RESEARCH ARTICLE Bhartiya Krishi Anusandhan Patrika, Volume 38 Issue 4: 360-368 (December 2023)

Comparative Analysis of Predicted SSR Sequences and CpG Islands to Discover Evolutionary Relics of Sex-chromosomes in Divergent Animal Species

Barinder Singh Grewal[1], Shilpa Tewari[2], Håkon Hægland[3], C.S. Mukhopadhyay[4]

10.18805/BKAP673

ABSTRACT

Background: DNA markers have high occurrence and mutation rates and are generally located around the controlling regions of some tissue-specific genes and housekeeping genes that can change the expression pattern. Microsatellites and CpG islands are stretches of DNA with repeats and are known to influence gene expression.

Methods: In the present study, these DNA markers are mined and an *In silico* comparison was carried out to understand their occurrence pattern and distribution frequency in sex chromosomes (X and Y) of 12 different animal species using Perl and R programming pipelines.

Result: It was found that female-dominant X chromosomes had higher occurrence and distribution frequencies for these DNA markers than that of male-dominant sex chromosome *i.e.* Y which means that the former has a higher number of the evolutionary sites.The density of DNA markers however, showed remarkable variation for different animal species. The results obtained need validation through wet-lab experimentation. Tri- and hexa-nucleotide repeats are more abundant in exons, whereas other repeats are more abundant in non-coding regions.

Key words: CpG island, DNA markers, Microsatellite, Sex chromosomes.

INTRODUCTION

In mammals, the sex chromosomes are generally dimorphic. The X chromosomes are usually of large size and gene-rich while Y chromosomes are comparatively of smaller in size and heterochromatic in nature and are almost completely different but they at small homologous region (pseudo autosomal region) they paired with each other. Genetic markers such as CpG and microsatellites plays important role in evolution of sex chromosomes. Many biological processes significantly affect the functionality of DNA. One such process is methylation that is involved in X-chromosome inactivation (XCI) especially at promoter-proximal regions that are enriched with CpG islands (Duncan *et al*., 2018). The Y chromosome accumulates repeat sequences that are epigenetically repressed, results an epigenetic dispute with Y gene expression and hence possible accelerates the Y chromosome degeneration. Ageing causes the loss of Y heterochromatin, which activates transposable elements and reduces male lifespan. In placental mammals namely eutherians and marsupials X-chromosome inactivation has evolved *via* two different non-coding RNA molecules (Muyle *et al.,* 2021).

Both SSRs and CpG islands are present in most of the organisms and are key elements in structural organization of genomes and their function and may be related with disease states, their systematic analysis has not been reported. The study of repeat density and its distribution pattern in the genome is expected to help in understanding their significance. The accumulating evidences suggested that SSRs plays role in gene expression regulation (Kunzler *et al.,* 1995). In the present study, *in silico* mining of the nucleotide motifs (SSR regions and CpG islands) has been targeted in the entire genome to explore the evolutionary relics of sex-chromosome constitute in divergent species of animals. The accessibility of complete genome sequences for many organisms through nucleotide databases has made it possible to carry out genome-wide analysis. *In silico* comparative analysis of DNA markers may be helpful in understanding their role and abundance in the coding, as well as non-coding, regions of the genome may give us some clue to the function of SSRs in gene regulation.

[1]Centre For One Health, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana-141 004, Punjab, India.
[2]College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana-141 004, Punjab, India.
[3]Uni Research, CIPR, P.O. Box 7810, 5020 Bergen, Norway.
[4]Department of Bioinformatics, College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana-141 004, Punjab, India.

Corresponding Author: C.S. Mukhopadhyay, Department of Bioinformatics, College of Animal Biotechnology, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana-141 004, Punjab, India. Email: csmbioinfo@gmail.com

How to cite this article: Grewal, B.S., Tewari, S., Hægland, H. and Mukhopadhyay, C.S. (2023). Comparative Analysis of Predicted SSR Sequences and CpG Islands to Discover Evolutionary Relics of Sex-chromosomes in Divergent Animal Species. Bhartiya Krishi Anusandhan Patrika. 38(4): 360-368. doi: 10.18805/BKAP673.

Submitted: 18-08-2023 **Accepted:** 03-01-2024 **Online:** 19-01-2024

MATERIALS AND METHODS

The nucleotide sequences of sex-chromosomes of twelve selected mammalian species, namely, *Gallus gallus*, *Meleagris gallopavo*, *Anopheles gambie, Drosophila melanogaster*, *Callithrix jacchus*, *Chlorocebus sabaeus*, *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus, Bos taurus* and *Sus scrofa* were downloaded in the Fasta format from the nucleotide database of National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/genome). The downloaded sex chromosomes were classified into five groups according to their order type (Table 1).

Microsatellite prediction

Microsatellite prediction was done with a Perl-based MISA (Microsatellite analysis) tool accessed under (https://webblast.ipk-gatersleben.de/misa/). Fig 1 showing the flowchart of microsatellite prediction.

