DOI:10.11984/j.issn.1000-7083.20170070

红尾蚺和原矛头蝮基因组微卫星分布特征比较分析

聂虎， 曹莎莎， 赵明朗， 杜林方*

(四川大学生命科学学院生物资源与生态环境教育部重点实验室，成都 610064)

摘要：本研究分析比较了红尾蚺 *Boa constrictor* 和原矛头蝮 *Protobothrops mucrosquamatus* 基因组完美型微卫星的分布特征，通过 MISA 工具分别鉴定出 398 860 个和 422 364 个微卫星，其总长分别为 8 550 741 bp 和 12 243 226 bp，分别占基因组大小的 0.59%和 0.73%，在各自基因组中的丰度分别为 275.46 个/Mbp 和 252.33 个/Mbp。红尾蚺基因组中单碱基重复类型微卫星最多，其次是四碱基、二碱基、三碱基、五碱基和六碱基，最丰富的 5 种微卫星类型是 A、AC、AAAT、AG、AAT；原矛头蝮基因组中单碱基重复类型微卫星最多，其次是三碱基、四碱基、二碱基、五碱基和六碱基，最丰富的 5 种微卫星类型是 A、AAT、AC、C、AAAT。红尾蚺和原矛头蝮微卫星在基因组不同区域丰度不同，基因间区丰度最高，其次是内含子区和外显子区，编码区微卫星丰度最低，表明编码区微卫星受到的选择压力最大。红尾蚺和原矛头蝮在基因中微卫星密度分布的位置特征相似，即微卫星在基因上下游 500 bp 密度最高，在内含子区次之，在外显子区密度最低。红尾蚺和原矛头蝮基因编码区所有 6 种重复类型的微卫星中，三碱基重复类型的微卫星占绝对优势。红尾蚺和原矛头蝮基因组中含有微卫星的编码序列分别有 1 480 条和 1 397 条，被 GO 注释的分别有 736 条和 733 条。它们的 GO 功能归类结果类似，但是与其它物种相比存在种系差异。本研究结果为后续开发 2 种蛇类高质量微卫星标记提供了方便，也为进一步探索这些微卫星在它们基因组中的生物学功能提供了有意义的基础数据。

关键词：红尾蚺； 原矛头蝮； 基因组微卫星； 密度分布

中图分类号：Q959.6； Q915.864 **文献标志码：**A **文章编号：**

Comparative Analysis of Microsatellite Distributions in Genomes of *Boa constrictor* and *Protobothrops mucrosquamatus*

NIE Hu, CAO Shasha, ZHAO Minglang, DU Linfang*

(Key Laboratory of Bio-resource and Eco-environment of Ministry of Education, College of Life Sciences，Sichuan University, Chengdu 610064, China)

Abstract：In this study, we analyzed and compared the distributions of perfect microsatellites in the genomes of *Boa constrictor* and *Protobothrops mucrosquamatus*. Using the MISA tool, a total of 398 860 and 422 364 microsatellites were identified in genomes of *B. constrictor* and *Protobothrops mucrosquamatus*, respectively. The total length of the identified microsatellites was 8 550 741 bp in *Boa constrictor* and 12 243 226 bp in *P. mucrosquamatus*, accounting for 0.59% and 0.73% of each genome, respectively. The abundance of microsatellites was 275.46 no./Mbp in *B. constrictor* and 252.33 no./Mbp in *P. mucrosquamatus*. In *B. constrictor* genome, mono-nucleotide repeat was the most abundant, followed by tetra-nucleotide, di-nucleotide, tri-nucleotide, penta-nucleotide and hexa-nucleotide repeat, and A、AC、AAAT、AG、AAT were the five most abundant repeat units. In *P. mucrosquamatus* genome, mono-nucleotide repeat was the most abundant, followed by tri-nucleotide, tetra-nucleotide, di-nucleotide, penta-nucleotide and hexa-nucleotide, and A、AAT、AC、C、AAAT were the five most abundant repeat units. In both species, the abundances of microsatellites in intergenic region was the highest, followed by intron region and exon region, and the lowest was in coding region. These phenomena indicated that microsatellites in coding sequences were subject to the greatest selective pressure. The positional specificity of microsatellite density distributions in these two snakes were similar, that is, the density of microsatellites was the highest in the upstream and downstream 500 bp regions of genes, followed by intron regions and exon regions. Tri-nucleotide repeat was dominant among the six repeat units in the coding sequences of both genomes. The number of coding sequences containing microsatellites were 1480 and 1 397, among which 736 and 733 were assigned with GO terms of known function in genomes of *B. constrictor* and *P. mucrosquamatus*, respectively. These coding sequences resulted the similar GO classification outputs, but behaved in a lineage manner comparing with other species. This study made a great convenience to develop large number of high-quality microsatellite markers for these two snakes and provided meaningful underlying data for further exploration of the biological function of microsatellites in their genomes.

