Int.J.Curr.Microbiol.App.Sci (2017) *6*(11): 2244-2252

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 6 Number 11 (2017) pp. 2244-2252**
Journal homepage: http://www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2017.611.266

In Silico Characterization of Selected Microsatellite Loci Reported in Zebrafish Genome

Mujahidkhan A. Pathan*, E.A. Nesnas, Aditya Pratap Acharya, Rameez Roshan, Thushar P. Kumar, P. Gireesh-Babu, Shrinivas Jahageerdar, Aparna Chaudhari and Gopal Krishna

Fish Genetics and Biotechnology Division, ICAR-Central Institute of Fisheries Education, Mumbai-400061, India
**Corresponding author*

Keywords

Danio rerio, SSR, Genome distribution, Repeats, Genome coordinates.

Article Info

Accepted:
17 September 2017
Available Online:
10 November 2017

A B S T R A C T

Microsatellite loci of zebrafish are an important genetic resource owing to widespread use of microsatellites for a number of applications and the vertebrate model status of zebrafish. Data available in the public domain was mined for microsatellites present in and near transcribed regions. Two thousand loci reported earlier were individually traced to their genome coordinates in order to record type and structure of repeats, location with respect to exonic, intronic and noncoding regions and distribution of each type across linkage groups. Location of 1781 loci could be ascertained using Ensembl BLAT against the zebrafish (GRCz10) genome database or Zv9 (Ensembl release 79), of which a total of 981 loci (55%) were located in the genic regions and 800 (45%) were mapped in the noncoding regions of the zebrafish genome. Within the genes, 700 (39%), 36 (2%) and 245 (14%) were in introns, exons and exon-intron junctions, respectively. None of the microsatellites present in the noncoding region were within +/- 300 bp of the open reading frames, implying lack of STR loci in the 5' and 3' regulatory regions in the available data. Analysis with online Microsatellite Repeat Finder tool revealed 1597 loci with well-defined repeat structures, of which 1524, 19, and 54 were di, tri and tetra repeats, respectively. The genes spanning the loci were recorded and more than ninety percent had biological roles assigned. New PCR primers were designed for selected loci and robust primers are reported for 42 tri, tetra, non-coding and exonic loci. The information generated in this work will facilitate further investigations related to role elucidation of STRs and enable other applications like genotyping of zebrafish strains and cross-species amplification of loci.

Introduction

The zebrafish, *Danio rerio* (Teleostei infraclass and Cyprinidae family) is a monophyletic group that is thought to have arisen approximately 340 million years ago from a common ancestor (Amores *et al.*, 2013). It has emerged as a popular vertebrate model organism for various biological studies due to excellent traits like external fertilisation, fast development, optical clarity during embryogenesis, high fecundity, short generation time and ease of maintenance (Kimbery and Leonard 2000, Zhoa *et al.*, 2015). Zebrafish is widely used for modelling of human diseases because of high genetic

homology, physiology, and developmental similarity with humans (Bakkers 2011; Liu and Stainier 2012).

The whole genome sequence of zebrafish has been mined for various kinds of information and the recorded overall repeat content of 52.2%, is the highest reported so far in any vertebrate (Howe *et al.,* 2013). Microsatellites or simple sequence/ tandem repeats (SSRs/ STRs) are short tandemly arrayed di-, tri-, or tetranucleotide repeat sequences with repeat sizes of 1–6 bp (Tautz, 1989) that are scattered across prokaryotic and eukaryotic genomes.

The distribution and density of microsatellites within chromosomes varies from species to species, but the size of the repeat unit is largely inversely related to its genomic frequency (Koreth 1996; Oliveira *et al.,* 2006). Microsatellite loci can be perfect, compound and imperfect based on the repeat structure (Urquhart 1994). Due to their abundance, codominance, small size, high variability and even distribution across the genome, microsatellite loci are widely used as DNA markers for various genetic studies (Muneer *et al.,* 2009). The majority of STRs are found in the noncoding regions and only about 8% are located in coding regions (Ellegren 2000).

Shimoda *et al.,* (1999) reported a microsatellite map of zebrafish including 2000 microsatellites spaced about 1.2 cM apart (2295 cM/ 2000 markers). An average of 79 loci per linkage group (LG) was reported with a maximum of 97 on LG 7 and a minimum of 50 on LG 5. ZFIN (Zebrafish Information Network) provides some details of these loci like the length, linkage group on which they are located, primer pairs for PCR amplification, and hyperlinks for NCBI-BLAST, Ensembl, etc. Usually, a researcher requires more detailed information about microsatellite loci, including the repeat type and structure, exact coordinates and location with respect to genes.