CpG island prediction

For the prediction of CpG islands, we assumed a minimum length of 200 nt, the minimum content of C+G 55% and the ratio between the frequency of observed and expected CpG sites to be at least 0.65. The downloaded chromosome

Table 1: Species used for microsatellite and CpG island prediction.

Animal species	Common name	Group	Order	Y_chr size*	X_chr size**
Gallus gallus	Red junglefowl	Avian	Galliformes	82363669$	1248174$$
Meleagris gallopavo	Wild turkey		Galliformes	68461266$	260627$$
Anopheles gambiae	Mosquitoes	Insects	Diptera	10,429	24393108
Drosophila melanogaster	Fruit fly		Diptera	3667352	23542271
Callithrix jacchus	New world monkey	Primates	Primates	2,853,901	142,054,208
Chlorocebus sabaeus	Green monkey		Primates	6181219	130038232
Homo sapiens	Humans		Primates	57,227,415	156040895
Pan troglodytes	Chimpanzee		Primates	263,42,871	156848144
Mus musculus	House mouse	Rodents	Rodentia	91,744,698	171,031,299
Rattus norvegicus	Brown rat		Rodentia	3,310,458	159,970,021
Bos taurus	Cattle	Even-toed ungulates	Artiodactyla	433,00,181	148823899
Sus scrofa	Wild boars		Artiodactyla	1,637,716	144,288,218

*Y Chromosome size; **X Chromosome size; $ W Chromosome size; $$ Z Chromosome size.

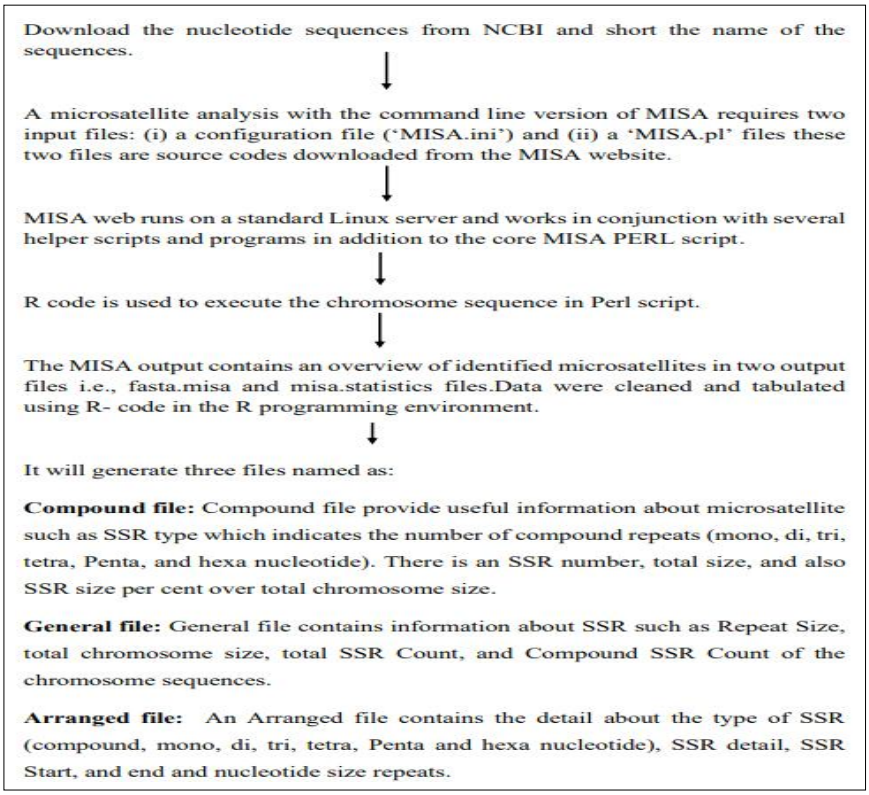


Fig 1: Flowchart showing the microsatellites prediction.

sequences were subjected to notepad++ for further modification. Then the sequences were subjected to Perl code for predicting the statistical data. The statistical data were subjected to the R-programming environment for further cleaning and getting the predicted data Fig 2 showing the flowchart of CpG island prediction.

RESULTS AND DISCUSSION

Sex chromosomes of different animal species *viz*. species belonging to ruminants, other mammals, avians, *etc* were analyzed for the distribution of microsatellites and CpG island.

Microsatellite prediction in the avian group

W chromosome

MISA has predicted higher monomeric and lower hexameric SSRs in W chromosomes of *Gallus gallus* and *Meleagris gallopavo* species (Fig 3).

Z chromosome

Gallus gallus has higher numbers of all predicted SSRs based on the Z chromosome and sums to 291 while *Meleagris gallopavo* comparatively has only 58 SSRs (Fig 4).

CpG island prediction in the avian group

W chromosome

Gallus gallus has a greater average island length of 569.12, Variation in island length is more as compared to *Meleagris gallopavo* (Table 2).