Keywords: *Boa constrictor*; *Protobothrops mucrosquamatus*; genomic microsatellites; density distribution

---

收稿日期：2017-03-08 **接受日期：**2017-05-24

作者简介：聂虎(1991—)， 男， 硕士研究生， 主要从事生物信息学研究

*通信作者 Corresponding author， E-mail:linfangdu@scu.edu.cn

微卫星是由 1~6 个核苷酸为基本重复单元构成的简单串联重复序列(simple sequence repeat，SSR)。微卫星广泛分布于动植物基因组中，但其在基因编码区，非翻译区(UTRs)和内含子区的分布并不随机，并且 5'UTR、3'UTR 和内含子区和外显子区微卫星的收缩或扩张可通过多种方式引起基因功能的改变，从而影响细胞功能，最终导致表型的变化和疾病发生(Li *et al.*，2004)。基因组层面微卫星分析有助于比较不同物种之间微卫星的分布特征、了解基因组功能，并为开发微卫星标记提供方便(李午佼等，2014；Wang *et al.*，2016)。

红尾蚺 *Boa constrictor* 又称红尾蟒，蚺科 Boidae 卵胎生无毒蛇，主要分布于中美洲、南美洲以及加勒比海附近的一些岛屿。在某些地区，红尾蚺在调节负鼠的种群规模中具有重要作用，能防止利什曼病传播给人类(Laurie & Janalee，2009)。原矛头蝮 *Protobothrops mucrosquamatus* 又称龟壳花，蝰科 Viperidae 原矛头蝮属 *Protobothrops* 管牙类毒蛇。原矛头蝮广泛分布于中国大陆以及印度、孟加拉、缅甸等地。原矛头蝮已被列入国家林业局 2000 年 8 月 1 日发布的《国家保护的有益的或者有重要经济、科学研究价值的陆生野生动物名录》。红尾蚺和原矛头蝮高质量的全基因组的测序和组装已经完成(Kajitani *et al.*，2014；Kerkkamp *et al.*，2016)，这为在基因组水平上开展红尾蚺和原矛头蝮微卫星的研究提供了可能。

本研究主要目的有：1）比较有毒蛇原矛头蝮和无毒蛇红尾蚺基因组层面微卫星数量、种类和丰度的异同；2）比较红尾蚺和原矛头蝮基因组不同区域(即基因间区、内含子和外显子)SSR的分布特征；3）比较红尾蚺和原矛头蝮基因区SSR密度分布的位置特征；4）探讨含有SSR的CDS的功能，分析含有SSR的编码基因在2种蛇中的差异。本研究有助于加深对蚺科和蝰科基因组的认识和了解，也为后续筛选和开发大量高质量的2种蛇类微卫星标记提供方便。

1 材料和方法

1.1 数据

原矛头蝮、人 *Homo sapiens* 和小鼠 *Mus musculus* 的基因组从 NCBI Genome 数据库下载，登录号分别为 GCF_001527695.2、GCF_000001405.36、GCF_000001635.25。红尾蚺的基因组从http://platanus.bio.titech.ac.jp/Snake.tgz 下载(Kajitani *et al.*，2014)。

1.2 微卫星鉴定

利用MISA软件在红尾蚺和原矛头蝮基因组搜索1~6个核苷酸重复类型微卫星(Thiel *et al.*，2003)。MISA运行时，misa.ini文件def设置为"1-12 2-7 3-5 4-4 5-4 6-4"，int设置为100。即单核苷酸重复次数不小于12次，二核苷酸重复次数不小于7次，三核苷酸重复次数不小于5次，四核甘酸、五核苷酸和六核苷酸重复次数都不小于4次。如果2个微卫星之间距离小于100 bp则认为是1个复合型微卫星。

1.3 微卫星分类

根据微卫星重复单元的序列，对微卫星进行分类。如果2个微卫星是循环排列，或者反向互补，则认为这2个微卫星归属于同一类微卫星。比如微卫星ACG包括了微卫星ACG、CGA、GAC、TGC、GCT和CTG(Jurka & Pethiyagoda，1995)。

1.4 计算微卫星在基因组中的分布

通过微卫星与特定区域的位置重叠，计算出微卫星在各个区域的分布。如果微卫星的位置与编码基因，或外显子，或内含子的位置完全重叠，则认为微卫星位于编码基因，或外显子，或内含子。否则，认为微卫星位于基因间区。另外，若微卫星位于基因上游500 bp或下游500 bp，则认为微卫星位于基因上游或下游。

1.5 计算微卫星在基因中的密度分布

为了计算微卫星在基因组中的密度分布，把基因中的外显子和内含子归为不同元件：基因上游、第一个外显子、第一个内含子、第二个外显子、第二个内含子等，中间左边外显子、中间内含子、中间右边外显子等，倒数第二个内含子、倒数第二个外显子、倒数第一个内含子、倒数第一个外显子和基因下游等。微卫星在某个类型的元件中的相对位置为微卫星到元件左端的距离除以元件的长度与微卫星长度的差。微卫星的密度为某个类型元件中 SSR 的数量除以该类型元件的总长，单位为个/Mb (Fujimori *et al.*，2003)。具体的计算过程如图一所示，图中"Upstream"表示基因上游 500 bp，"First"表示第一个外显子/内含子，"Second"表示第二个外显子/内含子，"Middle"表示中间左边的外显子/

中间内含子/中间右边外显子，“Last second”表示倒数第二个内含子/外显子，“Last first”表示倒数第一个内含子/外显子，“Downstream”表示基因下游 500 bp。