In several cases, researchers may be interested in microsatellite loci present specifically in the exonic, intronic, intron-exon boundary or regulatory regions. This study was conducted to mine such data, analyse it and make it easily available to future researchers, keeping in mind the importance of zebrafish model for genetic research.

Materials and Methods

Data mining for details of microsatellite loci

Accession IDs of the two thousand microsatellites reported earlier (Shimoda *et al.,* 1999) were obtained from ZFIN (https://zfin.org/). The full length microsatellite sequences with 1000 bp flanking regions at both ends were retrieved in FASTA format from the GenBank, NCBI (https://www.ncbi.nlm.nih.gov/). Each of these sequences was subjected to Ensembl BLAT (BLAST Like Alignment Tool) analysis individually against the zebrafish (GRCz10) genome database or Zv9 (Ensembl release 79) at Ensembl genome database (https://www.ensembl.org/Multi /Tools/Blast?db=core). The top hit hyperlink on the results page was followed to obtain the coordinates of each microsatellite, which were used to determine whether they were located in the non-coding regions, intronic regions, exonic regions or on exon-intron boundaries. Information was also recorded about genes spanning the locus or those in the vicinity along with gene orientation, its length, product and metabolic role, if any. The structure of repeat region was determined by Microsatellite Repeat Finder tool (http://insilico.ehu.es/mini_tools/microsatellites/).

Designing of PCR primers

New PCR primers with high stringency were designed for 50 selected microsatellite loci using Gene Runner v 3.0. The length of the primers was 22-26 bases, Tm was 45°C to 65°C and GC content was 40 to 60%. Primers were searched for the absence of dimerization, hairpin formation and secondary priming sites to avoid mispairing. Tm difference between primer pairs was < 5°C. Polynucleotide stretch was avoided.

PCR amplification for testing primers

PCR amplification was carried out with the primers designed above using zebrafish genomic DNA as the template. DNA was extracted by the standard Phenol-chloroform method and PCR was performed in 25 µl reaction volume containing 50 ng template DNA, 10 pmol of each specific primer, 200 µM of each dNTPs, 0.75 units of Taq DNA polymerase and 1× Taq buffer containing 1.5 mM MgCl2 (Sambrook *et al.,* 2001). The amplification reaction was carried out in 0.2 ml PCR tubes in a heated lid thermocycler. The PCR conditions included initial denaturation at 95° C for 5 min followed by 35 cycles of denaturation at 94°C for 30 s, annealing for 20 s using touch down conditions set at 65°C to 60°C, extension at 72° C for 30 s and final extension at 72°C for 8 min. The amplified products were visualised on 1.5 % agarose gel (Sambrook *et al.,* 2001).

Results and Discussion

Analysis of microsatellite loci

Out of the 2000 microsatellites reported earlier (Shimoda *et al.,* 1999), repeat regions and genomic locations could be identified for only 1781 loci. A total of 981 loci (55%) were located in the genic regions while 800 (45%) were mapped in the noncoding regions of the zebrafish genome. Within the genes, 700 (39%), 36 (2%) and 245 (14%) were in introns, exons and exon-intron junctions, respectively (Fig. 1). More than ninety percent of the genes harbouring microsatellite loci had biological roles assigned. None of the microsatellites present in the noncoding region were within +/- 300 bp of the open reading frames, implying lack of STR loci in the 5' and 3' regulatory regions in the available data. Figure 2 depicts the distribution of these 1781 loci across non-coding/ exonic/ intronic/ boundary regions and chromosomes. LGs 8, 9, 14, 19, 20, 21, 22, 23, 24, and 25 were found to have more microsatellite loci located in the genic regions compared to non-coding regions. LG 5 had the maximum number of 6 microsatellite loci in the exonic region, while no exonic microsatellite loci could be identified in the LGs 1, 4, 6, 17, 18 and 25. Details of individual microsatellite loci in terms of linkage group, locus number, microsatellite accession ID, co-ordinates, repeat sequence length are provided in supplementary sheet 1.

Nucleotide sequences of the loci, information about genes located on or near the loci, GenBank Accession IDs, genome coordinates, orientation of gene, gene product, gene function for tri and tetra repeats and for exonic microsatellites are provided in supplementary tables 1 – 3.

