动物科学

DOI:10.11721/cqnuj20190516

白鹭基因组微卫星分布规律研究*

黄　杰[1]，杨　波[2,3]，贾银平[1]，杨承忠[4]

(1. 商丘师范学院 生物与食品学院，河南 商丘 476000；2. 中国大熊猫保护研究中心，四川 都江堰 611830；
3. 大熊猫国家公园珍稀动物保护生物学国家林业和草原局重点实验室，四川 都江堰 611830；
4. 重庆师范大学 生命科学学院 重庆市动物生物学重点实验室，重庆 401331)

摘要:【目的】研究白鹭(*Egretta garzetta*)的全基因组微卫星分布规律。【方法】利用生物信息学方法对已报道的白鹭全基因组进行查询和分析。【结果】白鹭全基因组中 1～6 个碱基重复的微卫星有 255 630 个，序列总长度为 4 282 844 bp，占全基因序列长度的 0.37%。不同重复类型的微卫星中，单碱基的数量最多，有 207 108 个，占全部类型的 81.0%；然后依次是二碱基、三碱基、四碱基、五碱基、六碱基，分别占全部类型的 6.4%，5.3%，4.9%，2.0%和 0.4%。A，C，AC，AT 和 AG 是白鹭基因组微卫星序列中重复数量较多的拷贝类型，占全部重复类型的 87.4%。在白鹭基因组中，按降序排列，重复类型出现次数超过 1 000 次的有 17 个，分别是 T，A，AT，G，C，AC，GT，ATT，AAT，ATTT，AGG，AAAC，CCT，AAAT，CT，GTTT 和 AG，占全部微卫星数量的 92.7%。【结论】研究结果对白鹭微卫星的筛选及深入研究提供了数据支持。

关键词:白鹭；微卫星；基因组；分布规律

中图分类号:Q953　　**文献标志码**:A　　**文章编号**:1672-6693(2019)05-0066-06

白鹭(*Egretta garzetta*)为鹳形目(Ciconiiformes)鹭科(Ardeidae)白鹭属(*Egretta*)鸟类，又有春锄、白鸟等称谓；该鸟身形纤瘦，全身白色，为全球性分布的水鸟，常见于江河、湖泊及滩涂等地。白鹭主要分布于亚洲、非洲、欧洲中部和南部、大洋洲等区域，在中国则主要分布在四川、陕西南部、河南等地。目前，有关白鹭的研究主要集中于生态环境指示作用[1]、种群的繁殖习性研究[2]、个体识别[3]等方面。

微卫星(Microsatellite)又称简单重复序列(Simple sequence repeat)或短串联重复(Short tandem repeat)，指基因组中以少数几个(通常为 1～6 个)核苷酸为短重复单元构成的串联长度为数十个核苷酸的 DNA 重复序列。由于微卫星标记具有高效以及稳定的特点，因此被广泛应用于遗传连锁图谱的构建、遗传育种等领域。在研究有关动物遗传特征的研究中，可利用微卫星标记进行物种间、物种内甚至个体间的亲缘关系分析，如遗传育种[5]、亲子鉴定[6]、遗传多样性分析等[7]。目前最常见的获取微卫星的方法有如下 3 种：1) 通过查找公共数据库获取微卫星位点；2) 利用物种间遗传距离相临近的交叉转移扩增；3) 通过基因组 DNA 筛选获取微卫星。

目前鹭科鸟类中有关微卫星标记研究主要集中于位点的筛选[8-9]，有关基因组微卫星分布情况并未见报道。本研究利用生物信息学方法搜索白鹭的全基因组序列，并对基因组中的微卫星进行了分析，以便为今后白鹭微卫星标记的筛选、遗传多样性、遗传育种等方面的研究提供数据支持。

1 材料与方法

1.1 白鹭基因组序列

白鹭全基因组序列来源于网站 http://www.diark.org/diark/species，基因组大小为 1 151.9 Mb，以 FASTA 格式保存。

* **收稿日期**:2019-03-06　**修回日期**:2019-08-26　**网络出版时间**:2019-09-26 11:24

资助项目:国家自然科学基金(No. 31501845)；重庆市自然科学基金(No. cstc2017jcyjAX0165；No. cstc2018jcyjAX0738)；重庆市留学人员回国创业创新支持计划(No. cx2018108)；商丘师范学院高层次人才科研启动项目(No. 50013901)