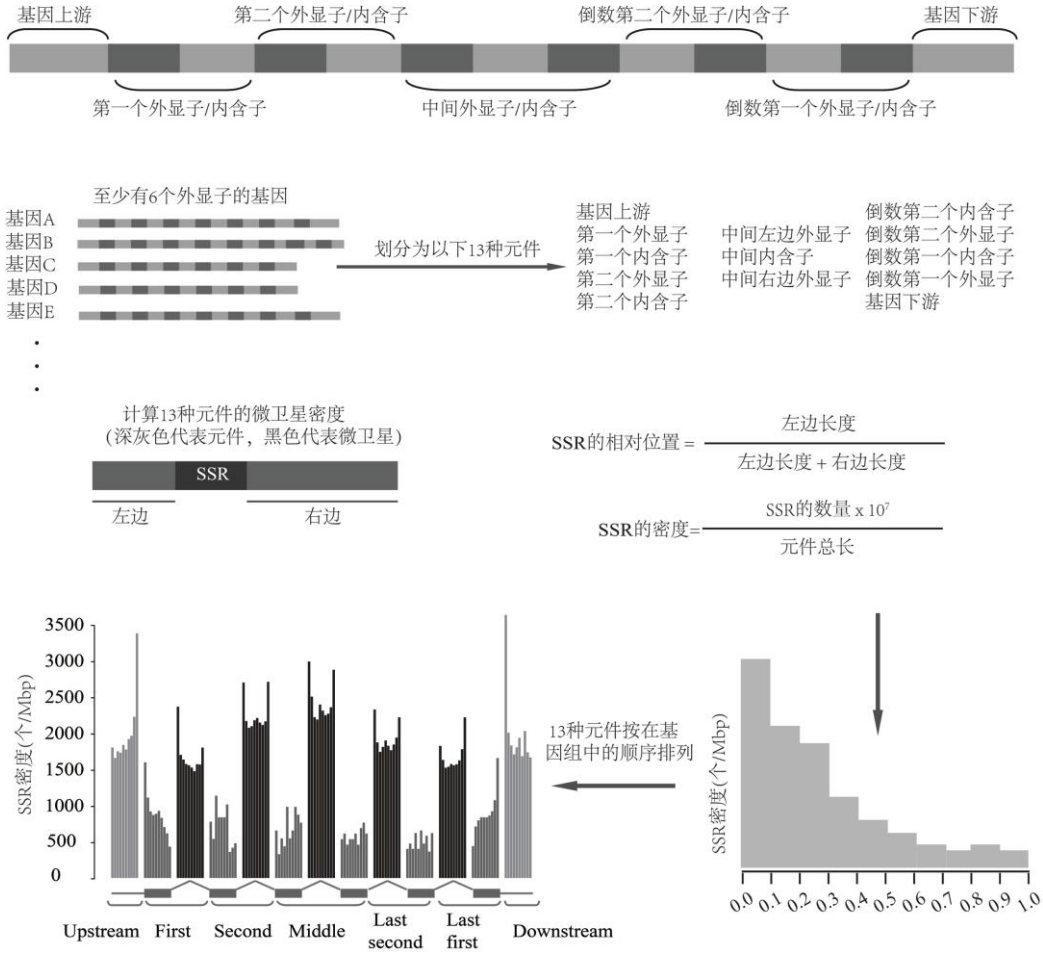

图1　微卫星密度分布的计算方法

Fig. 1　Method to calculate the distribution of microsatellites density

1.6 含微卫星的编码序列的功能分析

对微卫星坐标和基因编码序列的坐标进行重叠，如果某个基因编码序列含有了微卫星，则筛选出该编码序列。将筛选出来的编码序列通过BLASTx比对到NR蛋白质数据库，然后对注释出的蛋白进行GO功能分析(Conesa *et al.*，2005)。使用OrthoMCL对含有微卫星的CDS序列进行基因家族分析(Li *et al.*，2003)。

2 结果

2.1 红尾蚋和原矛头蝮基因组中微卫星的分布特征

利用MISA软件搜索微卫星，在红尾蚋1.5 G全基因组序列中共搜索到398 860个SSR，总长度为8 550 741 bp，占基因组序列的0.59%。在原矛头蝮1.67 G全基因组序列中共搜索到422 364个微卫星，总长度为12 243 226 bp，占基因组序列的0.73%。红尾蚋和原矛头蝮基因组中SSR的丰度分别为275.46个/Mb和252.33个/Mb(表1)，两者的丰度比较相似。红尾蚋基因组中最多的5种SSR类型为A、AC、AAAT、AG和AAT，原矛头蝮基因组中最多的5种SSR类型为A、AAT、AC、C和AAAT(表2)，两者最常见的SSR类型有所不同。红尾蚋基因组中6种重复类型的SSR中，最丰富的是单碱基重复类型，其次是四碱基、二碱基、三碱基、五碱基和六碱基重复类型。原矛头蝮基因组中6种重复类型的SSR中，最丰富的也是单碱基重复类型，其次是三碱基、四碱基、二碱基、五碱基、六碱基重复类型。2个物种单碱基重复类型最丰富的都是$(A)_n$，红尾蚋$(A)_n$类型SSR占单碱基重复类型的88.86%，原矛头蝮$(A)_n$类型SSR占单碱基重复类型的74.37%(表1，表3)。红尾蚋基因组四碱基SSR以$(AAAT)_n$、$(AAAC)_n$、$(AATG)_n$和$(AATG)_n$为主，原矛头蝮基因组三碱基SSR以$(AAT)_n$、$(AGG)_n$、$(AAC)_n$、和$(ATG)_n$为主。2个物种六碱基重复单元的微卫星丰度最低，都以$(ACATAT)_n$类型为主。

