First report of SSR identification in the new improved genome assembly of Duck (ZJU1.0)

Jayakumar Sivalingam[1]*, Chaudhari Mahesh Vishwas[2], Athe Rajendra Prasad[3], Sajeed Mohd[4], Bachamolla Shivani[5], Satya Pal Yadav[6], T.K. Bhattacharya[6] and M. Balakrishnan[7]

[1]Senior Scientist, ICAR-Directorate of Poultry Research, Hyderabad, India,

[2]Asssistant Professor, College of Veterinary Science, Guru Angad Dev Veterinary and Animal Sciences University (GADVASU), Bathinda, Punjab, India

[3]Ph. D scholar, ICAR-Directorate of Poultry Research, Hyderabad, India,

[4,5]PG scholar, ICAR-Directorate of Poultry Research, Hyderabad, India,

[6,7]Principal Scientist, ICAR-Directorate of Poultry Research, Hyderabad, India,

[8]Principal Scientist, ICAR-National Academy of Agricultural Research Management, Hyderabad, India

ABSTRACT

Ducks (Anas platyrhynchos) were among the earliest domesticated fowls in the world. The Illumina-based genome assembly of the duck, BGI1.0 reference (GCA_000355885.1), was produced by Huang et al (2013), but recently Li etal (2021) produced a new duck genome assembly (ZJU1.0) with overall improvement compared to BGI1.0 duck genome, the previous Sanger-based zebra finch, and is comparable to the latest version of chicken and VGP zebra finch genomes. ZJUI1.0 duck genome has 62-fold improvement of scaffold continuity and assembled majorities of micro chromosomes that were all unmapped in the BGI1.0 genome. In the present study, the new Duck genome assembly (ZJU1.0)was downloaded from the NCBI and was used for identification of SSRs using MISA software. A total number of 7,03,449 SSRs were identified in the Mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats. Without taking into consideration of the mononucleotide repeats, our catalogue provides the first duck SSR reference based on ZJU1.0 genome including 183,406SSR loci with motifs of 2-6 bp at ~163 SSRs/Mb density. Without considering mononucleotide repeats, we observed that dinucleotide repeats were most abundant. The correlation between repeat motifs and SSR frequency was negative in our study in accordance with trend observed by Fan et al., 2018 except tetranucleotide repeats.

Key words: Duck genome, SSR, Diversity

*Corresponding author: jeyvet@gmail.com

INTRODUCTION

Ducks (*Anas platyrhynchos*) were among the earliest domesticated fowls in the world (CASS 1979). A haploid genome size of duck was estimated to be 1.41 Gb (Nakamura et al., 1990; Tiersch &Wchtel, 1966). A karyotype consists of 40 pairs of chromosomes out of which 9 pairs are of macrochromosomes (chr1–chr8, chrZ/chrW) and 31 pairs of micro chromosomes (chr9 – chr39) (Takagi & Makino, 1966). The Illumina-based genome assembly of the duck, BGI 1.0 reference (GCA_000355885.1), was produced by Huang et al., 2013. Fan et al., 2018referred to this genome assembly to provide the first genome-wide duck SSR reference. Short sequence repeats (SSRs)(Tautz et al., 1986) or microsatellites or short tandem repeats (STR) or simple sequence length polymorphisms (SSLPs) (McDonald & Pott., 1997) are simple sequence stretches with tandemly repeated motifs of 1-6 bp that occur both in coding and non-coding regions(Gupta et al., 1996;Toth et al., 2000; Katti et al., 2001) of both prokaryotic and eukaryotic genomes. SSRs, by virtue of their co-dominant and multiallelic nature, prove to be efficient in genetic diversity studies, pedigree evaluation, genetic mapping and marker-assisted selection(Edwards et al.,1991; Edwards et al., 2000;Dorji et al .,2003;Ashwell et al., 2004;Williams, 2005;Rishichkowsky & Pilling,2007; Nguyen et al.,2007;Sun et al 2007;Mao et al., 2008; Zhang et al., 2007; Deng et al.,2016) . The Illumina-based Duck genome assembly BGI 1.0 has 25.9% of the assembled genome assigned to chromosomes, containing 3.17% of bases as gaps. Fan et. al. (2018) referred to this genome assembly to provide the first genome-wide duck STR reference including 198,022 STR loci with motif size of 2–6 base pairs distributed unevenly in the duck genome indicating a directional

selection pressure.