第一作者简介:黄杰，女，讲师，博士，研究方向为动物学，E-mail:huangjie66666@163.com；**通信作者**:杨承忠，男，副教授，博士，E-mail:drc-zyang@126.com

网络出版地址:http://kns.cnki.net/kcms/detail/50.1165.N.20190926.1123.006.html

1.2 微卫星相关术语的定义

为了完全统计重复序列，特将与重复序列有关的 4 个术语进行说明。

1）重复类型：指每个重复序列中核心重复单元是由多少碱基(bp)组成。依据核心重复单元的碱基数的多寡，可将微卫星分为单碱基、二碱基、三碱基、四碱基、五碱基和六碱基共 6 种类型。

2）重复拷贝数：指每个微卫星重复序列中核心单元出现的次数，比如$(AACA)_{28}$，这个重复序列的重复拷贝数就是 28。

3）重复拷贝类别：指每个重复类型由哪几个碱基构成，比如四碱基重复类型 AACA 和 AAGA 分别属于不同的重复拷贝类别，而 AAAC，AACA 则属于 1 个重复拷贝类别。

4）重复数目：指在基因组中，每个重复类型的碱基数目。

1.3 统计软件

对白鹭基因组微卫星的搜索通过微卫星搜索及统计软件 MSDB v2.4[10]来完成，此软件用 Perl 程序语言编程，界面友好，操作简便快捷，可以准确辨认并建立物种的微卫星序列资源库，并集成了微卫星分类及统计各种功能，在不同物种微卫星研究中具有广泛应用[11-17]。有关统计标准设置如下：单碱基重复序列中，筛选的拷贝数大于或等于 12；二碱基重复序列中，筛选的拷贝数大于或等于 7；三碱基重复序列中，筛选的拷贝数大于或等于 5；四碱基、五碱基和六碱基重复序列中，筛选的拷贝数大于或等于 4。

2 结果

2.1 基因组中不同重复类型微卫星的基本情况

表 1 显示：在白鹭基因组中共发现 255 630 个微卫星序列，总长度为 4 282 844 bp，占全基因长度的 0.37%；总丰度为 3 718.07 bp·Mb^{-1}，总频率为 221.92 个·Mb^{-1}。从表 1 可以明显看出：单碱基类型的数量最多；然后依次为二碱基、三碱基、四碱基和五碱基类型；数量最少的是六碱基类型。同样地，白鹭不同重复类型微卫星的总长、丰度和频率也遵循上述规律(表 1)。此外在白鹭的基因组中，通过降序规律排列，可以发现：微卫星重复基本类型出现频率超过 1 000 次的有 17 种，分别是 T，A，AT，G，C，AC，GT，ATT，AAT，ATTT，AGG，AAAC，CCT，AAAT，CT，GTTT 和 AG，它们占全部微卫星数量的 92.7%；在基因组中，微卫星重复基本类型出现频率超过 100 次的有 66 种，占基因组全部总数的 98.5%。

表 1　白鹭基因组中不同重复类型微卫星的基本信息

Tab.1　Distribution of SSR of different repeat type in genome of *E. garzetta*

重复类型	数量/个	总长/bp	频率/(个·Mb^{-1})	丰度/(bp·Mb^{-1})
单碱基	207 108(81.0%)	3 246 092	179.80	2 818.03
二碱基	16 346(6.4%)	282 546	14.19	245.29
三碱基	13 448(5.3%)	251 151	11.67	218.03
四碱基	12 614(4.9%)	243 324	10.95	211.24
五碱基	5 075(2.0%)	201 915	4.41	175.29
六碱基	1 039(0.4%)	57 816	0.90	50.19
总计	255 630(100.0%)	4 282 844	221.92	3 718.07