Z chromosome

Gallus gallus has a greater average island length of 743.57, Variation in island length is more as compared to *Meleagris gallopavo* (Table 3).

Microsatellite prediction in the insect group

X chromosomes

X-chromosomes of *Anopheles gambie* and *Drosophila melanogaster* are closer in size. *Anopheles gambie* has a smaller sized X chromosome than *Drosophila melanogaster* but comparatively contains a large number of SSRs (Fig 5).

Y chromosomes

Surprisingly, *Anopheles gambie* has only mononucleotide repeat motifs. All types of microsatellites were present in *Drosophila melanogaster* but they were found to be fewer in numbers (Fig 6).

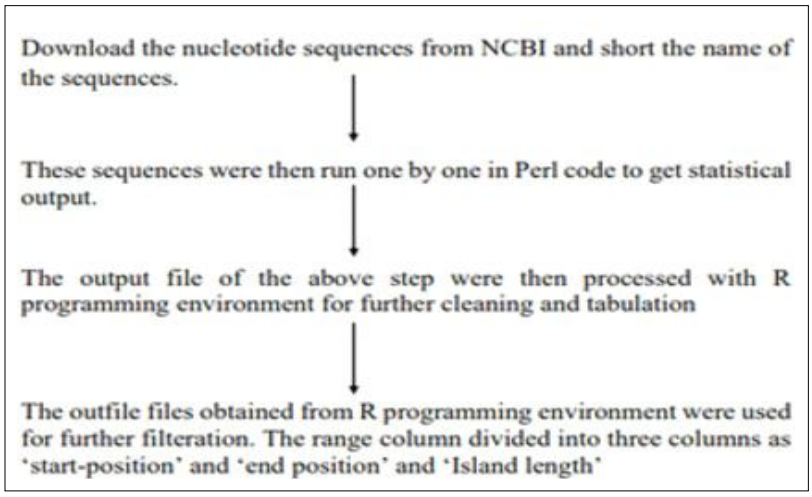


Fig 2: Flowchart showing the CpG Island prediction.

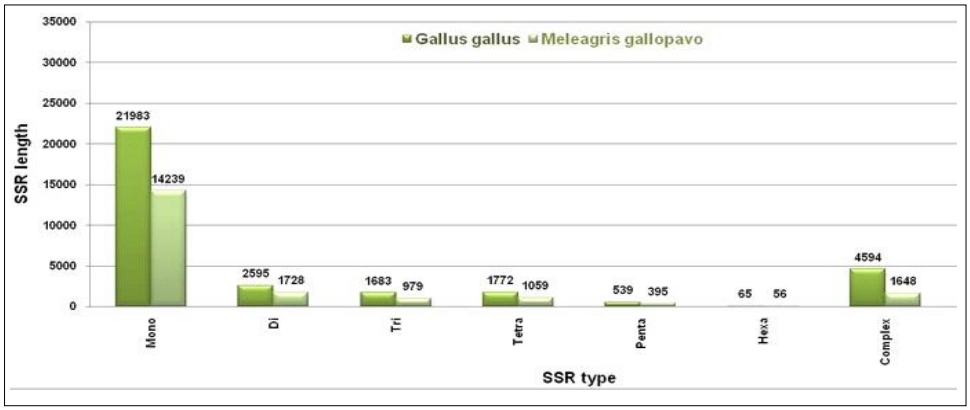


Fig 3: Predicted SSRs with respective length in W-chromosome.

Table 2: Final parameters of CpG island in W-chromosome.

Features	*Gallus gallus*	*Meleagris gallopavo*
Average island length	569.12	564.88
The standard error (Island length)	4.38	17.82
Island number	2433	83.00
Average G+C per cent	51.97	50.5
Standard error G+C per cent	0.06	0.14
Average CpG per cent	4.7	5.25
Standard error CpG per cent	0.01	0.12
Average ratio	0.72	0.9
Standard error ratio	0.00	0.03
Minimum island length	500.00	500.00
Maximum island length	5230.00	1653.00

Table 3: Final parameters of CpG island in Z-chromosome.

Features	*Gallus gallus*	*Meleagris gallopavo*
Average island length	743.57	630.51
The standard error (Island length)	7.1	5.63
Island number	4635.00	1799.00
Average G+C per cent	55.3	51.74
Standard error G+C per cent	0.08	0.08
Average CpG per cent	5.76	5.26
Standard error CpG per cent	0.02	0.02
Average ratio	0.76	0.81
Standard error ratio	0.00	0.00
Minimum island length	500.00	500.00
Maximum island length	6949.00	2330.00

CpG island prediction in insect group

X chromosomes

The average island length of *Anopheles gambiae* in the X-chromosome is 634.24 and *Drosophila melanogaster* has an average island length of 619.42 (Table 4).

CpG island prediction in insect group

Y chromosomes

The average island length of *Anopheles gambiae* in the Y-chromosome is 535 and *Drosophila melanogaster* has an average island length of 601.02 (Table 5).

