ECS Transactions

Statistical Analysis of Simple Sequence Repeats in Genome Sequence: A Case of *Acheta Domesticus* (Orthoptera: Gryllidae)

ECS Transactions, 107 (1) 14799-14806 (2022)
10.1149/10701.14799ecst ©The Electrochemical Society

Statistical Analysis of Simple Sequence Repeats in Genome Sequence: A Case of *Acheta domesticus* (Orthoptera: Gryllidae)

S. Homchan[a], P. Bhadola[b], and Y. M. Gupta[a*]

[a]Department of Biology, Faculty of Science, Naresuan University, Phitsanulok, 65000, Thailand.
[b]Centre for Theoretical Physics & Natural Philosophy "Nakhonsawan Studiorum for Advanced Studies", Mahidol University, Nakhonsawan Campus, Phayuha Khiri, Nakhonsawan, 60130, Thailand.

Herein, we have described genome wide screening of microsatellites (Simple Sequence Repeats: SSRs) and their distribution in the *Acheta domesticus* genome using a custom python script. A total of 232,179 microsatellites were identified, in which trinucleotide repeats were found to be the most abundant repeats in the genome, representing 60 % of the total microsatellite, followed by dinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides. Among trinucleotides, dinucleotides, and tetranucleotides, the most prevalent microsatellite motifs were AAT/ATT, TG/CA, and AAAT/ATTT, respectively. Notably, statistical analysis of the SSR repetition frequency distribution showed that SSR motifs that were repeated four times were the most recurrent. Additionally, the SSR length distribution showed that SSRs with a12 bp size were the most common. As a result, the statistical analysis of the SSR dataset in *A. domesticus* will be a useful resource for a better understanding of microsatellite distribution in crickets for evolutionary genetics.

Introduction

There are a number of instances and compelling theoretical explanations why repetitive DNA is necessary for genome function. A Simple sequence repeat (SSR) is one of the subclasses of tandem repeats that make up genomic repetitive regions (1). SSRs are also referred as microsatellites, opposite to minisatellites, which have longer nucleotide motifs. SSR variants are often caused by the addition or deletion of complete repeat motifs (2). As a result, different individuals show variations in repetition counts at specific loci, making them highly useful genetic markers that are experimentally reproducible and transferable among related species (3, 4). Aside from the use of SSR as a genetic marker for genotyping in prokaryotes and eukaryotes, it is also important to understand their distribution throughout the respective genome. Therefore, the focus of present research is to develop an alternative approach to search and examine SSRs occurrences from the assembled genome sequence: A case of *Acheta domesticus*.

Microsatellite DNA sequences were first studied almost four decades ago and found to be dispersed throughout the genome (5). Onwards, they have been found in a wide array of species, including crickets (4). SSR can provide more information more easily than other

conventional DNA-based genetic marker technologies, such as RFLP and RAPD. Moreover, microsatellite assays like; population genetics (4, 6), conservation biology (7, 8), and evolutionary biology (9), require only a small quantity of DNA and these markers are highly reproducible, and readily transferable to other species (10, 11). Despite the widespread use of SSRs and the development process, bioinformatical algorithms must evolve to take advantage of next-generation sequencing technology. Currently, SSR marker commonly developed by genome and transcriptome dataset. The recent review of development process of SSR marker using next generation sequencing technology explain the use of RNA-seq data (12). The review was limited to *de novo* transcriptome for SSR development, that also being used by many researchers. However, SSR development from genome data has been in practice (4), but tools have been developed to search transcriptome and genome for simple or complex repeats, but they do not provide information on SSR distribution.

The next generation sequencing (NGS) technology is rapidly evolving, which also increases the number of sequence data (13). As a result, NGS has aided SSR development by allowing for a faster search of SSR in genomic sequences (12). Initially, biotin-labelled oligonucleotide were used to capture microsatellite containing DNA fragment using Streptavidin-coated magnetic beads (14, 15). Since 2009, microsatellite development using next-generation sequencing data has become more common, with NGS data from the genome and transcriptome being used in a number of studies (16). The microsatellite marker development has become cheaper and faster due to advancement in NGS technology (17).