注："数量"一列中括号内的数据为某一重复类型微卫星数量占微卫星总数量的百分比。

2.2 不同重复类型微卫星重复序列的数量

白鹭基因组中不同重复类型微卫星的出现次数范围有较大差异，重复次数最少的仅 4 次，重复次数最多的(TCCTT)达 380 次。重复 4～6 次的有 855 个，占全部类型微卫星数量的 49.5%；重复 7～15 次的有 622 个，占全部类型微卫星数量的 36.0%；重复 16～30 次的有 143 个，占全部类型微卫星数量的 8.3%；重复 16～30 次的有 58 个，占全部类型微卫星数量的 3.4%；重复 52～380 次的有 49 个，占全部类型微卫星数量的 2.8%。在白鹭基因组中，不同重复类型微卫星的主要重复次数也不同：单碱基类型中以 12～26 个重复的居多，占全部单碱基类型的 96.2%；二碱基类型中以 7～10 个重复居多，占全部二碱基类型的 85.2%；三碱基类型中以 4～7 个重复的居多，占全部三碱基类型的 85.0%；四碱基类型中以 4～7 个重复的居多，占全部四碱基类型的 95.8%；五碱基类型中以 4～10 个重复居多，占全部五碱基类型的 88.9%；六碱基类型中以 4～8 个重复居多，占全部六碱基类型的 83.6%。但从整体上来看，白鹭基因组中微卫星序列重复次数在 4～40 之间的居多，占全部类型微卫星重复序列数量的 95.8%(图 1)。