通过分析和比较微卫星在红尾蚺基因组和原矛头蝮基因组中各个区域的分布，发现在红尾蚺和原矛头蝮基因组中，基因间区微卫星数量最多、丰度最高，其次是内含子和外显子，编码区微卫星的数量最少，丰富最低。另外，还发现红尾蚺基因组和原矛头蝮基因组非翻译区(UTR)微卫星的丰度比编码区微卫星要高(表4)。我们对人类和小鼠的基因组中的SSR进行了鉴定和分析，发现这2种蛇基因组与这2种哺乳动物相比，基因组编码区SSR的数量和丰度差异很小，而在基因间区、外显子区和内含子区SSR的数量和丰度差异较大(表4)。

分析红尾蚺基因组和原矛头蝮基因组编码区、外显子区和内含子区中微卫星的重复类型，发现两者编码区和外显子区主要是三碱基重复类型的微卫星，红尾蚺基因组编码区三核苷酸重复类型微卫星占编码区所有微卫星的84.07%，原矛头蝮编码区三碱基重复类型微卫星占编码区所有微卫星的95.11%(图2：B)。红尾蚺和原矛头蝮基因间区中SSR的类型主要是单碱基、四碱基、二碱基和三碱基类型，各种类型的SSR都不占据主导性优势(图2：D)，其分布和整个基因组中SSR的分布类似(图2：A)。比较编码区和外显子区SSR的重复类型(图2：B和图2：C)，发现外显子区单碱基类型的SSR比编码区单碱基类型的SSR要多，而外显子含有了编码区和非翻译区，说明了非翻译区中以单碱基重复类型SSR为主。

表1　红尾蚺和原矛头蝮基因组中微卫星的分布

Table 1　Distribution of microsatellites in the genomes of *Boa constrictor* and *Protobothrops mucrosquamatus*

微卫星类型	红尾蚺				原矛头蝮			
	数量/个	长度/bp	丰度/(个/Mb)	比例/%	数量/个	长度/bp	丰度/(个/Mb)	比例/%
单核苷酸	127 438	1 768 476	88.01	31.95	112 325	1 584 938	67.10	26.59
二核苷酸	64 318	1 298 646	44.42	16.13	86 241	2 386 200	51.52	20.42
三核苷酸	55 284	1 195 623	38.18	13.86	103 517	4 275 309	61.84	24.51
四核苷酸	120 262	3 101 836	83.05	30.15	90 822	2 861 504	54.26	21.50
五核苷酸	25 677	949 820	17.73	6.44	26 263	1 015 665	15.69	6.22
六核苷酸	5 881	236 340	4.06	1.47	3 196	119 610	1.91	0.76
总计	398 860	8 550 741	275.46	100	422 364	12 243 226	252.33	100
基因组大小/bp	1 447 999 364				1 673 876 332			

注：　丰度(个/Mb)=某类型微卫星数量/基因组大小；　比例(%)=某类型微卫星数量/微卫星总数。

Notes: Abundance (no./Mb)=Number of certain repeat microsatellites/genome size; Percentage (%)=Number of certain repeat microsatellites/total number of microsatellites.

表2　红尾蚺和原矛头蝮基因组中最丰富的10种微卫星类型

Table 2　Ten most abundant microsatellite repeats in the genomes of *Boa constrictor* and *Protobothrops mucrosquamatus*

红尾蚺		原矛头蝮	
Boa constrictor		*Protobothrops mucrosquamatus*	
类型	个数	类型	个数
A	113 242	A	83 531
AC	37 281	AAT	53 161
AAAT	36 774	AC	42 878
AG	17 121	C	28 794
AAT	15 617	AAAT	27 465
AAAC	14 588	AG	27 093
C	14 196	AGG	19 502
AT	9 782	AAGG	13 229
AAC	9 604	AAC	9 021
ATG	8 821	ATG	7 918

表 3　红尾蚺和原矛头蝮基因组中 6 种重复类型中最常见的 4 种微卫星

Table 3　The 4 common microsatellite repeats of 6 different repeat types in the genomes of *Boa constrictor* and *Protobothrops mucrosquamatus*