SSR makers in population genetic studies often focus on a small number of polymorphic SSR loci for genetic research (18, 19). In our previous research, the genome of *A. domesticus* was sequenced and deposited on NCBI (GenBank assembly accession: GCA_014858955.1). From the genomic sequences, only 91 SSR loci (91 sequences out of 709,385) were utilized for the study (4). As result, the present research is an extension of the previous study to conduct a genome wide search and investigate the distribution of SSRs. The complete protocol of sample collection, DNA extraction, and sequencing is also explained in previous paper (4). However, a genome-wide characterization of microsatellites remains unidentified in *A. domesticus.* Therefore, brief explanation and statistical analysis of SSRs is conducted in present study. It is important to understand the SSR distribution in the genome, therefore, genomic sequence data is readily utilized to min SSR (20).

There are numerous techniques available for searching SSR from nucleotide sequences. One of the tool to search SSR from the genomic sequences is MIcroSAtellite identification tool (MISA)(21). Herein, the preliminary SSR search was conducted using (MISA). Recently, the characterization of microsatellite DNA in genomes have been conducted to examine their abundance and frequencies (22). Similarly, we intend to conduct SSR distribution study for *A. domesticus* genome. The traditional SSR search tools are heavily reliant on searching SSR (21). Herein, we plan to give detailed information on each class of SSR, including the occurrence frequency of each distinct motif type and repeat count. Our statical data of SSR in *A. domesticus* genome could be beneficial to understand SSR structure and distribution of each SSR motif in different SSR types.

Materials and Methods

Data validation

To min the repeat sequence, *A. domesticus* genome was used for analysis purposed because it has been utilized for developing molecular markers like microsatellite markers but never been employed to conduct genome wide microsatellite repeat distribution analysis. Therefore, the assembled genome sequence of *A. domesticus* was used to inspect the occurrence of microsatellite (4).

Genome mining for SSR

The assembled genome sequence of *A. domesticus* was used to inspect the occurrence of microsatellite (4). In previous study, the microsatellites were examined using MIcroSAtellite identification tool (MISA) tool. Herein, SSR dataset was reexamined for validating the presence of each SSR type. SSR distribution analysis was performed on the SSR dataset using a custom python script to calculate SSR type, SSR sequence length, and SSR motif frequency. Herein only perfect microsatellite repeats were taken in account for distribution analysis. The genome dataset containing 709,397 nucleotide sequences were mined to identify SSR with minimum number of repeats for each class (SSR class/unit size: Di-/6, Tri-/4, Tetra-/4, Penta-/4, Hexa-/4).

Microsatellite analysis

Custom python script was written using NumPy, Pandas and Matplotlib python libraries. The file containing microsatellite dataset from genome of *A. domesticus* was analyzed using NumPy and Pandas library. Matplotlib library was used for data visualization.

Results and Discussion

We analyzed SSRs from draft genome assembly of *A. domesticus* genome. Total of 709,397 sequences containing 929,180,478 base pairs were employed for microsatellite distribution analyses. Microsatellites were analyzed using custom python scripts from SSR dataset. Total 2,28,632 SSRs were identified, in which 225,950 were present in perfect formation and 2,682 SSRs were in compound formation. Trinucleotide units were the most common microsatellite units, representing for 60%, followed by dinucleotide (22%), tetranucleotide (11%), pentanucleotide (5%), and other compound microsatellite units (1%). The distribution of different perfect repeat types in given in Table I**.**

TABLE I. Distribution to different perfect repeat type classes

Unit size	Number of SSRs
Dinucleotide	50,857
Trinucleotide	135,908
Tetranucleotide	25,296
Pentanucleotide	11,767
Hexanucleotide	2,122