After analysis with Microsatellite Repeat Finder tool, 1597 loci with well-defined repeat structures were selected out of which 1524, 19, and 54 were di, tri and tetra repeats, respectively (Fig. 3). The distribution of repeat types across chromosomes is depicted in Figure 4. Six linkage groups, namely 4, 13, 14, 15, 19 and 25 had exclusively dinucleotide repeats. As observed here, tetramers have been reported to be more abundant than trimers in vertebrate genomes (with the exception of tetraodon), and

Rouchka (2010) has reported them to be twice as many. LG 7 has the highest number of tri and tetra repeat microsatellite loci (Supplementary table 1 – 3).

Most of the microsatellite loci in the exonic region were dinucleotide repeats with AC/CA being the most frequent repeat followed by GT/TG in accordance with earlier reports. Among vertebrates, GT/TG and AC/CA are believed to be the most abundant dinucleotide microsatellites, the latter being more abundant in fish.

Also the tri-, tetra, penta and hexanucleotide repeats, CAT, ATCT, TTATC and CACACT have maximum occurrences in fish genomes (Nagpure *et al.*, 2013). Exons were found to harbour both simple and compound repeats and the locus z22084 present in melanocyte proliferation gene contained the compound repeat $(GATGAA)_3$ $(TG)_8$. The functions of genes harbouring microsatellite repeats in exonic regions varied from Wnt signalling, signal transduction, ubiquitin pathway, immune related, DNA binding, replication, heart development, protein trafficking, DNA repair, growth and transcription.

Out of the 19 tri-nucleotide microsatellite loci only five were located in non-coding regions and the rest were found in the introns or intron – exon junctions. Surprisingly, no trinucleotide repeats were found exclusively in the exonic regions, although an earlier work by Subramanian *et al.*, (2003) found a twofold greater density of tri repeats in exonic regions compared to the non-coding regions in all human chromosomes except Y. In yeast as well as humans, many proteins involved in transcriptional regulation contain glutamine-rich domains and trinucleotide repeats encoding series of polyglutamine (Escher 2000). The repeat length of trimers ranged from 3 to 12 with z20576 locus being the longest $(ATC)_{12}$. The microsatellite loci z9944 and z6973 had compound repeats whereas all others were simple repeats. The tri repeats were found associated with genes related to differentiation, transcription, signal transduction, ubiquitination, splicing, immune related etc. The tri repeats observed in the present study were found to be AT rich. This is in agreement with the study by Subramanian *et al.*, (2003) who reported that the trimers in human genome are dominated by AT rich sequences.