Microsatellite prediction in primates

X chromosomes

X-chromosomes of animals belonging to this group have comparable sizes. All of them have a higher number of mono nucleotide repeat motifs. Homo sapiens among all these animal species contains the highest number of all types of SSRs (Fig 7).

Y chromosomes

A comparable number of microsatellite motifs was found in Y-chromosomes irrespective of their chromosome sizes. The total number of microsatellites of different types follow decreasing order in the range of Mono>di>tri>tetra> penta>hexa (Fig 8).

CpG island prediction in primates

X chromosomes

Callithrix jacchus has a greater average island length of 634.76, Variation in island length is more as compared to

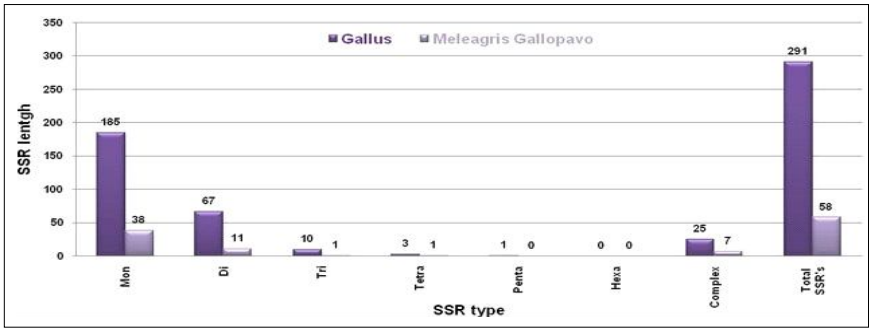


Fig 4: Predicted SSRs with respective lengths in Z-chromosome.

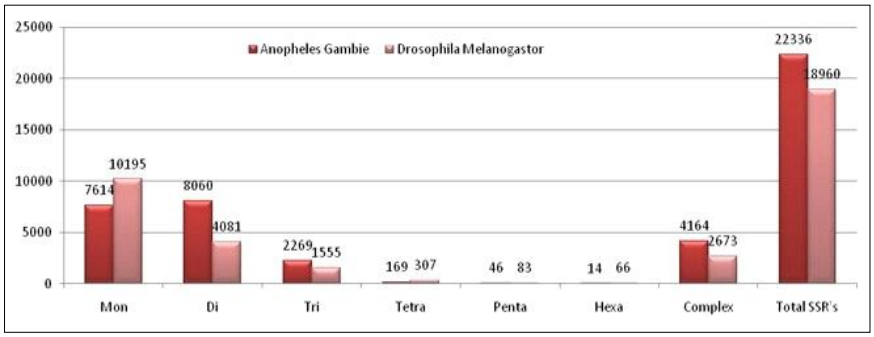


Fig 5: Predicted SSRs with respective lengths in the X-chromosome.

Table 4: Final parameters of CpG island in X-chromosome.

Features	*Anopheles gambiae*	*Drosophila melanogaster*
Average island length	634.24	619.42
The standard error (Island length)	1.78	1.94
Island Number	50388.00	31613.00
Average G+ C per cent	50.46	50.69
Standard error G+ C per cent	0.01	0.01
Average CpG per cent	6.5	5.64
Standard error CpG per cent	0.01	0.01
Average ratio	1.03	0.89
Standard error ratio	0.00	0.00
Minimum island length	500.00	500.00
Maximum island length	9249.00	7881.00

Chlorocebuss abaeus, Homo sapiens and *Pan troglodytes* (Table 6).

Y chromosomes

Callithrix jacchus has a greater average island length of 643.37, Variation in island length is more as compared to *Chlorocebus sabaeus, Homo sapiens and Pan troglodytes* (Table 7).

Microsatellite prediction in rodents

X chromosome

Both these animal species have comparable X-chromosome sizes and similarly, have higher monomeric and lower hexameric types of microsatellites. But *Mus musculus* significantly contains a five times higher number of SSRs than *Rattus norvegicus* (Fig 9).

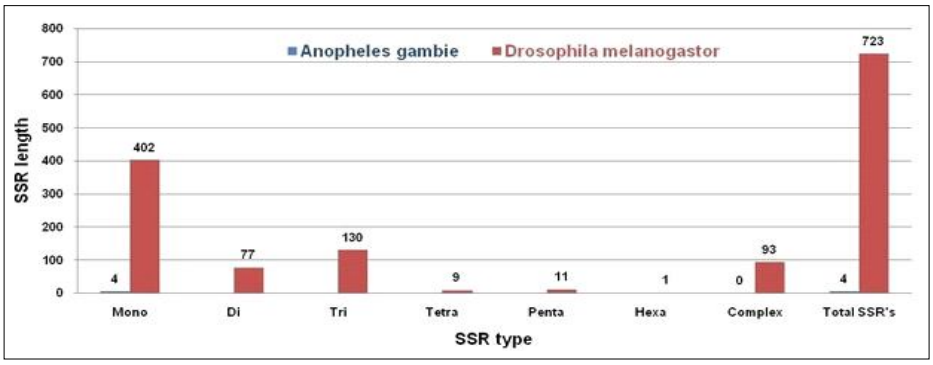


Fig 6: Predicted SSRs with respective lengths in the Y-chromosome.