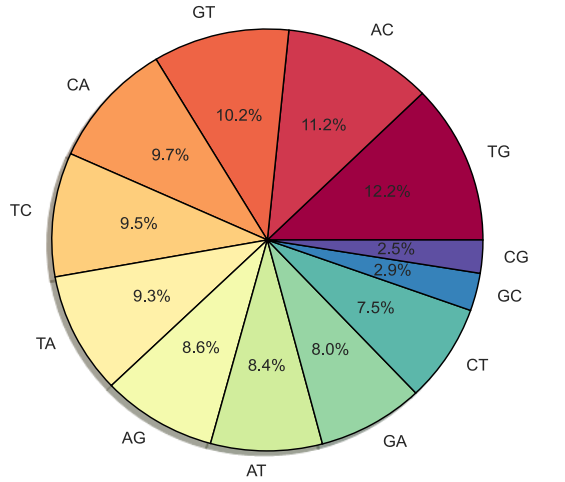
The frequency of distinct microsatellites units varies. In which, CA/TG units were the most common dinucleotides in the genome sequence, accounting for 43.30 %, followed by TC/GA (33.60%), TA/AT (17.68%), and GC/CG (5.42%). Among trinucleotides, AAT/TTA units were the most common among trinucleotide units, contributing for 27.41 %, followed by GGC/CCG (20.38 %), TAT/ATA (11.88 %), and CGC/GCG (10.37 %). The abundance of GTC/GAC trinucleotide units were least with 4.08 %. For other repeats including Tetra-, Penta- and Hexa- nucleotides, AAAT/TTTA (22.48 %), CGCCG/CGGCG (17.48 %), CCGCCC/GGGCGG (8.44 %) were most abundant units in each class. The frequency of each repeat types and their reverse complement and revers sequence also varies frequency varies which can due to variation in sequence repeats. Di-, Tri-, and Tetra-nucleotide repeat frequencies are shown in Figure 1.

Microsatellites that are extensively dispersed across the genome are useful genetic markers for measuring genetic diversity, building genetic maps, comparative genomics, and marker-assisted selective breeding (4, 22). The finding and characterization of microsatellite markers across the genome provides significant information for understanding gene accidents with repeats, genome stability, and evolution. Repeat distribution studies have proven that it may be a beneficial tool in a variety of genetics and plant breeding studies (23). As a result, we conducted the first SSR distribution study of the *A. domesticus* genome to investigate repeat frequency.

The $(CA/TG)_n$ repeat units are shown to be the most frequent dinucleotide in this study. These dinucleotide repeats have been shown to be preferentially associated to Alu elements (24), and these dinucleotide repeats were also shown to be highly frequent in the swamp eel genome (23). The most common trinucleotide repeats discovered in this study were AAT/TTA, which are also seen in swamp eels (25) and humans (26). The most frequent repeat units in trinucleotide and tetranucleotide are AAT and AAAT, showing the preponderance of A-rich repeats throughout cricket genome evolution, which is consistent with swamp eel (25). Furthermore, trinucleotide repeats have a much greater destiny than other repeat classes and include A-rich repeat units, which might be associated to Alu repeats. Likewise, Alu repeats are widely dispersed in the human genome, accounting for 10% of the overall genome size (27).