Fig.1 Microsatellite loci abundance in various genomic regions of zebrafish

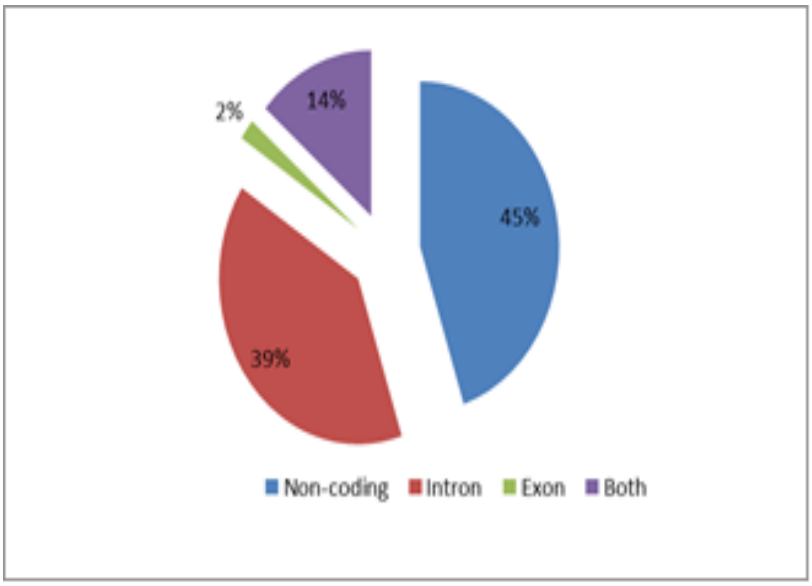


Fig.2 Location based distribution of microsatellites across linkage groups

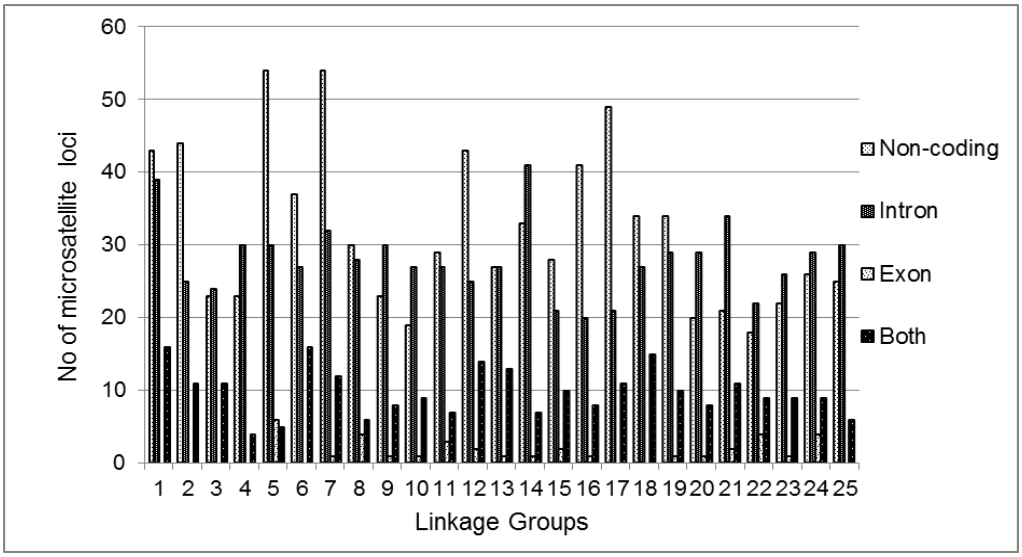


Fig.3 Abundance of microsatellite loci based on repeat unit length in zebrafish genome

