Jurnal Teknologi

Full Paper

DISCOVERY OF POTENTIAL SSR MARKERS THROUGH GENOME WIDE ANALYSIS OF MAIZE B73 GENOME

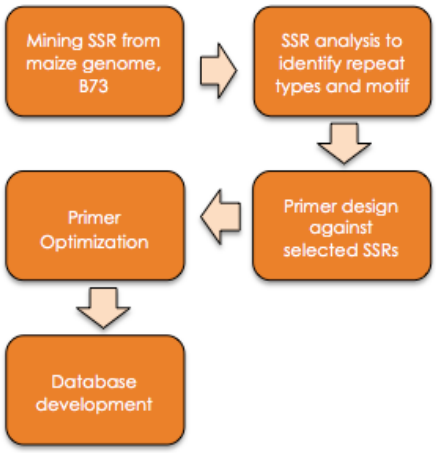
Rabiatul Adawiah Zainal Abidin*, Zulkifli Ahmad Seman, Shahril Ab Razak, Norzihan Abdullah, Umi Kalsom Abu Bakar

Centre for Marker Discovery & Validation, Biotechnology & Nanotechnology Research Centre, Malaysian Agriculture Research Development Institute, MARDI HQ, 43400 Serdang, Selangor, Malaysia

Article history
Received
26 June 2015
Received in revised form
5 September 2015
Accepted
15 October 2015

*Corresponding author
rabiatul@mardi.gov.my

Graphical abstract



Abstract

In silico analysis provides an economical approach in the development of simple sequence repeat (SSR) markers through utilization of genome sequences generated from high throughput sequencing platform. In this study, we present the potential SSR markers of maize mining from its reference genome of cultivar B73. In total, 94, 534 putative SSRs were detected in maize reference genome B73. Dinucleotide repeats (57.00%) were found the most frequent repeats in maize genome, followed by trinucleotide (38.90%), tetranucleotide (2.77%), pentanucleotide (0.85%) and hexanucleotide (0.48%) repeats. A total of 2239 primer pairs were successfully designed for experimental validation. Of these, 99 SSR markers were selected for optimization and only 71(71.71%) SSR primer pairs produced DNA amplification products and therefore validated as developed SSR markers for maize. This *in silico* approach through genome wide analysis of maize genome not only provides rapid discovery and cost effective methods in SSR markers development but also will act as useful tool for genetic diversity and marker-trait association in maize.

Keywords: In silico, maize, microsatellite, simple sequence repeats (SSR), SSR development

1.0 INTRODUCTION

Maize (*Zea mays*) belongs to the *Pocacea* family. According to Food and Agriculture Organization (FAO), maize has been ranked third in the world after rice and wheat due to high consumption in diets among populations in developing country [1]. In Malaysia, maize is a minor crop but important to feed livestock. Continuous breeding activities for genetic improvement in maize were conducted to fulfill the high demand of maize. Most of the breeding activities were focused on increment of yield and resistance towards the diseases. Therefore, research on breeding in maize is important to tackle this issue

through utilization of molecular markers such as simple sequence repeats (SSRs) and single nucleotide polymorphism (SNP).

SSRs are widely used for application in marker-assisted selection (MAS) and breeding in crops due to its advantages. The SSRs are tandem repeats of 2-6 nucleotides, multiallelic, co-dominant and show high level of polymorphism [2]. With the advances of high throughput sequencing technology and bioinformatics tools, it is now possible to develop the SSRs using computational tools and publicly available sequences from public databases. This approach is known as *in silico* analysis. The *in silico* analysis provides more efficient and cost effective approach compared to conventional method in SSRs development whereby the process are usually labour intensive and time consuming [3]. In the past few years, an increasing number of SSR markers from genome by using *in silico* approach are being adopted such as for castor oil [4], wheat rust [5], rice [6], potato [7] and chickpea [8].

Meanwhile, various bioinformatics tools have been developed by bioinformatics communities to mine SSRs such as SSRFinder [9], MIcroSAtellite (MISA) [10], GMATo [11] and SSR Locator [12]. These SSRs discovery tools have different platform dependent and allow rapid discovery of putative SSRs markers. For instance, SSR Locator is suitable for users that are not familiar with command line interface. Both MISA and GMATO are powerful tools for Linux users who familiar with command line interface and therefore permit large discovery of SSRs from bulk sequences data. Hence, it is highly depends on users to choose their SSR discovery tools regardless to their computer literacy and data size.