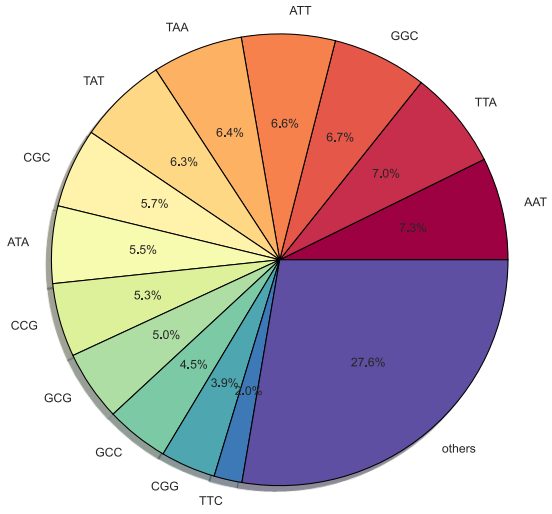
Microsatellites' frequency and density are most likely connected to genome size, with density being higher in large genomes than in short genomes throughout mammals. However, the frequency of microsatellites in plants is lower in larger genomes than in smaller genomes (28, 29). Aside from the frequency of microsatellites in genomic sequences, the composition and size of SSR units also vary. In terms of SSR unit repetition frequency, SSR repeated four times was the most common when compared to other SSR repetition frequencies (Figure 2). In terms of SSR size frequency, SSRs of 12 base pairs were the most common (Figure 3). This is connected to the number of trinucleotide SSR units and their four-fold repetition in the genome (e.g. $(AAT)_4$ = 12 base pair).

The genomic sequences are great sources for SSR mining and have been used in a variety of species (30). The current research focused on the insides of SSR distribution was restricted to genomic sequences in the *A. domesticus* genome. The search for SSR patterns was limited to a minimum of six units for dinucleotides and four units for tri-, tetra-, penta-, and hexanucleotides.



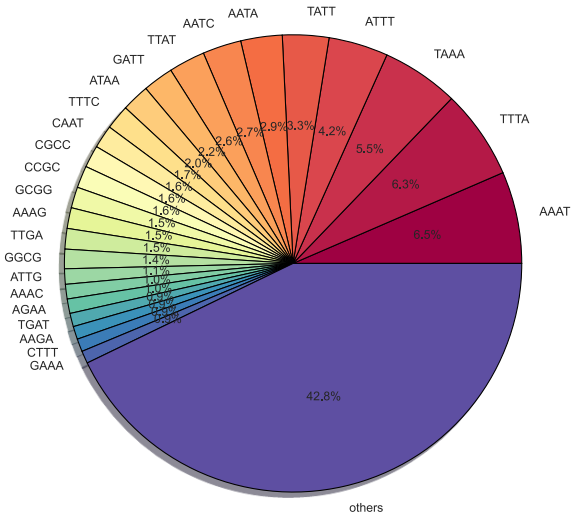
A.
Dinucleotide repeat distribution

CA/TG (AC/GT) = 43.30 %



B.
Trinucleotide repeat distribution

AAT/TTA (TAA/ATT) = 27.41 %



C.
Tetranucleotide repeat distribution

AAAT/TTTA(TAAA/ATTT)
= 22.48 %

Figure 1. Distribution frequency of three classes of repeats found in genome of *A. domesticus*: (a) dinucleotide repeat, (b) trinucleotide repeats (c) tetranucleotide repeats. The frequency of reverse compliment and reverse sequence is shown separately in the pie chart.