But recently Li et al (2021) produced a new duck genome assembly (ZJU1.0), (accession No. JACGAL000000000 and accession No. JABVCD000000000), with overall improvement compared to BGI1.0 duck genome, the previous Sanger-based zebra finch, and comparable to the latest version of chicken (Warren et al.,2017)and VGP zebra finch genomes (Rhie et al.,2021).

There is no report available on genome-wide duck SSR based upon the new Duck genome assembly (ZJU1.0) and therefore the objective of the study was to identify SSRs in the new improved new Duck genome assembly (ZJU1.0).

MATERIALS AND METHODS

Duck genome assembly (ZJU1.0), (accession No. JACGAL000000000 and accession No. JABVCD000000000) was downloaded from the NCBI and was used for identification of SSRs using MISA software (Beier et al., 2017). Then, the Perl script MISA tool was used to screen the polymorphic microsatellites with different thresholds level (1-10 2-6 3-5 4-5 5-5 6-5 i.e. Mononucleotide repeats with minimum 10 repeats, dinucleotide repeats with minimum 6 repeats, Trinucleotide repeats with minimum 5 repeats, Tetra nucleotide repeats with minimum 5 repeats, Pentanucleotide repeats with minimum 5 repeats, Hexa nucleotide repeats with minimum 5 repeats). The minimum distance between 2 compound SSRs was set at 100 bp in length.

RESULTS AND DISCUSSION

Duck genome, ZJU1.0, of size 1.189 Gb was downloaded successfully. As shown in Table 1 SSR loci were mapped on macrochromosomes (chr1–chr8, chrZ/chrW) and micro chromosomes (chr9 – chr33). A total of 703449number of SSR loci of Mono-, di-, tri-, tetra-, penta-, and hexa nucleotide repeat motifs were identified. Mono-nucleotide repeat motifs were highest in number (520043) followed by di-, tri-, tetra-, penta-, and hexanucleotide repeat motifs (Table 2). Compared with total size of ZJU1.0 genome examined density of 624.45 SSRs/Mb whereas without taking of mononucleotide repeats into account an average density of 162.81 SSRs/Mb was observed (Table 2 & 3).

Two genome assemblies of duck, BGI1.0 (Huang et al., 2013) and ZJU1.0 (Li et al., 2021), were produced within the span of past 8 years. The first duck SSR reference was produced by Fan and co-workers in the year 2018 based on BGI1.0 genome. But ZJU1.0 genome has improvements compared to BGI1.0. ZJUI1.0 duck genome has 62-fold improvement of scaffold continuity and assembled majorities of micro chromosomes that were all unmapped in the BGI1.0 genome. There is 7.1% increase in size of ZJUI1.0 and has a higher level of completeness measured by its almost gapless sequence composition (0.37% vs 3.17%) found to be enriched for repetitive elements and GC-rich sequences. Overall, ZJI1.0 genome predicted 15,463 protein-coding genes, including 71 newly annotated chicken W chromosome genes, identified 8,238 missing exons in the BGI1.0 assembly in 2,099 genes, including 745 genes that were completely missing and corrected 683 partial genes and merged them into 356 genes in the new assembly. Therefore, we produced improved SSR reference for duck based on ZJU1.0 genome.

Without taking mononucleotide repeats into account, our catalogue provides the first duck SSR reference based on ZJU1.0 genome including 183,406 SSR loci with motifs of 2-6 bp at ~163 SSRs/Mb density. The number and density of SSR identified in our study was lower not only than that earlier reported in duck (198022) (Fan et al., 2018) but also than in human (700000)Willems et al., 2014and porcine (600000)(Liu et al., 2017). It is in accordance with confirmation, by Fan et al., 2018, that SSR abundance is lower in avian than in mammals (Primmer et al., 1997; Brandstrom & Ellegren, 2008). Mononucleotide repeats outnumbered other SSR categories in our catalogue, which agreed with previous reports in Gallus gallus, Meleagris gallopavo, Taeniopygiaguttata, Geospiza fortis, Melopsittacus undulates, and Columba livia(Huang et al., 2016). We observed that dinucleotide repeats were most abundant. We observed the trend of negative correlation between repeat motifs and SSR frequency in accordance with trend observed by Fan et al., 2018 except tetranucleotide repeats.