Prior to this study, the maize genome had been sequenced and deposited in MaizeGDB [13] with about two thousands pairs of maize SSR primers have been deposited in maize genome database. This effort becomes a valuable resource for maize breeder and geneticist to accelerate their study on marker assisted selection (MAS) and marker assisted breeding (MAB) in maize.

Taking these advantages, we utilized maize genome sequences from MaizeGDB to mine the putative SSR markers using computational approach. The aim of this study was to develop a set of SSR markers from maize genome sequences for application in genetic diversity study in maize.

2.0 EXPERIMENTAL

2.1 SSR Identification and Selection of Genomic SSR Primers

The whole genome sequences of maize were retrieved from MaizeGDB database. A total of ten chromosomes from *Zea mays* cultivar B73 were available at the time of the study. We used MISA, a stand-alone software package to discover the SSRs from B73 genome sequences. We set the minimum number of repeats as follows; ten for dinucleotide, six for trinucleotide and five for each tetranucleotide, pentanucleotide and hexanucleotide. The maximum lengths of interruption between two adjacent SSR repeat unit was set at 100 base pair (bp). Further filtering of SSRs was performed using a custom Perl scripts and MySQL to examine the repeat distributions and screen the candidate SSRs. We extracted 200 bp of each side of repeat motif region using BEDTools version 2.18 [14]. The primer pairs for amplification of the SSR motifs were designed using Primer3 [15]. The input parameter for primer design were set as follows; 1) melting temperature, T_m ranged from 55 to 65 ^{0}C, 2) minimal primer length is 20 nucleotide bases with the maximum of 30 nucleotide bases 3) GC content were ranged from 40-60%, and 4) amplicon size were set in ranged of 100 to 450 bp.

2.2 Evaluation of SSR Primers

To evaluate the ability of potential SSRs, we conducted an experimental validation. DNA from young leaves was extracted according to the established DNA extraction protocol routinely used in Centre for Marker Discovery and Validation (CMDV). PCR was prepared in a total volume of 10 µl containing 50 ng of genomic DNA, 0.5 units of *Taq* DNA polymerase and 1X buffer (Invitrogen), 2.5 mM of $MgCl_2$, 0.2 µm of each dNTP and 0.4 µm of each primer. Thermocycler (Biorad, USA) was used to amplify the DNA which involved an initial denaturation at 94 °C/2 minutes followed by 34 cycles at 94 °C/30 seconds, range of annealing temperature between 41°C-65°C/30 seconds, 72°C/30 seconds and followed by final extension at 72°C/10 minutes. Electrophoresis was done using 2% agarose gel to visualize the amplification product.

3.0 RESULTS AND DISCUSSION

3.1 Frequency and Distribution of SSRs in the Maize Genome Sequences

SSR markers based on genome sequences are useful for diversity analysis due to high polymorphism in genome. In addition, SSRs are widely dispersed throughout the genome. To date, the publication of mining SSR markers from maize genome has provided abundant information that can be used to design SSR markers and perform wide range of studies on the genetic variation in maize [16,17]. However, it is a preference for us to develop the new SSR markers for in house application due to data privacy.

In this study, a total of ten chromosomes with size 2059.701728 Mbp were screened for putative SSR markers. A large number of SSRs were discovered where the SSR numbers varied among maize chromosomes. The mononucleotide repeats and compound were excluded from the analysis

because of the abundance of poly A/T repeats. Table 1 shows the summary of putative SSRs from the genome sequences of cultivar B73. A total of 198, 414 SSRs loci were detected with the highest SSRs distribution was from chromosome one with a total number of 13, 993 SSRs identified. Table 2 summarized the information of identified SSRs from maize chromosomes. The chromosome ten had the minimum number of SSRs distribution with a total of 6928 SSRs identified. This could be due to the chromosome length whereby it has the minimum length of chromosome when compared to others chromosome. From the analysis, there are no significant differences in the SSR abundance composition within the ten chromosomes. The SSR density of maize varies greatly among chromosomes and thus indicates a wide usage of maize genome sequences.