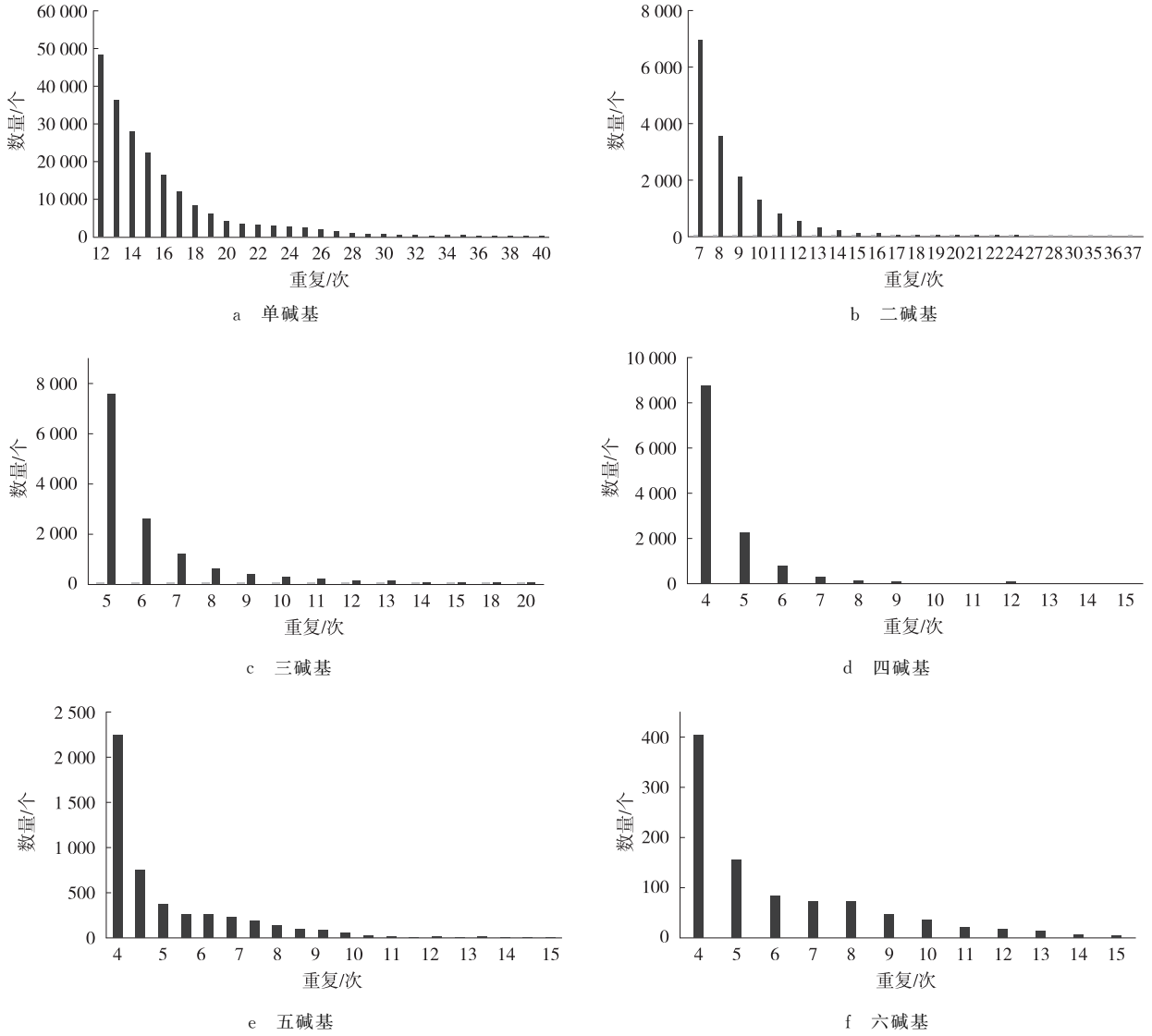

图1 单碱基至六碱基重复类型的拷贝数的数量分布

Fig. 1 The number distribution of copy numbers in 1～6 bp repeats

2.3 各重复拷贝类别的数量

表2显示:白鹭基因组中出现数量最多单碱基重复拷贝类别是A,有196 318个,C相对最少,仅有10 790个。二碱基重复类型中最多的是AC,有8 052个;其次是AT和AG,分别有5 560和2 721个;最少的是GC仅有12个。

表2 白鹭基因组中主要重复类型微卫星的拷贝类别数量

Tab. 2 Number of the main repeat type of SSR in genome of *E. garzetta*

重复类型	拷贝类别	数量/个	重复类型	拷贝类别	数量/个
单碱基	A	196 318(94.80%)	三碱基	AAT	4 623(34.38%)
	C	10 790(5.20%)		ACC	1 079(8.02%)
总计		207 108(100.00%)		ACG	592(4.40%)
二碱基	AC	8 053(49.26%)		ACT	645(4.80%)
	AG	2 721(16.65%)		AGC	625(4.65%)
	AT	5 560(34.02%)		AGG	2 970(22.09%)
	CG	12(0.07%)		AGT	671(4.98%)
总计		16 346(100.00%)		CCG	55(0.41%)
三碱基	AAC	1 663(12.37%)	总计		13 448(100.00%)
	AAG	525(3.90%)			

注:"数量"一列中括号内的数据为某一重复类型微卫星某一拷贝类别数量该重复类型微卫星总数量的百分比。

三碱基重复类型中，最多的是 AAT，有 4 623 个，其次是 AGG，有 2 970 个，最少的是 CCG，仅有 55 个。

3 讨论

3.1 全基因组上微卫星序列的分布特性

随着生物信息学技术飞速发展，从数据库中可获得的生物基因组序列也逐渐增多。在基因组水平上来分析各个物种的微卫星序列，不仅可以认识到各个物种的微卫星序列特征，而且还可以揭示微卫星在生物基因组内的作用。本研究通过在网站 http://www.diark.org/diark/species 中下载白鹭基因组，并通过 MSDB[10] 软件搜索和分析了白鹭基因组中存在的微卫星序列，在 1 151.9 Mb 全基因组序列中共统计出微卫星位点 255 630 个，所有的微卫星序列总长 4 282 844 bp，占全基因组序列长度的 0.37%，与虎皮鹦鹉（*Melopsittacus undulatus*）[11]、林麝（*Moschus berezovskii*）[12]、红原鸡（*Gallus gallus*）[13] 微卫星序列占全基因序列长度的比例相近（三者依次分别为 0.41%，0.42% 和 0.49%），但是与大熊猫（*Ailuropoda melanoleuca*）、北极熊（*Ursus maritimus*）[14] 以及猪（*Sus scrofa*）[15] 的这一比例相差较大（三者依次分别为 0.64%，0.79% 和 0.85%）。