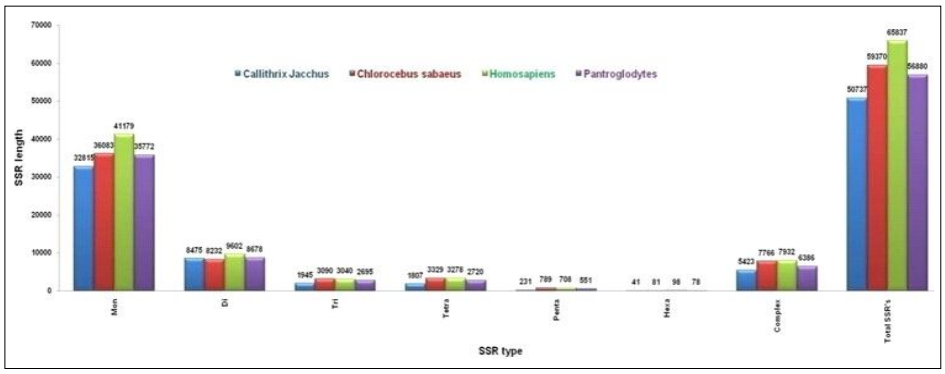


Fig 7: Predicted SSRs with respective lengths in X-chromosome.

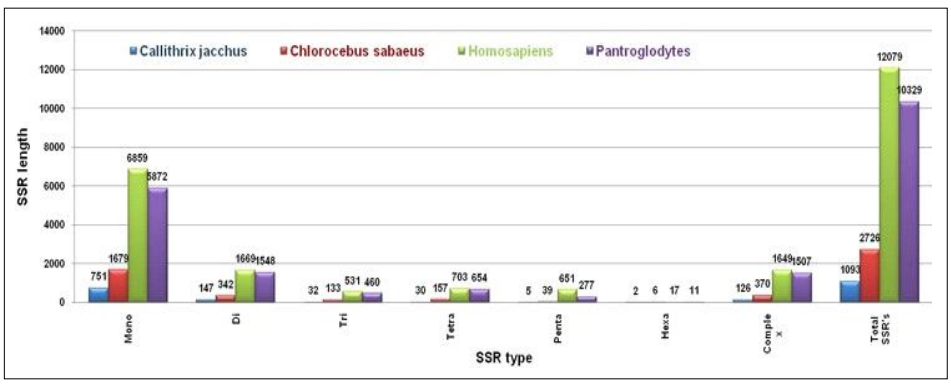


Fig 8: Predicted SSRs with respective lengths in the Y-chromosome.

Y chromosome

On the other side, the Y chromosome of *Mus musculus* is approximately 2 and half times that of *Rattus norvegicus* but both of them contain a comparable number of all types of SSRs (Fig 10).

X chromosome

The average island length of *Mus musculus* in the X-chromosome is 588.17 and *Rattus norvegicus* has an average island length of 596.19 (Table 8).

Table 5: Final parameters of CpG island in Y-chromosome.

Features	*Anopheles gambiae*	*Drosophila melanogaster*
Average Island length	535.00	601.02
The standard error (Island length)	34.00	7.03
Island number	3.00	3270.00
Average G+C per cent	50.41	50.37
Standard error G+C per cent	0.21	0.02
Average CpG per cent	6.07	5.73
Standard error CpG per cent	0.13	0.02
Average ratio	0.97	0.91
Standard error ratio	0.03	0.00
Minimum island length	500.00	500.00
Maximum island length	603.00	7406.00

Y chromosome

The average island length of *Mus musculus* the Y-chromosome is 548.88 and *Rattus norvegicus* is 560.46 (Table 9).

Microsatellite prediction in Even-toed ungulates

X chromosome

Different type nucleotide repeats were exceptionally higher in numbers in the Y chromosome of *Bos taurus*, while the microsatellite repeat motifs were comparable in X-chromosomes (Fig 11).

Y chromosome

SSRs mined from Y chromosomes follow the decreasing order of several repeats in terms of mon>di>tri>tetra> penta>hexa. Complex type SSRs were considerably lower in *Sus scrofa* (Fig 12).

X chromosome

The average island length of *Bos taurus* in the X-chromosome is 701.35and *Sus scrofa* has an average island length of 580.44 (Table 10).

Y chromosome

The average island length of *Bos taurus* in the Y-chromosome is 545.86 and *Sus scrofa* has an average island length of

Table 6: Final parameters of CpG island in X-chromosome.