Table 1 Summary of putative maize SSRs

Total number of sequences examined (bp)	2059701728
Total number of identified SSRs	198, 414
Total number of identified SSRs	94, 534
(excluded mononucleotide & compound)	

Table 2 Distribution of SSRs on each maize chromosome

Chr	Chr length (Mbp)	Total number of SSRs	di	tri
1	301.43	13,993	8086	5320
2	237.89	11,055	6335	4252
3	232.22	10,631	6086	4131
4	242.02	11,063	6326	4299
5	217.93	10,077	5672	3994
6	169.38	7859	4424	3106
7	176.81	7969	4508	3122
8	175.35	7842	4475	3052
9	157.02	7117	4075	2755
10	149.63	6928	3901	2742
Total	2059.68	94,534	53,888	36,773

Among the derived SSR repeats, the dinucleotide repeat were the most abundant repeat with a frequency of 53,888, 57%, followed by trinucleotide (36,773, 38.9%), tetranucleotide (2622, 2.77%), pentanucleotide (801, 0.85%) and hexanucleotide repeats (450, 0.48%). These number of repeats showed that the frequency distribution of different types of repeats in maize. Our findings are in agreement with the situation in castor oil [4] and wheat rust [5] where dinucleotide repeats were also the most prevalent repeat motif detected.

Over 274 different repeat motifs were identified, of which the most frequent were AG/CT (48%). The top 10 most abundant of SSR repeat motif with different levels of repeats is shown in Table 3. The higher percentage of AG/CT and AT/AT in the non-coding sequences shows a higher probability of being mutated.

From the analysis of dinucleotide repeats motif shows that four types repeat motifs; AG/CT, AT/AT, AC/GT and CG/CG were found in the genome. Of the dinucleotide, AG/CT was the most common repeat motif, representing 48%, followed by 37% AT/AT repeat motif, 11% AC/GT repeat motif and 4% CG/CG repeat motif. Figure 1 shows the distributions and percentages of different motifs in dinucleotide repeats.

The predominance repeat motif in trinucleotide was observed in AGC/CTG motifs, accounted for 22% of the total trinucleotide repeats. Ten types of trinucleotide repeat motifs were found in the genome. Figure 2 shows the distributions of repeat motifs in trinucleotide repeats.

The differences in the SSR criteria affected the abundance of repeat motif across plant taxa. Nevertheless, AT/TA is not usually used to develop markers due to the self-complementary nature to form dimers [18].

Based on the length of SSRs repeat motif, the SSRs were categorized into two groups; Class I SSRs contained repeats ≥20 bp and class II contained repeats ≥10 nucleotides and <20 nucleotides in length [19]. Out of 94,534 SSR repeats, 20, 133 repeats were categorized in class I whereas 74,401 repeat were

categorized in class II. The longer SSR repeats, the higher chances to obtain polymorphic SSRs [19].

Table 3 Top 10 SSR motifs with repeat number

Repeats Motif	Total
AG/CT	31085
AT/AT	24170
AGC/CTG	9628
AC/GT	6949
CCG/CGG	5912
ACG/CGT	4945
ATC/ATG	4258
AAG/CTT	4230
AAT/ATT	3922