白鹭基因组中单碱基重复类型在所有的微卫星重复类型中所占的比例最高，为 81.0%，并且单碱基类型中以（A）*n* 和（T）*n* 为主，占 94.8%，这与其他物种如猪[15]、人（*Homo sapiens*）[16]、大鼠（*Rattus norvegicus*）[17] 等哺乳动物的有关研究结果相符合；而果蝇（*Drosophila melanogaster*）[16] 中两碱基类型所占比例最高，黑粉菌（*Ustilago maydis*）等[18] 低等生物中三碱基类型所占比例最高。白鹭二碱基类型微卫星数量有 16 346 个，占全部微卫星重复序列数量的 6.4%；其中 AC 数量最多，为 8 053 个；GC 含量最少，仅有 12 个。这与大鼠[17]、大黄鱼（*Larimichthys crocea*）[19]、牛（*Bos taurus*）[20]、绵羊（*Ovis aries*）[20]、小鼠（*Mus musculus*）[21] 二碱基类型微卫星中 AC 的分布规律相同，与红原鸡[13] 的二碱基类型微卫星中 AT 所占比例最高的结果不相同。此外，白鹭三碱基类型微卫星数量有 13 448 个，占全部微卫星重复序列的 5.3%，其中 AAT 所占比例最高，同酵母[18] 的三碱基类型微卫星中 AAT 所占比例最高的结果相同。然而，人类基因组的三碱基类型微卫星含有较多的 AAT 和 AAC，果蝇[16] 基因组的该类微卫星含有较多的 AGC。因此，不同物种中微卫星的分布规律存在着一定的差异，但这一点是否与物种的进化有关，以及它和不同生物的基因表达与调控是否存在关联，仍需进一步研究。

3.2 二碱基重复类型的分布特征

对小鼠[21]、猪[15]、牛和绵羊[20] 的基因组微卫星分析结果表明：在二碱基类型微卫星中，AC 所占比例最高，而红原鸡[13] 的基因组这一类型的微卫星中 AT 所占比例最高。本研究结果则显示，白鹭二碱基类型微卫星中 AC 所占比例最高，它的数量占全部二碱基类型微卫星数量的 49.26%。上述差异可能由不同的物种中占优势的二碱基类型有所差异而导致。而二碱基类型微卫星中 GC 所占比例在已研究的物种中都是较少的。有研究发现，GC 所占比例高的物种，它的基因组内微卫星所占比例也较低。Shorderet 等人[22] 研究了 6 种脊椎动物的基因组后，给出的解释是：由于基因组中的 CpG 甲基化，因而使它成为一个易突变的点；而甲基化的胞苷酸 C 又容易在脱氨基的作用下转变为胸腺嘧啶 T；然而 GC 又是保证 DNA 热力学所必备的。由此导致的结果是：GC 在其中越来越少，相对应的 TG 所占比例逐渐增加。这个解释可以在某些程度上阐明为何在人类及某些生物基因组二碱基类型微卫星中，AC 重复较多。虽然微卫星中 GC 所占比例相对较低，但是研究表明，这一拷贝类别在基因组中依然起着重要的作用，Griffiths 等人[23] 发现，胸苷激酶的 G（*n*）突变参与了单纯疱疹病毒的再活化。由此推断，本研究中白鹭二碱基类型微卫星中 GC 所占比例较少可能与上述原因有一定的关系，因为与之相对应的突变后的 AC 所占比例在二碱基类型中是最高的。然而，白鹭基因组的二碱基类型微卫星中 GC 所占比例较少的具体原因仍需要进一步研究。

参考文献：

[1] 夏秋烨，倪才英，赵中华，等. 鄱阳湖夏候鸟小白鹭对环境样品中重金属的富集研究[J]. 长江流域资源与环境，2014，23(11)：1540-1544.
XIA Q Y，NI C Y，ZHAO Z H，et al. Research on enrichment of heavy metals in little egret of the Poyang lake[J]. Resources and Environment in the Yangtze Basin，2014，23(11)：1540-1544.

[2] 韩庆，梁瑜，何超. 湖南花岩溪白鹭繁殖习性研究[J]. 四川动物，2008，27(4)：594-598.
HAN Q，LIANG Y，HE C. Breeding characteristics of *Egretta garzetta* in Huayanxi，Hunan province[J]. Sichuan Journal of Zoology，2008，27(4)：594-598.