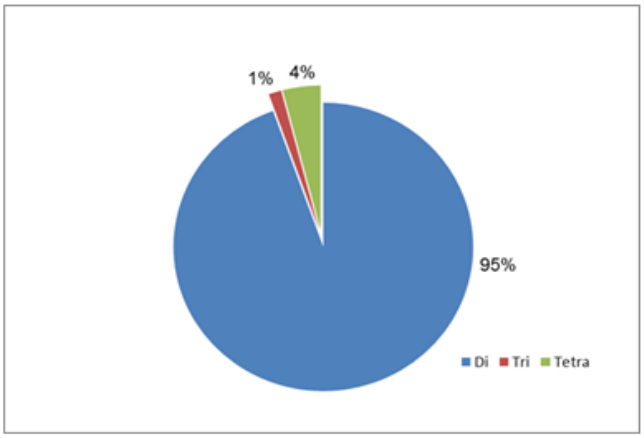


Fig.4 Distribution of microsatellite repeat types across linkage groups

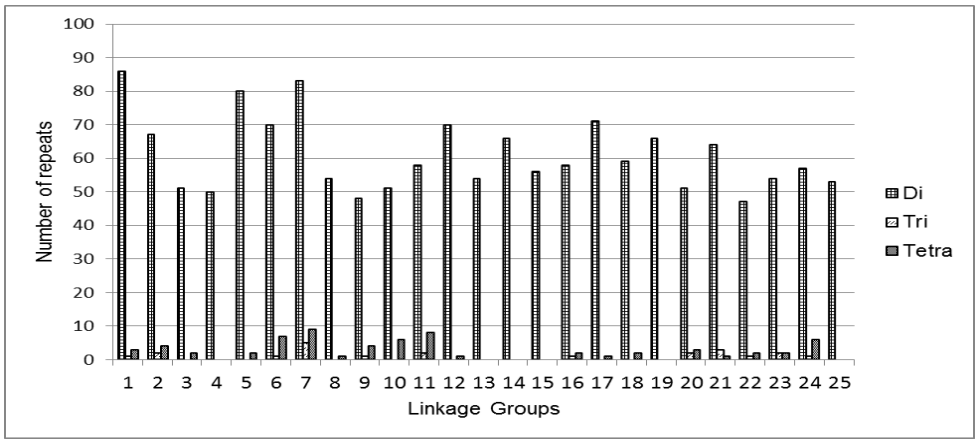
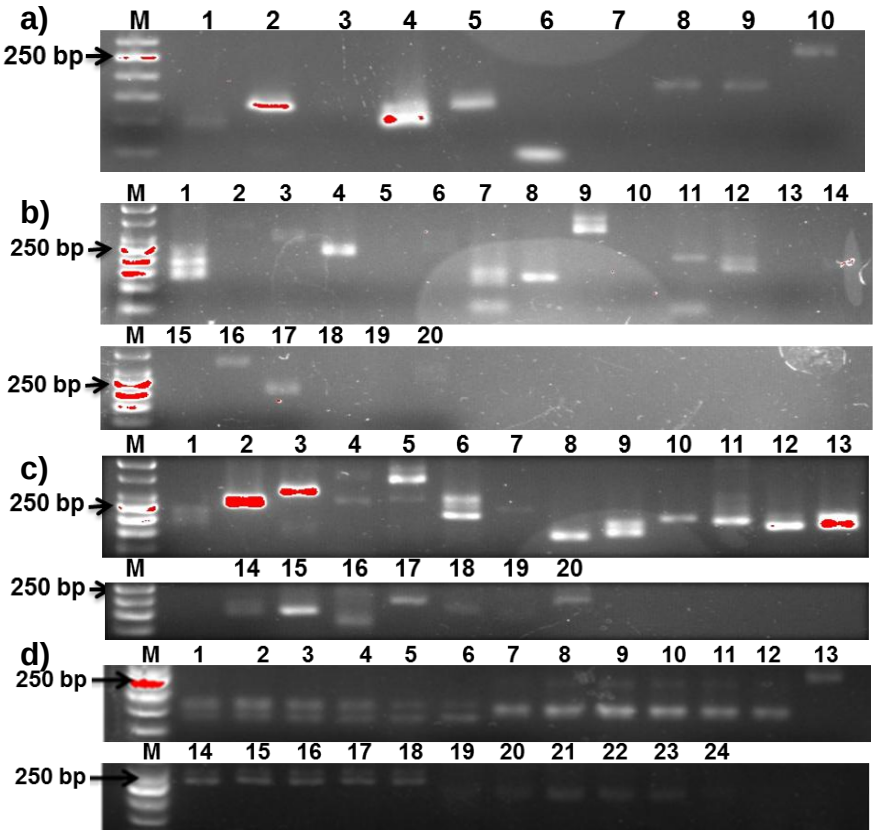


Fig.5 PCR amplification of microsatellite loci from zebrafish genomic DNA



Lane M: Generuler 50 bp DNA ladder (Thermoscientific, USA); a) lanes 1, 2, 4, 5, 8, 9 & 10: Z20576 (101bp), Z13685 (131bp), Z4329 (105bp), Z13253 (150bp), Z8602 (385bp), Z5022 (186bp), Z10838 (204bp) & Z9871 (235bp), lanes 3, 6, 7: Z7156, Z8602 & Z6850 did not show desired amplification product; b) lanes 1, 3, 4, 7, 8, 9, 11, 12, 16 & 17: Z7490 (180bp), Z9125 (262bp), Z7632 (194bp), Z7141(168bp), Z13685 (131bp), Z21115 (132bp), Z21123 (223bp), Z10324 (142bp), Z720 (324bp) & Z7925 (245bp), lanes 2, 5, 6, 10, 13, 14, 15, 18, 19 & 20: Microsatellite loci Z3602, Z8517, Z15457, Z10279, Z1984, Z3901, Z7125, Z131214, Z7235 & Z3558 did not show desired amplification product; c) lanes 1-20: Z6010 (169bp), Z9511 (267bp), Z11701 (317bp), Z9720 (241bp), Z1412 (309bp), Z4325 (212bp), Z7642 (250bp), Z7807 (149bp), Z8146 (144bp), Z4761 (207bp), Z6981 (180bp), Z10215 (186bp), Z1525 (214bp), Z22516 (309bp), Z11894 (151bp), Z4349 (159bp), Z10551 (115bp), Z20966 (168bp), Z8718 (176bp) & Z21679 (173bp); d) Gradient PCR from 55 to 65°C. Lanes 1-6: Z7156 (141bp), lanes 7-12: Z21123 (223bp), lanes 13-18: Z10324 (142bp), lanes 19-24: Z720 (324bp)

Out of 54 tetranucleotide microsatellite loci, 24 were located in the non-coding regions and the rest were found in intron, exon and intron – exon junction. No tetramers could be identified in the linkage groups 4, 13, 14, 15, 19 and 25. Only one tetramer, z21115 $[(CGTG)_{14}]$ was found in the exonic region of the *pcsk5b* (proprotein convertase subtilisin) gene. The repeat lengths ranged from 3 to 22 and the microsatellite locus z13223 showed the highest number of tetra repeats $(GTTT)_{22}$.