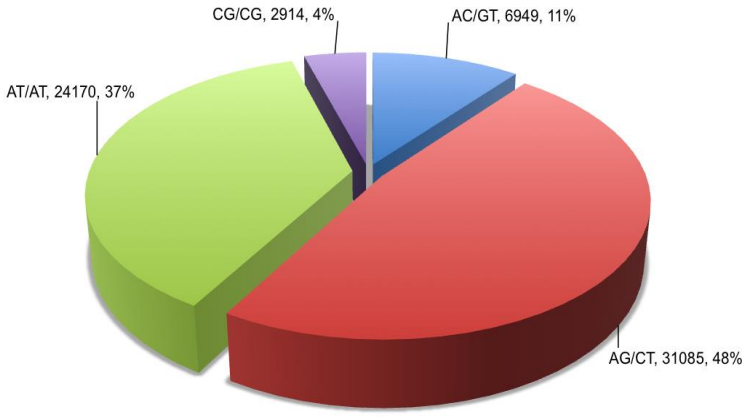


Figure 1 Percentage of different motifs in dinucleotide repeats

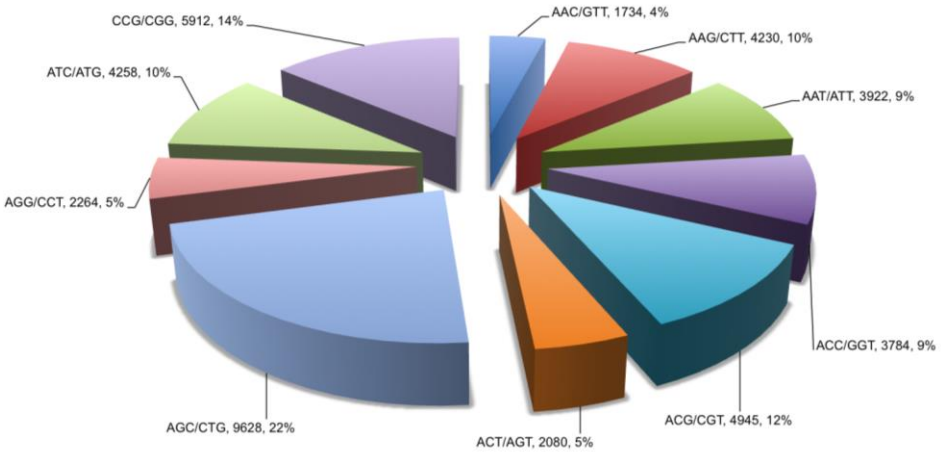


Figure 2 Percentage of different motifs in trinucleotide repeats

3.2 SSRs Primer Design and Database Development

In this study, we mainly focus on dinucleotide and trinucleotide repeats for SSRs primer development. From the analysis, a subset of 2239 putative SSR sites were randomly chosen to validate the usefulness of SSR primers in developing SSRs marker. Of these, a total of 99 primer pairs were randomly chosen for primer validation. Of the markers tested, 70 (70%) primer pairs were successfully produced allelic amplification product and therefore were validated as candidate SSR markers. Description of the top ten primer pairs and their ability to convert into SSR markers is showing in Table 4. It has been predicted that either the short flanking sequences or high GC content in SSRs loci affected the 30% of non suitable SSRs primer pairs [20]. The proportion of SSR primers that successfully amplified the tested DNA can be

used to measure the rate of conversation of SSR primer into SSR markers [21, 22]. Thus, the proportion rate obtained from this study was 71.71% and this value might be correlated with genome size as observed by Garner, 2002 [23]. However, further assessment to present polymorphism of those SSR markers in maize genotypes will be carried out before it can be utilized in genetic diversity study of maize. The 30% SSRs primer pairs are not suitable for primer design either because of the flanking sequences are too short or containing high GC content.

Table 4 Description of top ten validated primer pairs from maize SSRs loci

PrimerID	SSR Repeats & Num of repeats	Forward primer (5'-3')	Reverse primer (5'-3')	Tm° C	Range size of frag. (bp)
C2-6	(GA)8	CAATAGTGGGACAGTGCGG	AATCTCCGATGGTCTGGGC	58.3	100-200
C2-8	(GAG)10	AGGCGGATGTGGAAGCC	GAGTGACGAGGTACGCCC	51	100-200
C2-9	(GA)19	GGCTCTGTGATGGCGAATG	GAGGGCCAGTCCATCGG	51	200-300
C2-10	(ATC)6	CCTGCTCGGAGGCCATAAG	TCGCACAACTGATCTTGCC	55	300-400
C2-11	(GA)7	GGATTGGGCGGTAACCAAG	TGGCTTGCCATCACTATTCTG	44.8	200-300
C2-12	(GA)13	AGGAAATCCCGCAGATCCC	GTCCACAGAATCCCGGAGG	44.8	300-400
C2-13	(AGA)9	CCCTGTGGATGTCCCTCG	TTCCCAACGCATGCAGAAC	58.3	300-400
C2-14	(TC)16	AGGTTGCTTTGTACAGGGTTC	CTGACGCTAAATTTCATGGTAAATG	45.5	200-300
C2-15	(AT)18	CCAGTACTTGGCTGTTGCC	ATTCGCGAGGAGGTAATGC	51	200-300
C2-16	(TAA)12	CCACTAGAGCTAACGTGGC	AGATAGGAGGTGCCGACTTG	58.3	300-400

All the information on 2239 putative SSRs loci were stored in an in house database using Biomart version 0.7 [24]. The SSRs information can be search by chromosome, motif and primer identification (ID). Others characteristics such as repeats, motif, temperature, and SSR position can be retrieved using this database. Figure 3 shows the graphical user interface of maize SSRs database.