Table 1: Chromosome wise SSR identification in the new improved genome assembly of Duck (ZJU1.0)

Chromosome no.	Total size of examined sequences (bp) SSRs	Total number of identified SSRs	Number of SSRs present in compound formation	Mono nucleotide SSRs	Di nucleotide SSRs	Tri nucleotide SSRs	Tetra nucleotide SSRs	Penta nucleotide SSRs	Hexa nucleotide SSRs
Chromosome 1	207238429	145630	27560	102422	16184	9200	10213	6329	1282

Chromosome no.	Total size of examined sequences (bp) SSRs	Total number of identified SSRs	Number of SSRs present in compound formation	Mono nucleotide SSRs	Di nucleotide SSRs	Tri nucleotide SSRs	Tetra nucleotide SSRs	Penta nucleotide SSRs	Hexa nucleotide SSRs
Chromosome 2	164862000	114871	21434	82140	12450	7112	7413	4864	892
Chromosome 3	120086012	81129	14887	58968	8608	4768	4933	3337	515
Chromosome 4	76269206	49592	7937	37809	5331	2621	2324	1269	238
Chromosome 5	66856311	38172	5958	29482	3893	2051	1625	974	147
Chromosome 6	37939480	21134	3158	16873	2132	1025	732	314	58
Chromosome 7	39896445	22880	3350	18081	2393	1123	815	399	69
Chromosome 8	32618258	18051	2684	14223	1821	912	672	319	104
Chromosome 9	26788864	14027	1991	11169	1308	755	518	229	48
Chromosome 10	22484652	11919	1714	9330	1173	674	479	220	43
Chromosome 11	22598276	11758	1719	9490	1024	629	386	196	33
Chromosome 12	21610006	12326	1847	9912	1114	628	383	224	65
Chromosome 13	23150841	12103	1741	9811	1145	640	311	164	32
Chromosome 14	20984523	10683	1813	8279	919	654	341	233	257
Chromosome 15	17872457	8707	1273	7075	784	437	237	140	34
Chromosome 16	16222536	7974	1127	6143	686	450	332	124	239
Chromosome 17	1402342	733	185	520	49	65	44	47	8
Chromosome 18	12065260	5353	768	4326	434	348	129	94	22
Chromosome 19	13129675	6273	975	5064	455	413	172	133	36
Chromosome 20	11966879	5215	730	4273	388	325	138	63	28
Chromosome 21	16960004	8584	1324	6917	717	472	249	192	37
Chromosome 22	8397549	4103	651	3404	302	251	84	44	18
Chromosome 23	5328910	2847	475	2372	183	168	61	51	12
Chromosome 24	7636233	3783	696	2979	265	342	86	93	18
Chromosome 25	7658487	3508	506	2875	283	228	63	42	17
Chromosome 26	3037671	2178	595	1739	91	193	41	39	75
Chromosome 27	6799408	3270	520	2650	210	294	55	45	16
Chromosome 28	7028699	3817	879	2954	286	273	131	146	27
Chromosome 29	6031659	3730	860	3182	187	228	52	63	18
Chromosome 31	214689	226	71	194	5	20	1	6	-
Chromosome 33	112901	252	73	224	3	19	2	1	3
Chromosome W	16693329	5173	659	3580	420	753	215	146	59
Chromosome Z	84547829	63447	14272	41582	7180	4442	5449	4021	773
Mt	16604	1	0	1	-	-	-	-	-

bp – base pair
SSR -Short sequence repeats
mt- Mitochondria

Table 2: Comparison of the BGI1.0 and ZJU1.0 genomes

		Duck Genome	
		BGI1.0*	**ZJU1.0**
1	Size (Gb)	1.105	1.189
2	Total number of SSR loci identified	1,98,022	7,03,449
3	Repeat motifs	di-, tri-, tetra-, penta-, and hexa nucleotide	Mono-, di-, tri-, tetra-, penta-, and hexa nucleotide
4	An average density (SSRs/Mb)	182.0	162.81**
5	Correlation between SSR frequency and number of nucleotides	Negative exception tetra-nucleotide	Negative