Features	*Callithrix jacchus*	*Chlorocebus sabaeus*	*Homo sapiens*	*Pan troglodytes*
Average island length	634.76	628.83	611.57	608.97
The standard error (Island length)	4.3	4.93	3.62	3.8
Island number	4426.00	4232.00	6770.00	4379.00
Average G+C per cent	55.66	55.17	55.68	54.53
Standard error G+C per cent	0.09	0.1	0.08	0.09
Average CpG per cent	5.38	5.26	5.32	5.24
Standard error CpG per cent	0.02	0.02	0.02	0.02
Average ratio	0.7	0.7	0.7	0.72
Standard error ratio	0.00	0.00	0.00	0.00
Minimum island length	500.00	500.00	500.00	500.00
Maximum island length	4007.00	4473.00	4472.00	2991.00

Table 7: Final parameters of CpG island in Y-chromosome.

Features	*Callithrix jacchus*	*Chlorocebus sabaeus*	*Homo sapiens*	*Pan troglodytes*
Average island length	643.37	559.51	569.93	570.2
The standard error (Island length)	20.72	10.69	5.41	6.38
Island number	268.00	257.00	1756.00	997.00
Average G+C per cent	56.92	52.07	53.89	55.66
Standard error G+C per cent	0.39	0.24	0.14	0.19
Average CpG per cent	5.58	4.78	4.89	5.29
Standard error CpG per cent	0.08	0.05	0.03	0.04
Average ratio	0.7	0.72	0.71	0.69
Standard error ratio	0.00	0.01	0.00	0.00
Minimum island length	500.00	500.00	500.00	500.00
Maximum island length	3242.00	1950.00	3420.00	1987.00

Table 8: Final parameters of CpG island in X-chromosome.

Features	*Mus musculus*	*Rattus norvegicus*
Average island length	588.17	596.19
The standard error (Island length)	3.46	3.68
Island number	4545.00	4465.00
Average G+C per cent	54.68	53.03
Standard error G+C per cent	0.09	0.07
Average CpG per cent	5.26	4.94
Standard error CpG per cent	0.02	0.01
Average ratio	0.72	0.73
Standard error ratio	0.00	0.00
Minimum island length	500.00	500.00
Maximum island length	3476.00	4484.00

Table 9: Final parameters of CpG island in Y-chromosome.

Features	*Mus musculus*	*Rattus norvegicus*
Average island length	548.88	560.46
The standard error (Island length)	3.39	12.67
Island number	1516.00	110.00
Average G+C per cent	52.21	52.7
Standard error G+C per cent	0.11	0.37
Average CpG per cent	4.79	4.73
Standard error CpG per cent	0.02	0.06
Average ratio	0.71	0.7
Standard error ratio	0.00	0.01
Minimum island length	500.00	500.00
Maximum island length	1568.00	1059.00

567.28 which means *Sus scrofa* has a greater average island length (Table 11).

CpG islands are found almost everywhere in vertebrate genomes. Even though many tissue-specific genes lack CpG islands, it is becoming clear that they do exist in all commonly expressed genes, as well as a large number of tissue-specific genes with CpG islands can be found at the 5′ or 3′ ends of genes. CGIs are a fragmented but unified DNA sequence family whose members serve as genomic platforms for controlling transcription at their associated promoters. These characteristics are based on common DNA sequences traits, such as CpG richness and a higher-than-usual G+C concentration (Thomson *et al*., 2010). In addition, SSR sequences possess most of the desirable attributes of molecular markers, including information content, unambiguous designation of alleles, neutral selectively (although they can be subjected to hitch-hiking effects), high reproducibility, codominance and fast and easy assaying of genotypes and therefore microsatellite markers or SSR have proved to be very useful for cultivar identification, pedigree analysis and the evaluation of genetic distance between organisms (Priolli *et al*., 2002) and genetic mapping (Yu *et al*., 2000). To date, most macropod microsatellites have been isolated using laboratory-based techniques, including standard bacteria screening and microsatellite enrichment libraries (Karagyozov *et al*., 1993; Hakki and Akkaya, 2000). These methods can be time-consuming and unpredictable, with no guarantees of obtaining the numbers or types of markers desired. These approaches are effectively random samples of the genome

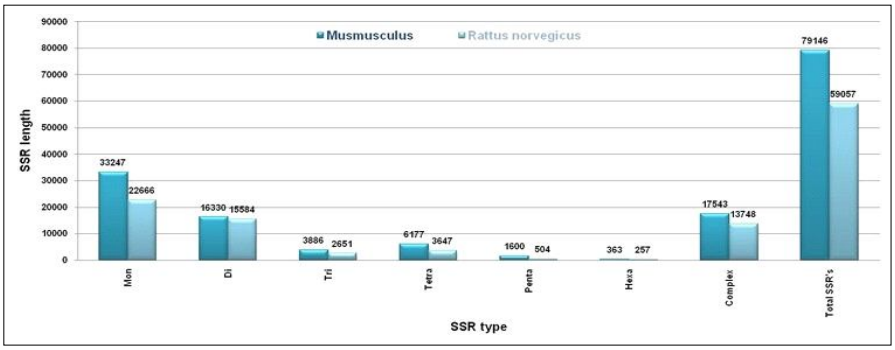


Fig 9: Predicted SSRs with respective lengths in the X-chromosome.