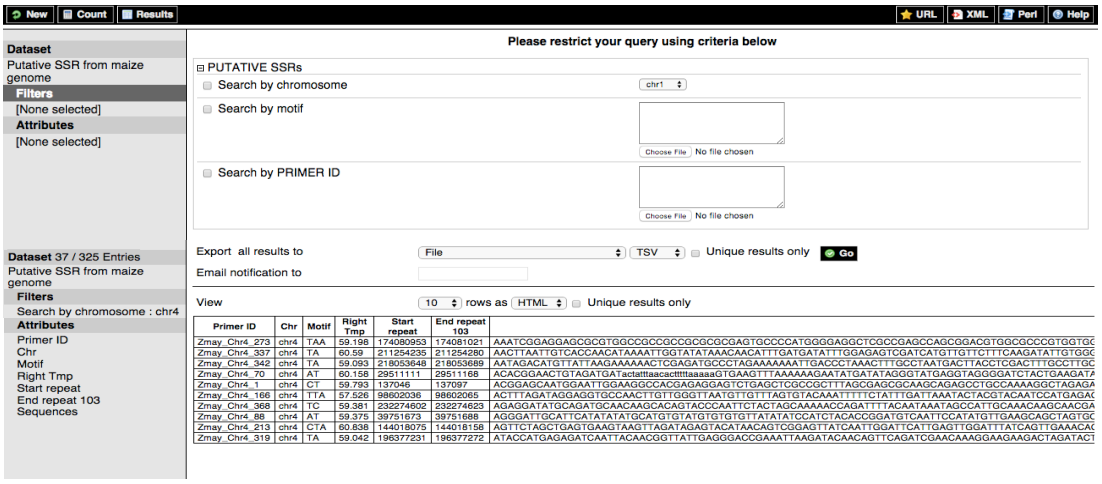


Figure 3 The Biomart user interface of in house maize SSRs database

4.0 CONCLUSION

In summary, this study provides a brief idea on the approach to develop SSR markers using *in silico* approach. In addition, by combining sequencing technology and bioinformatics, we were able to identify an abundance of SSRs in maize B73 genome. The developed pipeline provides a cost and time effective method in developing SSR markers from large amount sequences data. The candidate SSRs described here will be genotyped using genotyping platform such as ABI 373XL. This maize SSRs will be served as a useful resource for maize genetic studies such as the assessment of genetic diversity, the identification of germplasm, construction of linkage maps and marker trait association.

Acknowledgement

This author would like to acknowledge and thank EPP14 Pembangunan Industri Benih for financial support.

References

[1] FAOSTAT 2010. [Online] From: http://faostat3.fao.org/home/E. [Acessed on 8 September 2015].

[2] Ibitoye, D. O. and Akin-Idowu, P. E. 2011. Marker-assisted-selection (MAS): A Fast Track to Increase Genetic Gain in Horticultural Crop Breeding. *African Journal Biotechnology*. 9(52): 8889-8895.

[3] Duran, C., Singhania, R., Raman, H., Batley, J. and Edwards, D. 2013. Predicting Polymorphic EST-SSRs in Silico. *Molecular Ecology Resources.* 13: 538-545.

[4] Meilian, T., Kun, W., Lei, W., Mingfang, Y., Zhidan, Z., Jing, X,. Yang, Z., Xuekun, Z., Chunling, F., Jianfeng, X., Lijun, W. and Xingchu, Y. 2014. Developing and Characterising Ricinus Communis SSR Markers by Data Mining of Whole-Genome Sequences. *Mol Breeding*. 34: 893-904.

[5] Rajender, S., Bharati, P., Mohd, D., Sonia, S., Pradeep, S. and Ravish, C. Mining and Survey of Simple Sequence Repeats In Wheat Wheat Rust Puccinia sp. *Bioinformation*. 7(6): 291-295.

[6] Susan, R., M., Leonid, T., Yunbi, X., Katarzyna, B., L., Karen, C., Mark, W., Binying, F., Reycel, M., Zhikang, Li., Yongzhong, X., Qifa, Z., Izumi, K., Masahiro, Y., Robert, F., Genevieve, D., David, S., Samuel, C., Doreen, W. and Lincoln, S. 2002. Development and Mapping of 2240 New SSR Markers for Rice (Oryza sativa L.). *DNA Research*. 9 (6): 199-207.

[7] The Potato Genome Sequencing Consortium. 2011. Genome Sequence and Analysis of the Tuber Crop Potato. *Nature*. 475(7355): 189-195.

[8] Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., Cannon, S., Baek, J., Farmer, A. D., Gaur, P. M., Soderlund, C., Penmetsa, R. V., Xu, C., Bharti., A. K., He, W., Winter, P., Zhao, S., Hane, J. K., Carrasquilla-Garcia, N., Condie, J. A., Upadhayaya, H. D., Luo, M. C., Thudi, M., Gowda, C. L., Singh, N. P., Lichtenzveig, J., Gali, K. K., Rubio, J., Nadarajan, N., Dolezel, J., Bansal, K. C., Xu, X., Edwards, D., Zhang, G., Kahl, G., Gil, J., Singh, K. B., Datta, S.K., Jackson, S. A., Wang, J., Cook, D. R. Draft Genome Sequence of Chickpea (Cicer Arietinum) Provides a Resource for Trait Improvement. 2013. *Nature Biotechnology*. 31(3): 240-246.