*SSR data regarding BGI1.0 from Fan *et al* (2018)
**Without taking of mononucleotide repeats into account

Table 3: Comparison of the number of SSRs, its frequency and density in the BGI1.0 and ZJU1.0 genomes

		Number of SSR		Frequency of SSR		Density of SSR (No./Mb)	
Reference Genome		**BGI1.0**	**ZJU1.0**	**BGI1.0**	**ZJU1.0**	**BGI1.0**	**ZJU1.0***
Repeat motifs (nucleotide)	Mono-	-	520043	-	73.93	-	461.64
	di-	54,347	72423	27.44	10.30	49.95	64.29
	tri-	31,711	42513	16.01	6.04	29.15	37.74
	tetra-	76,100	38686	38.43	5.50	69.95	34.34
	penta-	25,750	24561	13.00	3.49	23.67	21.80
	Hexa-	10,114	5223	5.11	0.74	9.30	4.64
Total		198,022	703,449	100.00	100.00	182.02	624.45

*Compared with total size examined

**SSR data regarding BGI1.0 from Fan*et al*

ACKNOWLEDGEMENTS

The authors thank the Directors of ICAR-Directorate of Poultry Research, Hyderabad, India and ICAR-National Academy of Agricultural Research Management, Hyderabad, India for providing the facilities to analyse the genome sequence data.

FINANCIAL SUPPORT

For the present study, no grant was obtained from any of the funding agency.

ETHICAL STATEMENT

Data already available in the public domain GRCg7b (https://www.ncbi.nlm.nih.gov/genome/111?genome_assembly_id=1543395) and GRCg6a (http://asia.ensembl.org/Gallus_gallus/Info/Annotation) were used for the present study and hence no ethical committee approval was obtained.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Ashwell, M. S., Heyen, D. W., Sonstegard, T. S., Van Tassell, C. P., Da, Y., VanRaden, P. M., ... & Lewin, H. A. 2004. Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. *Journal of dairy science*, 87(2): 468-475.

Brandström, M., &Ellegren, H. 2008. Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome research*, 18(6): 881-887.

CASS (Chinese Academy of Social Sciences) 1979. Archaeological perspectives on Yin ruins, Anyang excavations in 1969–1977, China. Acta Archaeological Sinica, Institute of Archaeology.

Deng, P., Wang, M., Feng, K., Cui, L., Tong, W., Song, W., &Nie, X. 2016. Genome-wide characterization of microsatellites in Triticeae species: abundance, distribution and evolution. *Scientific reports*, 6(1): 1-13.

Dorji, T., Hanotte, O., Arbenz, M., Rege, J. E., &Roder, W. 2003. Genetic diversity of indigenous cattle populations in Bhutan: Implications for conservation. Asian-australasian *journal of animal sciences*, 16(7): 946-951.

Edwards, A., Civitello, A., Hammond, H. A., & Caskey, C. T. 1991. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American journal of human genetics*, 49(4): 746.

Edwards, C. J., Gaillard, C., Bradley, D. G., &MacHugh, D. E. 2000. Y-specific microsatellite polymorphisms in a range of bovid species. *Animal Genetics*, 31(2) 127-130.

Fan, W., Xu, L., Cheng, H., Li, M., Liu, H., Jiang, Y., & Hou, S. 2018. Characterization of duck (Anas platyrhynchos) short tandem repeat variation by population-scale genome resequencing. *Frontiers in genetics*, 9: 520.

Gupta, P. K., Balyan, H. S., Sharma, P. C., & Ramesh, B. 1996. Microsatellites in plants: a new class of molecular markers. *Current science*, 45-54.

Huang, J., Li, W., Jian, Z., Yue, B., & Yan, Y. 2016. Genome-wide distribution and organization of microsatellites in six species of birds. *Biochemical Systematics and Ecology*, 67: 95-102.