类型	红尾蚺 *Boa constrictor*	原矛头蝮 *Protobothrops mucrosquamatus*
单碱基重复	A (113 242) 88.86%	A (83 531) 74.37%
	C (14 196) 11.14%	C (28 794) 25.63%
二碱基重复	AC (37 281) 57.96%	AC (42 878) 49.72%
	AG (17 121) 26.62%	AG (27 093) 31.42%
	AT (9 782) 15.21%	AT (16 083) 18.65%
	CG (134) 0.21%	CG (187) 0.22%
三碱基重复	AAT (15 617) 28.25%	AAT (53 161) 51.35%
	AAC (9 604) 17.37%	AGG (19 502) 18.84%
	ATG (8 821) 15.96%	AAC (9 021) 8.71%
	AGG (8 252) 14.93%	ATG (7 918) 7.65%
四碱基重复	AAAT (36 774) 30.58%	AAAT (27 465) 30.24%
	AAAC (14 588) 12.13%	AAGG (13 229) 14.57%
	AAGG (8 693) 7.23%	AGGG (6 410) 7.06%
	AATG (7 882) 6.55%	AAAC (6 240) 6.87%
五碱基重复	AAAAT (4 473) 17.42%	AATAG (10 540) 40.13%
	AAAAC (2 763) 10.76%	AAAAT (2 050) 7.81%
	AAGGG (1 477) 5.75%	AAAAC (1 207) 4.60%
	AAATT (1 435) 5.59%	AAGAT (984) 3.75%
六碱基重复	ACATAT (799) 13.59%	ACATAT (286) 8.95%
	AACCCT (677) 11.51%	AAGGAG (204) 6.38%
	AAATAT (382) 6.50%	AACCCT (177) 5.54%
	AAGGAG (305) 5.19%	ATATAG (148) 4.63%

注： 括号内数字表示该重复类型在基因组中出现的次数， 后面的百分数表示该重复类型占所有所在重复类型的百分比。

Notes: number in parenthesis indicates occurrence number of the repeat type and the following percentage means the percentage of the repeat type in all repeat types.

表4　红尾蚺、原矛头蝮、人和鼠基因组不同区域微卫星的数量和丰富

Table 4　The number and abundance of microsatellites in different genomic regions of *Boa constrictor*, *Protobothrops mucrosquamatus*, *Homo sapiens* and *Mus musculus*

物种	基因区				基因间区
	编码区	非翻译区	外显子	内含子	
红尾蚺 *Boa constrictor*	1 638 (51.60)	1 225 (213.17)	2 863 (76.36)	111 031 (259.45)	284 927 (290.09)
原矛头蝮 *Protobothrops mucrosquamatus*	1 512 (48.77)	1 187 (152.92)	2 699 (69.63)	120 432 (243.49)	299 174 (262.73)
人类 *Homo sapiens*	1 794 (47.76)	8 210 (252.68)	10 004 (142.81)	426 480 (374.05)	584 031 (302.48)
小鼠 *Mus musculus*	1 558 (41.02)	8 915 (296.99)	10 473 (154.02)	358 091 (383.36)	592 071 (339.30)

注： 括号内数字表示微卫星的丰度， 单位为个/Mbp。

Note: number in parenthesis indicates the abundance of microsatellites, the unit is no./Mb.

2.2 基因中微卫星密度分布的位置特征

为了研究微卫星在基因以及基因上下游的密度分布的位置特征，参考方法 1.5 将微卫星定位到基因的不同区域，然后计算每个元件中微卫星的密度。红尾蚺基因组中有 1 552 个 CDS 含有有微卫星，其中只含有 1 个、2 个、3 个和 4 个微卫星的 CDS 分别有 1 480 个、61 个、8 个和 3 个。原矛头蝮中有 1 397 个 CDS 含有有微卫星，其中只含有 1 个、2 个、3 个、4 个和 5 个微卫星的 CDS 分别有 1 308 个、69 个、15 个、4 个和 1 个。红尾蚺中含有 4 个 SSR 的 CDS 有 3 个，分别来源于基因 ZFP36L2、H1C1、JUND。原矛头蝮中含有 5 个 SSR 的 CDS 来源于基因 WNK2，含有 4 个 SSR 的 CDS 分别来源于基因 PRDM2、H1C1、LOC107297696、SKOR2。计算每个区域微卫星的密度，我们发现红尾蚺和原矛头蝮基因中微卫星密度分布相似。红尾蚺基因上游 500 bp、外显子、内含子和基因下游 500 bp 各个区域微卫星的密度分别为 318.40 个/Mb、83.41 个/Mb、255.15 个/Mb 和 320.79 个/Mb。原矛头蝮基因上游 500 bp、外显子、内含子和基因下游 500 bp 各个区域微卫星密度分别为 392.34 个/Mb、70.17 个/Mb、242.66 个/Mb 和 380.36 个/Mb。在转录起始位点(TSS)附近微卫星密度最高，而且越靠近 TSS，微卫星密度越高。在基因上游 500 bp 和下游 500 bp 内微卫星呈现出对称密度分布，内含子微卫星密度比外显子微卫星密度高，而且在内含子区分布比较均匀，内含子 5’-和 3’-微卫星密度要比内含子内部区域密度高。我们也计算了人和小鼠基因及其上下游 SSR 的密度分布，人类基因上游 500 bp、外显子、内含子和基因下游 500 bp 各个区域微卫星密度分别为 307.79 个/Mb、71.92 个/Mb、379.03 个/Mb，324.40 个/Mb，小鼠基因上游 500 bp，外显子、内含子和基因下游 500 bp 各个区域微卫星密度分别为 388.44 个/Mb、115.23 个/Mb、391.02 个/Mb、386.06 个/Mb。人和小鼠内含子微卫星密度比外显子微卫星密度高，这与红尾蚺和原矛头蝮基因区 SSR 的密度分布类似。4 个物种的基因组中，基因的第一个外显子和最后一个外显子区域 SSR 的密度比内部外显子区域 SSR 的密度高。红尾蚺和原矛头蝮基因上下游 500 bp 内 SSR 的密度比内含子区域 SSR 的密度要高，而人和小鼠基因上下游 500 bp 内 SSR 的密度和内含子区域 SSR 的密度比较接近。

