10 |
| AY528725 NW_003383909 | (AG)20 (AG)14 | (AG)14 |
| AY528726 NW_003383570 | (CT)10 (CT)14 | (CT)14 |
| AY528727 NW_003383545 | (AG)7 (AG)16 | (AG)7 |
| AY528728 NW_003383554 | (CA)4 (AC)4 Not found |
| AY528729 NW_003384067 | (GT)6 | (GT)6 |
| AY528730 NW_003399631 | (TG)7 | (TG)7 |
| AY528731 NW_003383549 | (GCT)8 (GCT)8 | (GCT)8 |
| AY528732 NW_003384067 | (AC)5 (AC)4 Not found |
| AY528733 NW_003383549 | (AC)8 | (AC)8 |
| AY528734 NW_003384434 | (CA)10 | (CA)10 |
| AY528735 NW_003383752 | (CA)16 | (CA)16 |
| AY528736 NW_003384150 | (CA)7 (TA)3 (CA)3 | (TG)7 |
| AB162918 NW_003383507 | (ATA)5 Not found |
| AB162919 NW_003383919 | (GC)5 (TTA)7 | (TG)12|
| AB162920 NW_003383501 | (GA)11 | (GA)11 |
| AB162921 NW_003383764 | (AT)9 Not found |
| AB162922 NW_003383818 | (AC)7 (AC)5 Not found |
| AB162923 NW_003383520 | (TG)8 | (TG)8 |

4. Discussion

Here, we reported the mining of 3,40,233 microsatellite markers, which is almost double that are reported by Behura and Severson [24]. The percentage of mono- was higher followed by di-, tri-, tetra, penta, and hexa-nucleotide repeats respectively. A similar trend was observed by Sharma et al. supporting the fact that an increase in repeat length is proportional with the decrease in repeat numbers [31]. The distribution of repeat length showed a good coverage in the range of 11–15 bp long repeats. However, low coverage (0.13%) was observed in the case of repeats of length 46–50 bp. In 2001, Katti et al. observed that tri-nucleotide repeat seems to be much longer as compared to other repeats in
Drosophila [32]. This is highly correlated with our study of pea aphid that belongs to the same phylum. A significant correlation with the previously identified marker suggests the application of this database. Despite the improvement in pea aphid assembly from version 1.0 to version 2.0 still the assembly existed at the scaffold level. This indicates a gap in the knowledge of SSR markers in pea aphids and suggested that there must be a much more SSRs marker that could only be resolved only at the chromosome level.

5. Data maintenance

APMicroDB will be regularly maintained by our team. We will welcome any scientific suggestion from the readers via ‘Contact’ link on the database. In future, we will upgrade the database whenever the new assembly from different strain/race of pea aphid will be reported. The update will be helpful in study species-specific primer and establish an evolutionary relationship.

Fig. 4. The overall flow of user Blast query, and its link to database and primers.
6. Conclusion

STRs are the most extensively studied marker having wide application in genetic diversity, evolution, and genome mapping. Despite the great importance of microsatellite makers, no database exists to store and compile the genome-wide information of SSR markers from pea aphid. Therefore, in the present work, an effort has been made to develop first whole genome based SSRs database of pea aphid that will be useful in phylogenetic analysis, and evolutionary insight on pea aphid.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

The author is thankful to Mr. Amit Pandey for their help in database designing and also thankful to ICAR-IASRI for providing RA support. No separate funding is provided for publication of this article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2017.03.014.