Huang, Y., Li, Y., Burt, D. W., Chen, H., Zhang, Y., Qian,

W., ... & Li, N. 2013. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nature genetics*, 45(7): 776-783.

Katti, M. V., Ranjekar, P. K., & Gupta, V. S. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular biology and evolution*, 18(7): 1161-1167.

Li, J., Zhang, J., Liu, J., Zhou, Y., Cai, C., Xu, L., ... & Zhou, Q. 2021. A new duck genome reveals conserved and convergently evolved chromosome architectures of birds and mammals. *GigaScience*, 10(1): 142.

Liu, C., Liu, Y., Zhang, X., Xu, X., & Zhao, S. 2017. Characterization of porcine simple sequence repeat variation on a population scale with genome resequencing data. *Scientific reports*, 7(1): 1-10.

Mao, Y., Chang, H., Yang, Z., Zhang, L., Xu, M., Chang, G., ... & Ji, D. 2008. The analysis of genetic diversity and differentiation of six Chinese cattle populations using microsatellite markers. *Journal of Genetics and Genomics*, 35(1): 25-32.

McDonald, D. B., & Potts, W. K. 1997. DNA microsatellites as genetic markers at several scales. Avian molecular evolution and systematics. Academic Press, San Diego, 29-49.

Nakamura, D., Tiersch, T. R., Douglass, M., & Chandler, R. W. 1990. Rapid identification of sex in birds by flow cytometry. *Cytogenetic and Genome Research*, 53(4): 201-205.

Nguyen, T. T., Genini, S., Bui, L. C., Voegeli, P., Stranzinger, G., Renard, J. P., ... & Nguyen, B. X. 2007. Genomic conservation of cattle microsatellite loci in wild gaur (Bos gaurus) and current genetic status of this species in Vietnam. *BMC genetics*, 8(1): 1-8.

Primmer, C. R., Raudsepp, T., Chowdhary, B. P., Møller, A. P., &Ellegren, H. 1997. Low frequency of microsatellites in the avian genome. *Genome Research*, 7(5): 471-482.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., ... & Jarvis, E. D. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856): 737-746.

Rischkowsky, B., & Pilling, D. 2007. The state of the world's animal genetic resources for food and agriculture. Food & Agriculture Org.

Sebastian Beier, Thomas Thiel, Thomas Münch, Uwe Scholz, & Martin Mascher. 2017. MISA-web: a web server for microsatellite prediction, *Bioinformatics* 33(16): 2583–2585.

Sun, W., Chen, H., Lei, C., Lei, X., & Zhang, Y. 2007. Study on population genetic characteristics of Qinchuan cows using microsatellite markers. *Journal of Genetics and Genomics*, 34(1): 17-25.

Takagi, N., & Makino, S. 1966. A revised study on the chromosomes of three species of birds. *Caryologia*, 19(4): 443-455.

Tautz, D., Trick, M., & Dover, G. A. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, 322(6080): 652-656.

Tiersch, T. R., & Wachtel, S. S. 1991. On the evolution of genome size of birds. *Journal of Heredity*, 82(5): 363-368.

Tóth, G., Gáspári, Z., & Jurka, J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome research*, 10(7): 967-981.

Warren, W. C., Hillier, L. W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., ... & Cheng, H. H. 2017. A new chicken genome assembly provides insight into avian genome structure. G3: *Genes, Genomes, Genetics*, 7(1): 109-117.

Williams, J. L. 2005. The use of marker-assisted selection in animal breeding and biotechnology. *Revue Scientifique et Technique-Office International des Epizooties*, 24(1): 379.

Willems, T., Gymrek, M., &Highnam, G. 2014. 1000 Genomes Project Consortium Mittelman D., Erlich Y. The landscape of human STR variation. *Genome Res*, 24: 1894-1904.

Zhang, J., Xiong, Y., Zuo, B., Lei, M., Jiang, S., Zheng, R., & Li, J. 2007. Genetic analysis and linkage mapping in a resource pig population using microsatellite markers. *Journal of Genetics and Genomics*, 34(1): 10-16.