Seven microsatellite loci (z7824, z7303, z4375, z13244, z13678, z8456 and z8602) harboured compound repeats while rest were simple repeats. The tetra repeats were found in genes related to replication, cell cycle, signal transduction, immune related, metabolism and reproduction.

Microsatellite loci with pentamers and hexamers include z4375 $[(TTCAG_{)3}$ $(TTCAG)_3]$, z1182 $(TGCG)_7$, z9199

$(CTGAG)_3$ and z13244 $[(TTTAG)_3$ $(AGTTT)_8$ $(TAGTT)_4]$. The microsatellite locus z13678 had compound repeats comprising both tetramers and hexamers $[(TATC)_3$ $(TGAGGG)_3]$. Only one hexamer could be identified among all the loci tested. It has been reported earlier that zebrafish has a low frequency of hexamers in its genome (Rouchka, 2010).

PCR amplification of selected microsatellite

Out of the 50 primer sets tested, reproducible amplification showing bands of expected sizes was obtained for 42 loci (Fig. 5). The details of 42 microsatellite loci that could be successfully amplified from zebrafish genomic DNA are given supplementary tables 4-7.

STRs have come a long way since they were referred to as 'junk DNA' and are thought to have biological significance in the regulation of gene expression (Hamada 1984; Naylor 1990), recombination (Wahls 1990), generation of nucleosome positioning signals (Wang and Griffin 1995) and maintenance of chromatin spatial organization (Heale and Petes 1995). In many disease-causing bacteria, some "contingency genes" reside in STR sequences that allow frame-shift mutations for promoting adaptation and better ability to survive host defence (Fan and Chu 2007). A polymorphic dinucleotide repeat in intron 1 has been shown to modulate transcription of epidermal growth factor receptor gene *in vitro* (Gebhardt *et al.,* 1999). A tetranucleotide polymorphic microsatellite present in the first intron was found to regulate the transcription of the tyrosine hydroxylase gene *in vitro* (Meloni, *et al.*, 1998). Biological functions of STRs are still being explored and zebrafish is a good vertebrate model for these investigations. The zebrafish (Zv9) genome size is 1,412,464,843 and it possesses 26,206 protein-coding genes with the highest number of 3,23,599 exons among previously sequenced vertebrates (Collins *et al.,* 2012). It has high number of species-specific genes in its genome compared with human, mouse or chicken. Other sequenced teleost fish have an average repeat content of less than 30%, but zebrafish has 52% (Howe *et al.,* 2013). The information generated in this work will facilitate further investigations related to role elucidation of STRs and enable ease of application of microsatellites as DNA markers for genotyping of zebrafish strains and cross-species amplification of repeat loci.

This work provides researchers with detailed information about the microsatellite loci associated with transcribed regions. Robust primers have also been reported for 42 tri, tetra and exonic repeats that can be applied for genotyping of variant strains and cross amplification in other species, etc. Studies on elucidation of biological roles of repeats will also be facilitated.

Acknowledgements

The authors acknowledge Director/Vice Chancellor, ICAR-CIFE, Mumbai, India for providing all facilities and the Indian Council of Agricultural Research, New Delhi, India for financial support.

References

Amores, A., Catchen, J., Ferrara, A., Fontenot, Q. and Postlethwait, J.H., 2011. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, *188*(4), pp.799-808.

Bakkers, J., 2011. Zebrafish as a model to study cardiac development and human

cardiac disease. *Cardiovascular research*, *91*(2), pp.279-288.

Collins, J.E., White, S., Searle, S.M. and Stemple, D.L., 2012. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome research*, *22*(10), pp.2067-2078.

Escher, D., Bodmer-Glavas, M., Barberis, A. and Schaffner, W., 2000. Conservation of glutamine-rich transactivation function between yeast and humans. *Molecular and cellular biology*, *20*(8), pp.2774-2782.

Fan, H. and Chu, J.Y., 2007. A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics*, *5*(1), pp.7-14.

Gebhardt, F., Zänker, K.S. and Brandt, B., 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *Journal of Biological Chemistry*, *274*(19), pp.13176-13180.

