Biochemical Systematics and Ecology 67 (2016) 95–102

Contents lists available at ScienceDirect

Biochemical Systematics and Ecology

journal homepage: www.elsevier.com/locate/biochemsyseco

Genome-wide distribution and organization of microsatellites in six species of birds

Jie Huang [a], WuJiao Li [b], ZuoYi Jian [b], BiSong Yue [b], YongFeng Yan [a, *]

[a] College of Life Science, Shangqiu Normal University, Shangqiu, 476000, PR China
[b] Key Laboratory of Bio-resources and Eco-environment (Ministry of Education), College of Life Sciences, Sichuan University, Chengdu, 610064, PR China

ARTICLE INFO

Article history:
Received 21 March 2016
Received in revised form 24 May 2016
Accepted 29 May 2016
Available online 7 June 2016

Keywords:
Microsatellites
Six birds
Genome
Distribution

ABSTRACT

Simple sequence repeats (SSRs), or microsatellites, are important genetic markers and play a significant role in genome organization. The genome-wide analyses of microsatellite distributions in six bird species (*Gallus gallus*, *Meleagris gallopavo*, *Taeniopygia guttata*, *Geospiza fortis*, *Melopsittacus undulates*, and *Columba livia*) were performed using *in silico* data mining approach. Under our search criteria, the total numbers of 1–6 bp perfect microsatellites detected ranged from 90,346 to 282,728 and covered from 0.13 to 0.49% in the complete genomes of the six bird species. The SSR abundance was not correlated with genome size, and mononucleotide repeats outnumbered other SSR categories in all of the six species examined. However, there is was a little difference in the most common repeat motifs for each length among the six different bird species considered, with obvious relation to the AT-richness of their genomes. The distribution of SSRs in the different genomic regions of three bird genomes (*G. gallus*, *M. meleagris*, and *T. guttata*) indicated that the intergenic regions exhibited the highest relative abundance compared to the intron and exon regions in all the six motif lengths. In the genome of *G. gallus*, the motif of (AGGA)n made up the most iterated microsatellite locus (spanning 108 repetitions) and 83.6% of these most abundant SSR motifs had a repeat number less than 12. What's more, the number of microsatellites in different chromosomes was positively correlated with the size of the chromosomes in the genome of *G. gallus*.

1. Introduction

Microsatellite loci, also known as simple sequence repeats (SSRs), have been preferred as the most popular and versatile neutral genetic markers. During the past several years, SSRs have been widely employed in genome mapping, forensics, population structure, phylogenetics, linkage, and kinship relationships (Dib et al., 1996; Selkoe and Toonen, 2006; Li et al., 2010). In addition, SSRs have been found in recent studies to serve a functional role affecting gene regulation, transcription, protein function, and genome organization (Kashi and King, 2006; Lawson and Zhang, 2006). Moreover, some diseases, such as human neurodegenerative disorders and some human cancers, have been associated with microsatellite expansion or instability (Oda et al., 2005; Pearson et al., 2005). The importance of microsatellites stems from their hyper-variability and their abundance in some genomes (Ellegren, 2004). What's more, the most conventional procedure for SSR isolation (i.e.,

* Corresponding author.
E-mail address: yanyf01@sina.com (Y. Yan).

http://dx.doi.org/10.1016/j.bse.2016.05.023
0305-1978/

constructing SSR-enriched libraries, screening of the resulting library, and sequencing of the positive clones (Zane et al., 2002)) are labor-intensive, time-consuming, and costly. Such methods generally use one or a few specific repeated motifs that are most often selected without prior knowledge of their abundance in the genome (Castagnone-Sereno et al., 2010).