[9] Stieneke, D. L., Eujayl, I. Imperfect SSR Finder. [Online]. From: http://www.fresnostate.edu/ssrfinder/). [Accessed on 7 July 2015].

[10] Thiel, T., Michalek, W., Varshney, R. K. and Graner, A. 2003. Exploiting EST Databases for the Development of CDNA Derived Microsatellite Markers in Barley (*Hordeum vulgare* L.). *Theor Appl Genet*. 106(3): 411-422.

[11] Xuewen, W., Peng, L,. and Zhaopeng, L. 2013. GMATo: A Novel Tool for the Identification and Analysis of Microsatellites In Large Genomes. *Bioinformation*. 9(10): 541-544.

[12] Luciano, C., Dario, A. P., Velci, Q., D., S., Mauricio, M., K., Fernando, I., F., D., C. and Antonio, C., D., O. 2008. SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. *International Journal of Plant Genomics*. 28: 9.

[13] Schaeffer, M. L., Harper, L. C., Gardiner, J. M., Andorf, C. M., Campbell, D. A., Cannon, E. K. S., Lawrence, C. J. 2011. MaizeGDB: Curation and Outreach Go Hand-In-Hand. *Database: The Journal of Biological Databases and Curation*.

[14] Aaron R. Q. and Ira, M. H. 2010. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics*. 26(6): 841-842.

[15] Rozen, R. and Skaletsky, H. J. 2000. Primer3 on the WWW for General Users and for Biologist Programmers. In Bioinformatics Methods and Protocols. *Methods in Molecular Biology*.365-386

[16] Jie, X., Ling, L., Yunbi, X., Churun, C., Tingzhao, R., Farhan, A., and Shufeng, Z., Fengkai, W., Yaxi, L., Jing, W., Moju, C. and Yanli, L. 2013. Development and Characterization of Simple Sequence Repeat Markers Providing Genome-Wide Coverage and High Resolution in Maize. *DNA Research*. 20: 497-509.

[17] Jingtao, Q. and Jian, L. 2013. A Genome-Wide Analysis of Simple Sequence Repeats in Maize and the Development of Polymorphism Markers from Next-Generation Sequence Data. *BMC Research Notes.* 6: 403.

[18] Sosinski, B., Gannavarapu, M., Hager, L. D., Beck, L. E., King, G. J., Ryder, C. D., Rajapakse, S., Baird, W. V., Ballard, R. E. and Abbott, A. G. 2000. Characterization of Microsatellitemarkers in Peach [*Prunus persica (L.)*. *Theor Appl Genet*. 101(3): 421-428.

[19] Singh, R., Sheoran, S., Sharma, S. and Chatrath R. 2011. Analysis of Simple Sequence Repeats (Ssrs) Dynamics in Fungus Fusarium graminearum. *Bioinformation.* 5(10): 402-404.

[20] Yi, W., Chao, Y., Qiaojun, J., Dongjie, Z., Shuangshuang, W, Yuanjie, Y. and Long, Y. 2015. Genome-wide Distribution Comparative and Composition Analysis of the SSRs in Poaceae. *BMC Genetics.* 16: 18.

[21] Hendre, P. S, Phanindranath, R., Annapurna, V., Lalremruata, A. and Aggarwal, R. K. 2008. Development of New Genomic Microsatellite Markers from Robusta Coffee (Coffea canephora Pierre ex A. Froehner) Showing Broad Crossspecies Transferability and Utility in Genetic Studies. *BMC Plant Biol*. 8(51): 1-19

[22] Missio, R. F, Caixeta, E. T., Maciel-Zambolim, E., Zambolim, L. and Sakiyama, N. S. 2009. Development and Validation of SSR Markers for Coffea arabica L. *Crop Breed Appl Biotechnol*. 9: 361-371.

[23] Garner, T. W. 2002. Genome Size and Microsatellite: The Effect of Nuclear Size on Amplification Potential. *Genome* 45: 212-215.

[24] Damian, S., Syed, H., Benoit, B., Richard, H., Darin, L., Gudmundur, T. and Arek, K. 2009. BioMart–Biological Queries Made Easy. *BMC Genomics*. 10: 22.