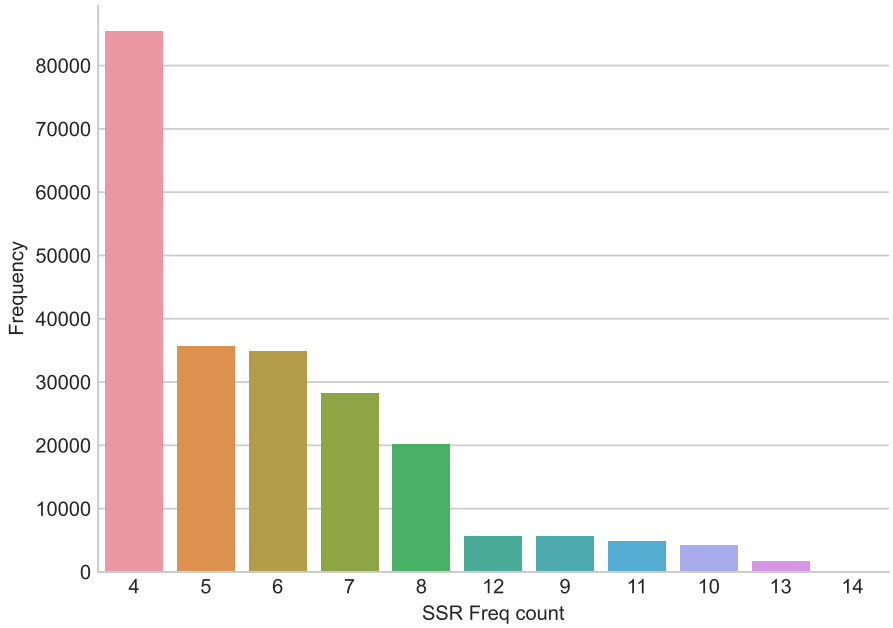


Figure 2. SSR repetition frequency distribution in *A. domesticus* genome sequences.

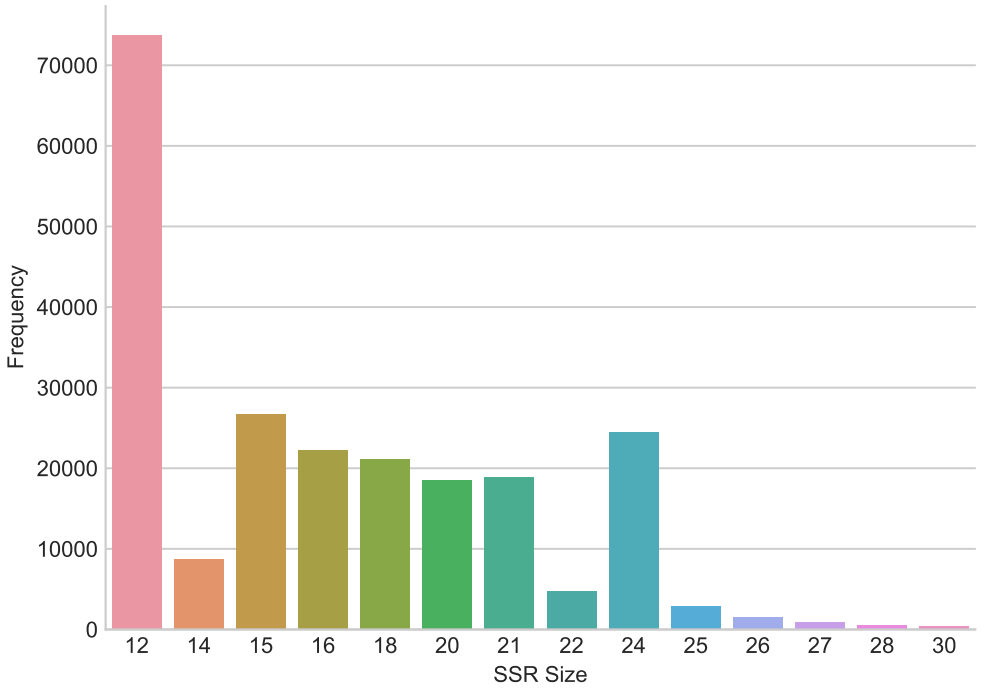


Figure 3. SSR size (length in base pair) frequency distribution in *A. domesticus* genome sequences.

Upon statistical analysis on SSR library generated from genome sequences, we discovered that trinucleotide abundance is significantly greater than that of other SSRs (Di-, Tetra-, Penta-, Hexa-nucleotides). However, *A. domesticus* transcriptome sequence will be necessary to acquire a better understanding of SSR connected to gene regulation and expression. To the best of our knowledge, this is the first study to explain SSR distribution in the genome of *A. domesticus*. It will provide a basis to understand more about microsatellite distribution in crickets for evolutionary genetics in general, as well as a foundation for understanding distribution of genic and intergenic nucleotide repeats.