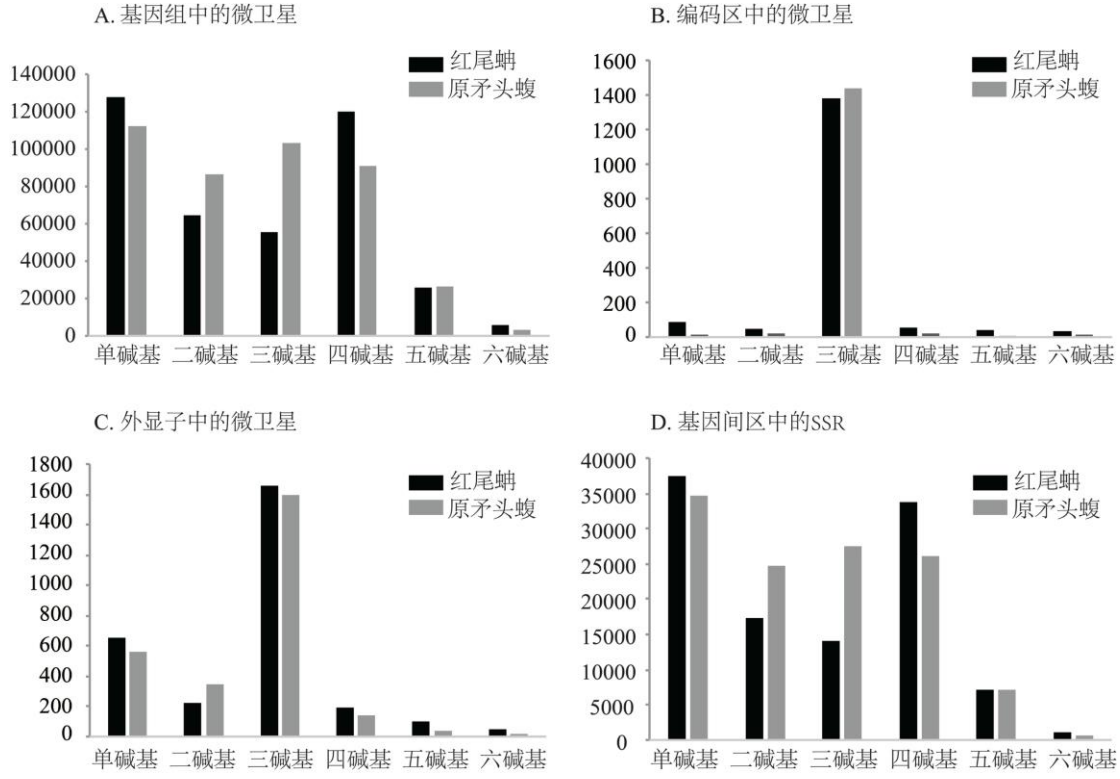

图 2　红尾蚺和原矛头蝮基因组不同区域的微卫星类型的分布

Fig. 2　The distribution of microsatellite types in different genomic regions of *Boa constrictor* and *Protobothrops mucrosquamatus*