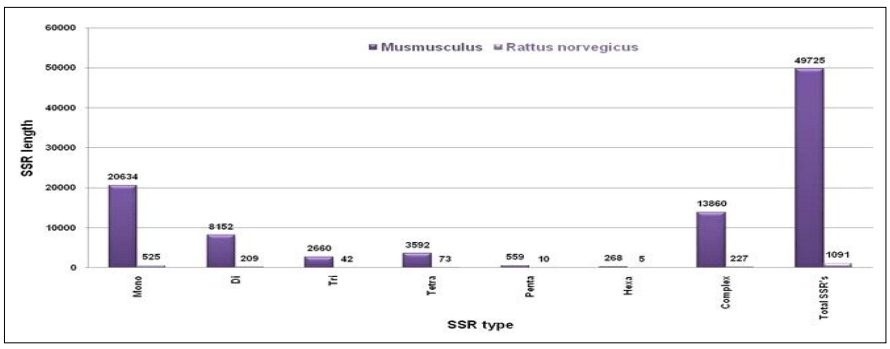


Fig 10: Predicted SSRs with respective lengths in the Y-chromosome.

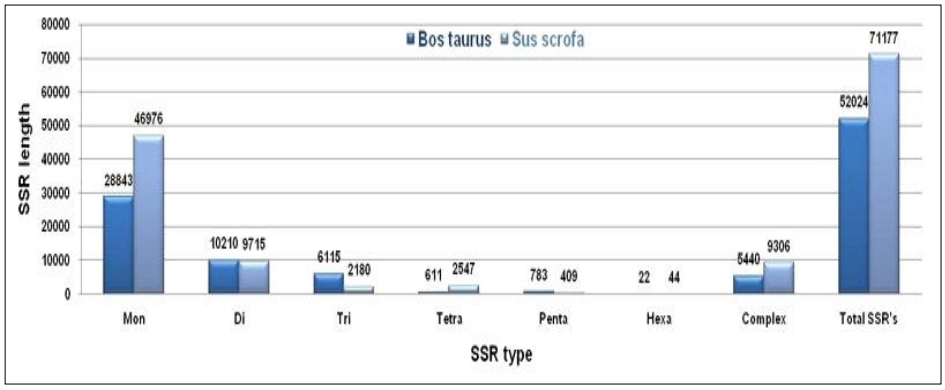


Fig 11: Predicted SSRs with respective lengths in the X-chromosome.

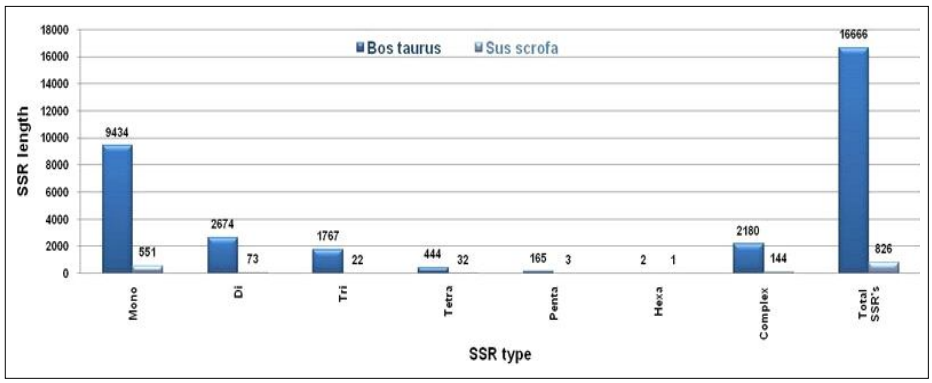


Fig 12: Predicted SSRs with respective lengths in the Y-chromosome.

Table 10: Final parameters of CpG island in X-chromosome.

Features	*Bos taurus*	*Sus scrofa*
Average island length	701.35	580.44
Standard error (Island length)	10.07	2.02
Island number	1516.00	13539.00
Average G+C per cent	57.61	54.5
Standard error G+C per cent	0.15	0.05
Average CpG per cent	5.84	5.13
Standard error CpG per cent	0.03	0.01
Average ratio	0.72	0.7
Standard error ratio	0.00	0.00
Minimum island length	500.00	500.00
Maximum island length	3832.00	5639.00

Table 11: Final parameters of CpG island in Y-chromosome.

Features	*Bos taurus*	*Sus scrofa*
Average island length	545.86	567.28
Standard error (Island length)	2.92	4.55
Island number	1994.00	1820.00
Average G+C per cent	53.36	53.86
Standard error G+C per cent	0.1	0.11
Average CpG per cent	4.93	4.96
Standard error CpG per cent	0.02	0.02
Average ratio	0.7	0.7
Standard error ratio	0.00	0.00
Minimum island length	500.00	500.00
Maximum island length	2360.00	2650.00

and do not permit the targeting of markers from particular chromosomes, or even the identification of the chromosomes of origin of known markers. Consequently, the availability of DNA sequences is now providing unprecedented opportunities to identify novel genetic markers for use.