References

[1] T.-H. Jun, A.P. Michel, M.A.R. Mian, Development of soybean aphid genomic SSR markers using next generation sequencing. Genome 54 (5) (May 2011) 360–367.
[2] S.K. Behura, D.W. Severson, Association of microsatellite pairs with segmental duplications in insect genomes. BMC Genomics 14 (1) (Jan 2013) 907.
[3] G. Tóth, Z. Gáspari, J. Jurka, Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res. 10 (7) (Jul. 2000) 967–981.
[4] J.M. Hancock, Simple sequences and the expanding genome. Bioessays 18 (5) (May 1996) 421–425.
[5] H. Ellegren, Microsatellites: Simple sequences with complex evolution. Nat. Rev. Genet. 5 (6) (Jun. 2004) 435–445.
[6] Y. Kashi, D.G. King, Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22 (5) (May 2006) 253–259.
[7] S.K. Behura, Molecular marker systems in insects: current trends and future avenues. Mol. Ecol. 15 (11) (Oct 2006) 3087–3113.
[8] K.S. Kim, T.W. Sappington, Microsatellite database analysis for population genetics. Methods Mol. Biol. 1006 (Jan 2013) 271–295.
[9] A.I. Putman, I. Carbone, Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecol. Evol. 4 (22) (Oct 2014) (p. n/a–n/a).
[10] S. Jing, B. Liu, L. Peng, X. Peng, L. Zhu, Q. Fu, G. He, Development and use of EST-SSR markers for assessing genetic diversity in the brown planthopper (Nilaparvata lugens Stål). Bull. Entomol. Res. 102 (1) (Feb 2012) 113–122.
[11] F. von H. Fontes, C.A. Colombo, A.L. Lourenço, Structure of genetic diversity of Bemisia tabaci (Genn.) (Hemiptera: Aleyrodidae) populations in Brazilian crops and locations. Sci. Agric. 69 (1) (Feb 2012) 47–53.
[12] K.P. Arunkumar, A.K. Sahu, A.R. Mohanty, A.K. Awasthi, A.R. Pradeep, S.R. Urs, J. Nagaraju, Genetic diversity and population structure of Indian golden silkmoth (Antheraea assama), Plus One 7 (8) (Jan 2012) e43716.
[13] A.R. Mahon, C.P. Arango, K.M. Halanych, Genetic diversity of Nympy (Aphrodispa: Pycnococcidae: Nympyidae) along the Antarctic Peninsula with a focus on Nympy australiae Hodgson 1902. Mar. Biol. 155 (3) (Jul 2008) 315–323.
[14] K.S. Kim, S.T. Ratcliffe, B.W. French, L. Liu, T.W. Sappington, Utility of EST-derived SSRs as population genetics markers in a beetle. J. Hered. 99 (2) (Jan 2008) 112–124.
[15] E. Stolle, J.H. Kidner, R.F.A. Moritz, Patterns of evolutionary conservation of microsatellites (SSRs) suggest a faster rate of generic evolution in Hymenoptera than in Dipitera. Genome Biol. Evol. 5 (1) (Jan 2013) 151–162.
[16] B.M.T. Brunet, D. Desser, R.R. Sturtevant, F.A.H. Sperling, Characterization of EST-based SSR loci in the spruce budworm, Choristoneura fumiferana (Lepidoptera: Tortricidae). Conserv. Genet. Resour. 5 (2) (Jan 2013) 541–544.
[17] E. Shiferaw, M.E. Pè, E. Forcédade, M. Ponniah, Exploring the genetic diversity of Ethiopian grass pea (Lathyrus sativus L) using EST-SSR markers. Mol. Breed. 30 (2) (Aug 2012) 789–797.
[18] E. Shiferaw, M.E. Pè, E. Forcédade, M. Ponniah, Exploring the genetic diversity of Ethiopian grass pea (Lathyrus sativus L) using EST-SSR markers. Mol. Breed. 30 (2) (Aug 2012) 789–797.
[19] T. Will, A.C.U. Furch, M.R. Zimmermann, How phloem-feeding insects face the challenge of phloem-located defenses. Front. Plant Sci. 4 (Jan 2013) 336.
[20] J. Pecoudd, A. Ollivier, M. Plantegenest, J.-C. Simon, A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. Proc. Natl. Acad. Sci. U. S. A. 106 (18) (May 2009) 7495–7500.
[21] J. Stavrinides, J.K. McInnesHy, K. Ochman, Pea aphid as both host and vector for the phytopathogenic bacterium Pseudomonas syringae. Appl. Environ. Microbiol. 75 (7) (Apr 2009) 2230–2235.
[22] T. Kurokawa, I. Yao, S.-i. Akimoto, E. Hasegawa, Isolation of six microsatellite markers from pea aphid, Acrystosiphon pisum (Homoptera: Aphididae). Mol. Ecol. Notes 4 (3) (Sep 2004) 523–524.
[23] M.C. Caillaud, G. Mondor-Genson, S. Levine-Wilkinson, L. Mieuzet, A. Frantz, J.C. Simon, A. Coeur D’acier, Microsatellite DNA markers for the pea aphid Acrystosiphon pisum. Mol. Ecol. Notes 4 (3) (Sep 2004) 446–448.
[24] W. Yung, P. Aghazadeh, B. G. Atlas, J. C. Budd, Cross-species transferability of microsatellite markers from six aphid (Hemiptera: Aphididae) species and their use for evaluating biotypic diversity in two cereal aphids. Insect Mol. Biol. 16 (5) (Oct 2007) 613–622.
[25] S.K. Behura, D.W. Severson, Genome-wide comparative analysis of simple sequence coding repeats among 25 insect species. Gene 504 (2) (Aug 2012) 226–232.
[26] S. Archak, E. Meduri, P.S. Kumar, J. Nagaraju, InSatDb: a microsatellite database of fully sequenced insect genomes. Nucleic Acids Res. 35 (Database issue) (Jan 2007) D36–D39.
[27] V. Aishwarya, A. Grover, P.C. Sharma, EuMicroSatDb: a database for microsatellites in the sequenced genomes of eukaryotes. BMC Genomics 8 (1) (Jan 2007) 225.
[28] Genome sequence of the pea aphid Acrystosiphon pisum, Plus One 8 (2) (Feb 2010) e1000311.
[29] A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B.C. Faircloth, M. Remm, S.G. Rozen, Primer3—new capabilities and interfaces. Nucleic Acids Res. 40 (15) (Aug 2012) 797–806.
[30] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 (6) (Mar. 2010) 841–842.
[31] S. Temnykh, G. DéClerck, A. Lakshavha, L. Lipovich, S. Cartinhour, S. McCouch, Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11 (8) (Aug. 2001) 1441–1452.
[32] P.C. Sharma, A. Grover, S. Kahl, Mining microsatellites in eukaryotic genomes. Trends Biotechnol. 25 (11) (Nov. 2007) 490–498.
[33] M.V. Katti, P.K. Ranjekar, V.S. Gupta, Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18 (7) (Jul. 2001) 1161–1167.