Conclusion

Microsatellite distribution was never been conducted for *A. domesticus.* Herein, we have conducted deep analysis on microsatellite occurrence in the genome of *A. domesticus.* We have found that trinucleotide occurred the most in the genome sequence followed by Di-, Tetra-, Penta-, and Hexa- nucleotide units. Notably, among all SSR types, A/T- rich repeats were most abundant in the genome sequence. This may indicate the predominance of this motif during field cricket genome evolution, although more research into microsatellite distribution in insect genomes is needed to fully comprehend repeat occurrences and evolution.

Future direction

Herein, we have provided insides of microsatellite distribution in *A. domesticus* genome using custom python script from SSR dataset generated using MISA (21). However, the traditional SSR search tool like MISA are heavily reliant on only searching SSR (21). In the future, we intend to create a Python script to search for SSR motifs and provide thorough information on each class of SSR, including the occurrence frequency of each different motif type and repeat count.

Acknowledgments

The authors hereby acknowledge the Department of Biology, Faculty of Science, Naresuan University, Thailand for awarding a research grant [R2564C033], and also for their laboratory facilities. PB wants to thank Mahidol University for research support "MRC-MGR 04/2565".

References

1. M. L. Vieira, L. Santini, A. L. Diniz and F. Munhoz Cde, *Genet. Mol. Biol.*, **39**(3), 312 (2016). doi: 10.1590/1678-4685-GMB-2016-0027.
2. R. Gemayel, J. Cho, S. Boeynaems and K. J. Verstrepen, *Genes-Basel*, **3**(3), 461 (2012). doi: 10.3390/genes3030461.
3. J. Qu and J. Liu, *BMC Res Notes*, **6**(1), 403 (2013). doi: 10.1186/1756-0500-6-403.
4. Y. M. Gupta, S. Tanasarnpaiboon, K. Buddhachat, S. Peyachoknagul, P. Inthim and S. Homchan, *Biodiversitas Journal of Biological Diversity*, **21**(9) (2020). doi: 10.13057/biodiv/d210921.
5. H. Hamada, M. G. Petrino and T. Kakunaga, *Proceedings of the National Academy of Sciences*, **79**(21), 6465 (1982).
6. M. Płecha, H. Panagiotopoulou, D. Popović, A. Michalska-Parda, R. Gromadka, P. Węgleński and A. Stanković, *Fisheries & Aquatic Life*, **27**(1), 33 (2019). doi: 0.2478/aopf-2019-0004.
7. S. Roques, P. Berrebi, P. Chèvre, E. Rochard and M.-L. Acolas, *Conserv. Genet. Resour.*, **8**(3), 313 (2016). doi: 10.1007/s12686-016-0538-7.
8. C. Spitzweg, P. Praschag, S. DiRuzzo and U. Fritz, *Salamandra*, **54**(1), 63 (2018).