CONCLUSION

In the present study, 12 different animal species were organized into five groups and targeted for microsatellite and CpG mining in sex chromosomes. Microsatellite data have been analyzed by considering the simple and complex repeats. Simple repeats comprise of six classes of repeats including mono-, di-, tri-, tetra-, penta- and hexamers. The density of each class of repeat is comparable across various genomic regions. However, there is often tremendous variation in density in different genomic regions among different SSR types, sometimes even in a chromosome-specific manner. Based on X-chromosomes analysis *Mus musculus* of primates group contains highest number of microsatellites *i.e.* 79146 while *Meleagris gallopavo* of the avian group had the least number (*i.e.* 58) of microsatellites. Complex microsatellites also followed same pattern of occurrence and were highest in the primates group and least in avian growing-type type SSRs were reported highest in *Bos taurus* of the even-toed ungulates group and lowest in *Meleagris gallopavo* of avian group. Based on Y chromosomes analysis *Mus musculus* of primates group scored highest with

total of 49725 microsatellites. *Anopheles gambie* of insect group contained the least microsatellites with total of 4 numbers. *Gallus gallus* of avian group contained highest and *Drosophila melanogastor* of insect group contained lowest mono type microsatellites respectively. Complex type SSRs were reported highest in *Mus musculus* rodent group and lowest in *Anopheles gambie i.e*. 0.

Mining of CpG island in female dominant chromosomes revealed the highest numbers of 50388 in *Anopheles gambie* of the insect group and the least in *Meleagris gallopavo* of avian group with 83 CpG islands. Based on male dominant chromosome analysis (*i.e.* Y chromosome) CpG islands were found highest in *gallus gallus* of the avian group *i.e*. 4635 and least in *Anopheles gambie* of the insect group *i.e*. 3 respectively. It was concluded from this study that female dominant chromosome (*i.e.* X chromosome) contained highest number of both microsatellites and CpG islands as compared to male dominant Y chromosomes. It could be hypothesized that the female sex could be more prone to mutations and involved in evolution more importantly than males. Mutation rate could depend upon species type, age, sex of the individual, type of chromosome and type of allele loci. The knowledge obtained from this study can be used to understand various aspects and functions of genome organization, for marker-assisted selection in breed improvement, characterization, conservation and DNA fingerprinting. This analysis left a few questions, for example, why some repeats are in huge numbers and others extremely rare? What is the structural and functional basis for specific SSRs' chromosome-specific differential abundance? To understand the genome-wide gene structural and functional studies other kinds of DNA sequences and repeats will be needed to be analyzed and evaluated.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support provided by the Department of Biotechnology, Government of India, through the collaborative research project "Parentage Determination and Cytogenetic Profiling in Dogs" [DBT-19I].

Conflict of interest

All authors declared that there is no conflict of interest.

REFERENCES

Borstnik, B. and Pumpernik, D. (2002). Tandem repeats in protein coding regions of primate genes. Genome Resarch. 12: 909-915.

Duncan, C.G., Grimm, S.A., Morgan, D.L., Bushel, P.R., Bennett, B.D., Roberts, J.D., Wade, P.A. (2018). Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver. Scientific Reports. 8(1): 1-17.

Hakki, E.E. and Akkaya, M.S. (2000). Microsatellite isolation using amplified fragment length polymorphism markers: No cloning, no screening. Molecular Ecology. 9: 2152-2154.

Karagyozov, L., Kalcheva, I.D., Chapman, V.M. (1993). Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Research. 21: 3911-3912.

Kunzler, P., Matsuo, K., Schaffner, W. (1995). Pathological, physiological and evolutionary aspects of short unstable DNA repeats in the human genome. Biol. Chem. Hoppe Seyler. 4: 201-211.

Muyle, A., Bachtrog, D., Marais, G.A., Turner, J.M. (2021). Epigenetics drive the evolution of sex chromosomes in animals and plants. Philosophical Transactions of the Royal Society B. 376(1826): 20200124. doi: 10.1098/rstb.2020.0124.

Priolli, R.H.G., Mendes-Junior, C.T., Arantes, N.E., Contel, E.P.B. (2002). Characterization of Brazilian soybean cultivars using microsatellite markers. Genetics and Molecular Biology. 25: 185-193.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Bird, A. (2010). CpG islands influence chromatin structure *via* the CpG-binding protein Cfp1. Nature. 464 (7291): 1082-1086.

Yu, K., Park, S.J., Poysa, V., Gepts, P. (2000). Integration of simple sequence repeat (SSR) markers into a molecular linkage map of common bean (*Phaseolus vulgaris* L.). Journal of Heredity. 91(6): 429-434.