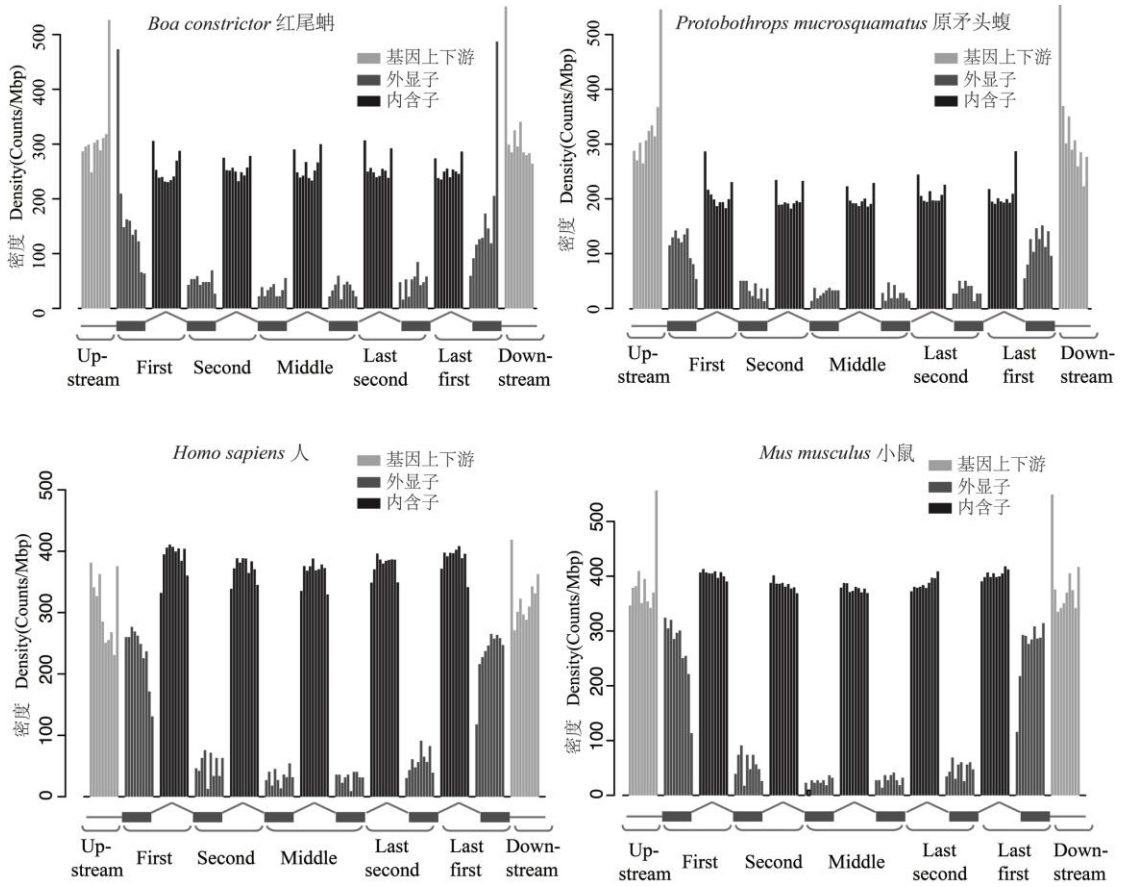

图 3　红尾蚺、原矛头蝮、人和鼠基因区及其上下游微卫星的密度分布

Fig. 3　The microsatellite density in gene regions and their upstream and downstream regions of *Boa constrictor*, *Protobothrops mucrosquamatus*, *Homo sapiens* and *Mus musculus*

2.3 含有微卫星的编码序列的功能分析

红尾蚺和原矛头蝮基因组中含有微卫星的 CDS 分别有 1 552 条和 1 397 条，分别来源于 1 431 个和 1 291 个蛋白编码基因。提取红尾蚺和原矛头蝮基因组中含有 SSR 的 CDS 序列，然后使用 BLASTx 比对到 NR 数据库，其中分别有 1 066 (68.69%)条和 1 047 (74.95%)条能比对到 NR 数据库，然后对结果进行 GO 注释，分别有 736 条和 773 条 CDS 能够被 GO 功能归类。红尾蚺含有微卫星的编码序列被分配到 3 142 个 GO 条目，原矛头蝮含有微卫星的编码序列被分配到 3 268 个 GO 条目。图 4 展示了红尾蚺和原矛头蝮含有微卫星的编码区的 GO 功能注释的比较。“Biology Process”本体中，“biological regulation”和“cellular process”分配的 CDS 数量最多；“Cellular Component”本体中，“organelle”、“cell part”和“cell”分配的 CDS 数量最多；“Molecular Function”本体中，“binding”和“catalytic activity”分配的 CDS 数量最多。红尾蚺和原矛头蝮基因组中分配到“biological regulation” (GO:0065007)条目的 CDS 序列最多，分别有 185 个和 175 个，占各自总数的 25.14%和 22.64%。相比之下，人和小鼠基因组中含有微卫星的 CDS 分别有 1 644 条和 1 458 条，分别来源于 1 443 个和 1 331 个编码基因，其中分别有 1 320 条和 1 155 条 CDS 能比对到 NR 数据库，分别有 1 116 条和 954 条 CDS 能够被 GO 功能归类。人和小鼠基因组中分配到“biological regulation” (GO:0065007)条目的 CDS 序列也最多，分别有 321 个和 251 个，占各自总数的 28.76%和 26.31%。总体上来看，红尾蚺和原矛头蝮基因组含有 SSR 的 CDS 序列的功能归类相似，与人和小鼠相比存在一定差异。

对红尾蚺、原矛头蝮、人和小鼠 4 个物种含有微卫星的 CDS 序列使用 OrthoMCL 进行直系同源分析，发现一共可以归类到 494 个基因家族，其中红尾蚺和原矛头蝮含有微卫星的 CDS 可以归类到 263 个基因家族，人和小鼠含有微卫星的 CDS 可以归类到 328 个基因家族，并且只有 3 个基因家族在这 4 个物种之间共享。共享的 3 个基因家族分别为 ONECUT2 (one cut homeobox 2)基因家族、LOC107401594(cyclin-dependent kinase 8)基因家族和 HOXD8 (homeobox D8)基因家族。红尾蚺和原矛头蝮含有微卫星的 CDS 相比较，两者共享 155 个基因家族，红尾蚺特有的基因家族有 42 个，原矛头蝮特有的基因家族有 66 个。人和小鼠含有微卫星的 CDS 相比较，两者共享 141 个基因家族，

人特有的基因家族有 97 个，小鼠特有的基因家族有 90 个。两个蛇类物种和两个哺乳类物种含有微卫星的 CDS 相比较，两者共享的基因家族有 97 个，蛇类特有的基因家族有 166 个，哺乳类特有的基因家族有 231 个。