Hamada, H.I.R.O.S.H.I., Seidman, M.I.C.H.A.E.L., Howard, B.H. and Gorman, C.M., 1984. Enhanced gene expression by the poly (dT-dG). poly (dC-dA) sequence. *Molecular and cellular biology*, *4*(12), pp.2622-2630.

Heale, S.M. and Petes, T.D., 1995. The stabilization of repetitive tracts of DNA by variant repeats requires a functional DNA mismatch repair system. *Cell*, *83*(4), pp.539-545.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L. and McLaren, S., 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), p.498.

Koreth, J., O'Leary, J.J. and O'D McGee, J.A.M.E.S., 1996. Microsatellites and PCR genomic analysis. *The Journal of pathology*, *178*(3), pp.239-248.

Liu, J. and Stainier, D.Y., 2012. Zebrafish in the study of early cardiac development. *Circulation research*, *110*(6), pp.870-874.

Meloni, R., Albanèse, V., Ravassard, P., Treilhou, F. and Mallet, J., 1998. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. *Human Molecular Genetics*, *7*(3), pp.423-428.

Muneer, P.A., Gopalakrishnan, A., Musammilu, K.K., Mohindra, V., Lal, K.K., Basheer, V.S. and Lakra, W.S., 2009. Genetic variation and population structure of endemic yellow catfish, *Horabagrus brachysoma* (Bagridae) among three populations of Western Ghat region using RAPD and microsatellite markers. *Molecular biology reports*, *36*(7), pp.1779-1791.

Nagpure, N.S., Rashid, I., Pati, R., Pathak, A.K., Singh, M., Singh, S.P. and Sarkar, U.K., 2013. FishMicrosat: a microsatellite database of commercially important fishes and shellfishes of the Indian subcontinent. *BMC genomics*, *14*(1), p.630.

Naylor, L.H. and Clark, E.M., 1990. d (TG) n· d (CA) n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. *Nucleic Acids Research*, *18*(6), pp.1595-1601.

Oliveira, E.J., Pádua, J.G., Zucchi, M.I., Vencovsky, R. and Vieira, M.L.C., 2006. Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, *29*(2), pp.294-307.

Rouchka, E.C., 2010. Database of exact tandem repeats in the Zebrafish genome. *BMC genomics*, *11*(1), p.347.

Sambrook, S.J., Russel, D.W., Janssen, K.A. and Irwuin, N.J., 2001. Molecular Cloning, A Laboratory Manual (Third

edition), Cold Spring Harbor Laboratory Press.

Shimoda N., Knapik E W., Ziniti J., Shimoda C., N., Knapik, E.W., Ziniti, J., Sim, C., Yamada, E., Kaplan, S., Jackson, D., de Sauvage, F., Jacob, H. and Fishman, M.C., 1999. Zebrafish genetic map with 2000 microsatellite markers. *Genomics*, *58*(3), pp.219-232.16.

Subramanian, S., Mishra, R.K. and Singh, L., 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome biology*, *4*(2), p.R13.

Tautz, D. and Renz, M., 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic acids research*, *12*(10), pp. 4127-4138.

Tautz, D., 1989. Hypervariabflity of simple sequences as a general source for polymorphic DNA markers. *Nucleic acids research*, *17*(16), pp.6463-6471.

Urquhart, A., Kimpton, C.P., Downes, T.J. and Gill, P., 1994. Variation in short tandem repeat sequences—a survey of twelve microsatellite loci for use as forensic identification markers. *International journal of legal medicine*, *107*(1), pp.13-20.

Wahls, W.P., Wallace, L.J. and Moore, P.D., 1990. The Z-DNA motif d (TG) 30 promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. *Molecular and cellular biology*, *10*(2), pp.785-793.

Wang, Y.H. and Griffith, J., 1995. Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. *Genomics*, *25*(2), pp.570-573.

How to cite this article:

Mujahidkhan A. Pathan, E.A. Nesnas, Aditya Pratap Acharya, Rameez Roshan, Thushar P. Kumar, P. Gireesh-Babu, Shrinivas Jahageerdar, Aparna Chaudhari and Gopal Krishna. 2017. *In Silico* Characterisation of Selected Microsatellite Loci Reported in Zebrafish Genome. *Int.J.Curr.Microbiol.App.Sci.* 6(11): 2244-2252. doi: https://doi.org/10.20546/ijcmas.2017.611.266