The recent completion of genome sequencing projects, together with new methodological developments of *in silico* mining of microsatellites, have provided the opportunity to carry out genome-wide scans for different microsatellite motif markers. Study on the distribution of microsatellites will both describe the basic patterns of microsatellite distribution and diversity across whole genomes and began to address fundamental aspects of their evolution in a comparative setting (Pannebakker et al., 2010). Recent studies have focussed on the genome-wide distribution of SSRs in different species (Subramanian et al., 2003; Castagnone-Sereno et al., 2010; Hickner et al., 2010). A better insight into the occurrence of microsatellites in a range of taxa may help to understand the evolution of simple repeats. Previous studies have found the relative abundances of several repeat motifs differed in mammals, invertebrates, and plants (Grover et al., 2007; Sonah et al., 2011). Absolute numbers of microsatellites also tend to correlate positively with genome size (Toth et al., 2000; Katti et al., 2001). Among birds, microsatellite markers have been very poorly developed, and their practical application has been limited. Few data are currently available about SSR genomic distribution and abundance across different bird taxa. Recent advances in bird genomics, such as the release of the genome sequence assembly of the Red Jungle Fowl (*Gallus gallus*) (International Chicken Genome Sequencing Consortium, 2004) and the Zebra Finch (*Taeniopygia guttata*) genome assembly (Clayton et al., 2005), as well as studies of the Wild Turkey (*Meleagris gallopavo*), the Medium Ground Finch of the Galapagos Islands (*Geospiza fortis*), the Budgerigar (*Melopsittacus undulates*), and the Rock Dove (*Columba livia*), provided a range of species for examining the development of avian genomic distribution and abundance. These bird species that already had the needed genomic information available, here, we first report the survey and comparative analysis of microsatellites for these six birds. Contrasted patterns of microsatellite abundance and diversity were characteristic among the six bird genomes, even in the case of two species in the same family (*G. gallus* and *M. gallopavo* of the Phasianidae). The large set of genome-wide distributed markers, which are anchored in the bird reference genomes, may open up the possibility to build linkage maps, conduct comparative genomics, map quantitative trait loci, understand heterozygosity-fitness correlations, and reveal the underlying genetic basis of phenotypic variation.

2. Methods

2.1. Sequence data

The genome sequence data of *T. guttata*, *M. gallopavo*, and *G. gallus* were downloaded directly from the sequencing project websites at UCSC (http://hgdownload.cse.ucsc.edu/downloads.html), which were assembled into scaffolds or chromosomes. The exon, intron, inter-genic regions, and gene position information of these three genomes were downloaded from UCSC (http://genome.ucsc.edu/cgi-bin/hgTables). The genome sequences of *G. fortis*, *M. undulatus*, and *C. livia* were downloaded from http://www.diark.org/diark/species, which were available only as supercontigs. All the genome sequences were downloaded in FASTA format. The total number of bp searched and (G + C) content for each of the five genomes are indicated in Table 1. For *G. gallus*, the whole set of 23,392 protein-coding sequences (including splice variants) predicted from the whole-genome sequence was included in the analysis. The whole set of protein-coding sequences for Turkey (*M. gallopavo*) and the Zebra Finch (*T. guttata)* were 173,73 and 19,297, respectively.

2.2. Sequence analyses

The genome sequences were scanned for the presence of microsatellite motifs using the software MSDB v2.4.1 (Du et al., 2013) downloaded at http://msdb.biosv.com/. MSDB is a new Perl program providing a user-friendly interface for identification and building databases of microsatellites from complete genome sequences. Detection criteria in this study were restricted to perfect repeat motifs of 1–6 bp and a minimum repeat unit was defined as 12 for mono-, 7 for di-, 5 for tri-, and 4 for tetra-, penta- and hexa-nucleotides. For each motif type, these minimum numbers of repeats have been optimized as default parameters of the software to eliminate repeats which might be observed by chance. Therefore, all the scanning

Table 1
Global coverage and density of the microsatellite loci identified in the genomes of six avian species.

	G. gallus	*M. gallopavo*	*T. guttata*	*G. fortis*	*M. undulatus*	*C. livia*
Sequence analyzed (bp)	1,100,463,666	1,061,800,384	1,233,169,488	1,065,292,181	1,117,355,426	1,107,971,856
GC content (in %)	39.4	35.7	41.1	41.6	41.2	41.5
Number of microsatellite loci	282,728	177,733	272,794	159,614	90,346	227,758
Relative abundance (no./Mbp)	256.9	167.4	221.2	149.8	80.9	205.6
Total length of microsatellites (bp)	5,362,166	2,969,887	5,389,375	3,605,605	1,472,726	4,495,887
The density (length in bp/Mbp)	4872.6	2797.0	4370.3	3462.6	1318.0	4057.8
Genome content (in %)	0.49	0.28	0.44	0.34	0.13	0.41
Chromosome W SSR NO.	173	11	NULL			
Chromosome Z SSR NO.	23,713	7326	19,126			