[3] 管昊,林清贤,周晓平,等.白鹭脱落羽毛的微卫星个体识别研究[J].厦门大学学报(自然科学版),2013,52(5):710-717.
GUAN H,LIN Q X,ZHOU X P,et al. Microsatellite individual identification for moulted feathers in *Egretta garzetta*[J]. Journal of Xiamen University (Natural Science),2013,52(5):710-717.

[4] 林清贤,周晓平,方文珍,等.中国6种白色羽鹭科鸟类的系统归属研究[J].厦门大学学报(自然科学版),2010,49(1):83-86.
LIN Q X,ZHOU X P,FANG W Z,et al. Phylogenetic relationships of 6 species egrets with white plumes in China [J]. Journal of Xiamen University (Natural Science),2010,49(1):83-86.

[5] QI W,CHEN X,FANG P,et al. Genomic and transcriptomic sequencing of *Rosa hybrida* provides microsatellite markers for breeding,flower trait improvement and taxonomy studies[J]. BMC Plant Biology,2018,18(1):119.

[6] YAMAMOTO S,KOMASU H,KITAURA J,et al. Development of 11 microsatellite markers and paternity analysis in the invasive apple snail *Pomacea canaliculata*[J]. Venus (Journal of the Malacological Society of Japan),2018,76 (1/2/3/4):79-85.

[7] BARBIAN H J,CONNELL A J,AVITTO A N,et al. CHIIMP:an automated high-throughput microsatellite genotyping platform reveals greater allelic diversity in wild chimpanzees[J]. Ecology and Evolution,2018,8(16):7946-7963.

[8] HUANG X,ZHOU X,CHEN M,et al. Isolation and characterization of microsatellite loci in vulnerable Chinese egret (*Egretta eulophotes*:aves)[J]. Conservation Genetics,2010,11(3):1211-1214.

[9] HILL A,GREEN M C. Characterization of 12 polymorphic microsatellites for the reddish egret,*Egretta rufescens*[J]. Conservation Genetics Resources,2011,3(1):13-15.

[10] DU L M,LI Y Z,ZHANG X Y,et al. MSDB:A user-friendly program for reporting distribution and building databases of microsatellites from genome sequences[J]. Journal of Heredity,2013,104(1):154-157.

[11] 黄杰,原宝东,杨承忠.虎皮鹦鹉全基因组中微卫星分布规律研究[J].野生动物学报,2017,38(3):422-426.
HUANG J,YUAN B D,YANG C Z. Distribution regularities of microsatellites in the *Melopsittacus undulatus* genome[J]. Chinese Journal of wildlife,2017,38(3):422-426.

[12] 卢婷,王晨,杜超,等.林麝全基因组微卫星分布规律研究[J].四川动物,2017,36(4):420-424.
LI T,WANG C,DU C,et al. Distribution regularities of microsatellites in *Moscsus berezovskii* genome[J]. Sichuan Journal of Zoology,2017,36(4):420-424.

[13] 黄杰,杜联明,李玉芝,等.红原鸡全基因组中微卫星分布规律研究[J].四川动物,2012,31(30):358-363.
HUANG J,DU L M,LI Y Z,et al. Distribution regularities of microsatellites in *Gallus gallus* genome[J]. Sichuan Journal of Zoology,2012,31(30):358-363.

[14] 李午佼,李玉芝,杜联明,等.大熊猫和北极熊基因组微卫星分布特征比较分析[J].四川动物,2014,33(4):874-878.
LI W J,LI Y Z,DU L M,et al. Comparative analysis of microsatellites sequence distribution in the genome of giant panda and polar bear[J]. Sichuan Journal of Zoology,2014,33(4):874-878.

[15] 戚文华,蒋雪梅,肖国生,等.猪全基因组中微卫星分布规律[J].畜牧与兽医学,2014,46(8):9-12.
QI W H,JIANG X M,XIAO G S,et al. Distribution regularities of microsatellites in the pig genome[J]. Acta Veterinaria et Zootechnica Sinica,2014,46(8):9-12.

[16] KATTI M V,RANJEKAR P K,GUPTA V S,et al. Differential distribution of simple sequence repeats in eukaryotic genome sequences[J]. Molecular Biology and Evolution. 2001,18(7):1161-1167.