9. E. Guang-Xin, L.-P. Chen, D.-K. Zhou, B.-G. Yang, J.-H. Zhang, Y.-J. Zhao, Q.-H. Hong, Y.-H. Ma, M.-X. Chu and L.-P. Zhang, *Mol. Immunol.*, **124**(83 (2020). doi: 10.1016/j.molimm.2020.05.005.
10. A. J. Armstrong, C. L. Dudgeon, C. Bustamante, M. B. Bennett and J. R. Ovenden, *BMC research notes*, **12**(1), 233 (2019). doi: 10.1186/s13104-019-4270-8.
11. Y. Xiao, W. Xia, J. Ma, A. S. Mason, H. Fan, P. Shi, X. Lei, Z. Ma and M. Peng, *Frontiers in plant science*, **7**, 1578 (2016). doi: 10.3389/fpls.2016.01578.
12. S. Taheri, T. Lee Abdullah, M. R. Yusop, M. M. Hanafi, M. Sahebi, P. Azizi and R. R. Shamshiri, *Molecules*, **23**(2), 399 (2018). doi: 10.3390/molecules23020399.
13. B. E. Slatko, A. F. Gardner and F. M. Ausubel, *Curr. Protoc. Mol. Biol.*, **122**(1), e59 (2018). doi: 10.1002/cpmb.59.
14. D. Xin, C. Xin-Bo, L. Song-Hua, X.-C. Wang, G. Yuan, H. Dong-Feng, W. Jin and W. Yu-Fu, *Acta Agronomica Sinica*, **34**(12), 2099 (2008). doi: 10.1016/S1875-2780(09)60021-3.
15. H. Vignes and R. Rivallan, in *Molecular Plant Taxonomy*, p. 177, Springer (2014).
16. J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen and M. L. Blaxter, *Nature Reviews Genetics*, **12**(7), 499 (2011). doi: 10.1038/nrg3012.
17. J. Abdelkrim, B. C. Robertson, J.-A. L. Stanton and N. J. Gemmell, *BioTechniques*, **46**(3), 185 (2009).
18. D. Annapurna, R. R. Warrier, A. N. Arunkumar, R. Aparna, C. N. Sreedevi and G. Joshi, *3 Biotech.*, **11**(7), 310 (2021). doi: 10.1007/s13205-021-02858-w.
19. T. R. H. Kerkhove, B. Hellemans, M. De Troch, A. De Backer and F. A. M. Volckaert, *Mol. Bio.l Rep.*, **46**(6), 6565 (2019). doi: 10.1007/s11033-019-05026-9.
20. Sarika, V. Arora, M. A. Iquebal, A. Rai and D. Kumar, *BMC Genomics*, **14**(1), 1 (2013). doi: 4310.1186/1471-2164-14-43.
21. S. Beier, T. Thiel, T. Munch, U. Scholz and M. Mascher, *Bioinformatics*, **33**(16), 2583 (2017). doi: 10.1093/bioinformatics/btx198.
22. Y. Lei, Y. Zhou, M. Price and Z. Song, *BMC Genomics*, **22**(1), 421 (2021). doi: 10.1186/s12864-021-07752-6.
23. R. Kumari, D. P. Wankhede, A. Bajpai, A. Maurya, K. Prasad, D. Gautam, P. Rangan, M. Latha, K. J. John, S. A, K. V. Bhat and A. B. Gaikwad, *PLoS One*, **14**(12), e0226002 (2019). doi: 10.1371/journal.pone.0226002.
24. S. S. Arcot, Z. Wang, J. L. Weber, P. L. Deininger and M. A. Batzer, *Genomics*, **29**(1), 136 (1995). doi: 10.1006/geno.1995.1224.
25. Z. Li, F. Chen, C. Huang, W. Zheng, C. Yu, H. Cheng and R. Zhou, *Sci. Rep.*, **7**(1), 3157 (2017). doi: 10.1038/s41598-017-03330-7.
26. S. Subramanian, R. K. Mishra and L. Singh, *Genome Biol.*, **4**(2), 1 (2003). doi: 10.1186/gb-2003-4-2-r13.
27. E. Nadir, H. Margalit, T. Gallily and S. A. Ben-Sasson, *Proc. Nat.l Acad. Sci. U S A*, **93**(13), 6470 (1996). doi: 10.1073/pnas.93.13.6470.
28. H. Ellegren, *Nat. Rev. Genet.*, **5**(6), 435 (2004). doi:10.1038/nrg1348.
29. M. Morgante, M. Hanafey and W. Powell, *Nat. Genet.*, **30**(2), 194 (2002). doi: 10.1038/ng822.
30. Y. Wang, C. Yang, Q. Jin, D. Zhou, S. Wang, Y. Yu and L. Yang, *BMC Genet.*, **16**(1), 18 (2015). doi: 10.1186/s12863-015-0178-z.