[17] 涂飞云,刘晓华,杜联明,等.大鼠全基因组微卫星分布特征研究[J].江西农业大学学报,2015,37(4):708-711.
TU F Y,LIU X H,DU L M,et al. Distribution characteristic of microsatellites in the rat genome[J]. Acta Agriculture Universitatis Jangxiensis,2015,37(4):708-711.

[18] KARAOGLU H,LEE C M,MEYER W. Survey of simple sequence repeats in completed fungal genomes[J]. Molecular Biology & Evolution,2005,22(3):639-649.

[19] 叶华,任鹏,刘洋,等.大黄鱼微卫星标记的开发及其遗传方式分析[J].水生生物学报,2012,36(6):1156-1163.
YE H,REN P,LIU Y,et al. Isolation and genetic analysis of microsatellite markers for *Larimichthys crocea*[J]. Acta Hydrobiologica Sinica,2012,36(6):1156-1163.

[20] 戚文华,蒋雪梅,肖国生,等.牛和绵羊全基因组微卫星序列的搜索及其生物信息学分析[J].畜牧与兽医学报,2013,44(11):1724-1733.
QI W H,JIANG X M,XIAO G S,et al. Seeking and bioinformatics of microsatellites sequence in the genome of cow and sheep [J]. Acta Veterinaria et Zootechnica Sinica,2013,44(11):1724-1733.

[21] 童晓玲,代方银,李斌,等.小鼠基因组中的微卫星重复序列的数量、分布和密度[J].动物学报,2006,52(1):138-152.
TONG X L,DAI F Y,LI B,et al. Microsatellite repeats in mouse:abundance,distribution and density[J]. Acta Zoologica Sinica,2006,52(1):138-152.

[22] SCHORDERET D F,GARTER S M. Analysis of CpG

suppression inmethylated and species[J]. Proceedings of the National Academy of Sciences of USA, 1992, 89(3): 957-961.

[23] GRIFFITHS A, LINK M A, FURNESS C L, et al. Low-level expression and reversion both contribute toreavtivation of herpes simplex virus drug-resistant mutants with mutations on homopolymeric sequcnces in thymidine kinase[J]. Journal of Virology, 2006, 80(13): 6568-6574.

Animal Sciences

The Distribution Regularities of Microsatellites in *Egretta garzetta* Genome

HUANG Jie[1], YANG Bo[2,3], JIA Yinping[1], YANG Chengzhong[4]

(1. College of Biology and Food, Shangqiu Normal University, Shangqiu Henan 476000;
2. China Conservation and Research Centre for the Giant Panda, Dujiangyan Sihuan 611830;
3. Key Laboratory of State Forestry and Grassland Administration on Conservation Biology of Rare Animals in The Giant Panda National Park, Dujiangyan Sichuan 611830; 4. Chongqing Key Laboratory of Animal Biology, College of Life Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: [**Purposes**]It is focused on the distribution regularities of microsatellites in *Egretta garzetta*. [**Methods**]The whole genome sequences of *E. garzetta* were analyzed by bioinformatics methods. [**Findings**]The results showed that: Microsatellite (simple sequence repeat, SSR) with 1～6 bp nucleotide motifs were detected in the *E. garzetta* were 255 630 loci, 4 282 844 bp in length, which account for 0.36% of the whole genome sequences. These 207 108 mononucleotide microsatellites, which accounted for 81% of all the different repeat types, is the most abundant type. Followed by dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, hexanucleotide microsatellites which accounted for 6.4%, 5.3%, 4.9%, 2%, and 0.4%, respectively. A, C, AC, AT, and AG were predominate in *E. garzetta* genome, accounted for 87.4% all together. The repeat times of T, A, AT, G, C, AC, GT, ATT, AAT, ATTT, AGG, AAAC, CCT, AAAT, CT, GTTT, and AG were more than 1 000 times, which accounted for 92.7% all together. [**Conclusions**]This data provided here will give a light on screening and further research of *E. garzetta* microsatellites.

Keywords: *Egretta garzetta*; microsatellite; genomic; distribution regularities

(责任编辑 方 兴)