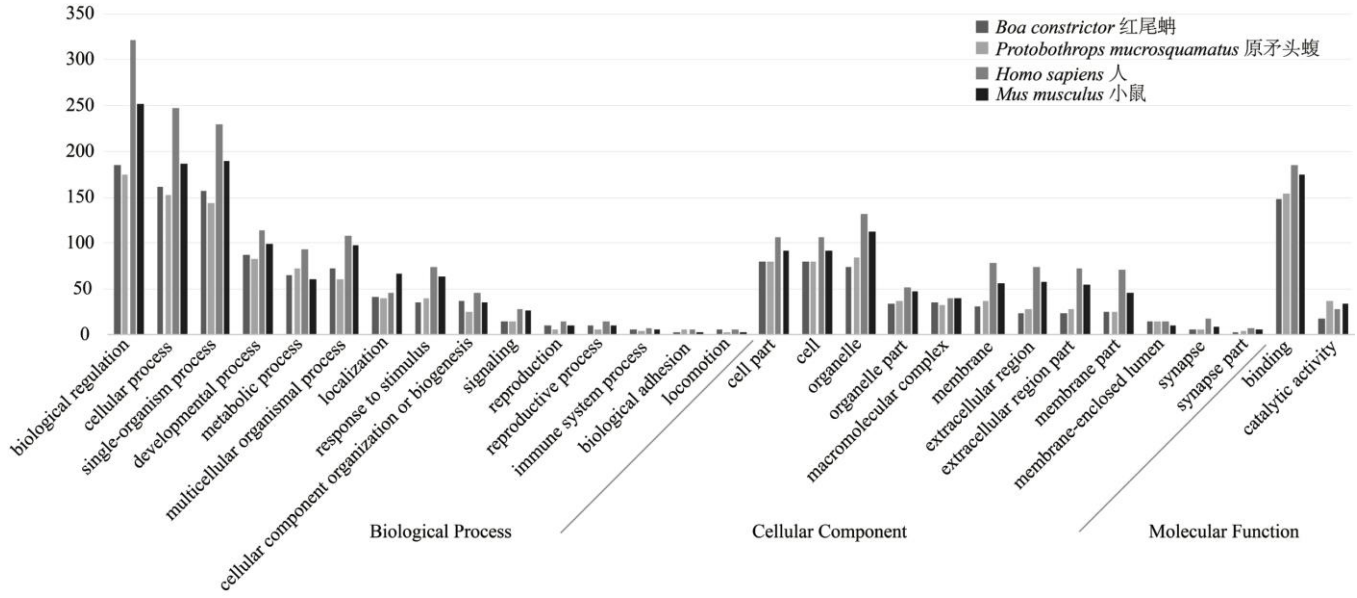

图4　红尾蚺、原矛头蝮、人和小鼠基因组中含有微卫星的编码区的GO功能归类

Fig. 4　GO classifications of coding sequences with microsatellites in the genomes of *Boa constrictor*, *Protobothrops mucrosquamatus*, *Homo sapiens and Mus musculus*

3 讨论

本研究从红尾蚺(基因组1.48 Gbp，Contig N50为47 kb)和原矛头蝮全基因组(基因组大小为1.67G，Contig N50为21 kb)中分别鉴定出398 860和422 364个SSR位点，数量的差异可能和基因组的大小、组装质量和物种基因组的特异性有关系。Wang等(2016)从亚利桑那州树皮蝎*Centruroides exilicauda*和马氏正钳蝎*Mesobuthus martensii*的全基因组中分别鉴定出114 026个和211 868个SSR，而亚利桑那州树皮蝎的基因组为926 Mb (Contig N50为5 kb)，马氏正钳蝎的基因组大小为925 Mb (Contig N50为45 kb)，两者基因组大小相近，而且是近源物种，但是鉴定出来的SSR数量差距很大，说明了测序的质量对基因组中SSR的识别有较大的影响。我们选取已测序蛇类物种中Contig N50最高的2个物种来做分析，主要是为了更加全面的鉴定出全基因组中的SSR。红尾蚺和原矛头蝮基因组中SSR的含量比较相近，分别占基因组大小的0.59%和0.73%，与大型哺乳动物大熊猫*Ailuropoda melanoleuca* (0.64%)和北极熊*Ursus maritimus*(0.79%)比较相似(李午佼等，2014)。红尾蚺和原矛头蝮基因组中SSR的丰度分别为275.46个/Mb和252.33个/Mb，与大熊猫(371.8个/Mb)、北极熊(405.6个/Mb)相比偏低(李午佼等，2014)，与人类(315.93个/Mb)、小鼠(342.68)相比也偏低，这是否暗示了蛇类物种基因组中SSR的丰度比哺乳类物种基因组SSR的丰度低有待进一步确认。

红尾蚺基因组中6种重复类型SSR所占的比例排序与原矛头蝮基因组6种重复类型SSR所占的比例排序不一致，并且最丰富的前5种微卫星也不一致(表1，表2)。研究发现大熊猫和北极熊基因组中6种重复类型的微卫星的比例排序和最丰富的前5种微卫星都表现出一致性(李午佼等，2014)。大熊猫和北极熊都是熊科Ursidae动物，相比之下，红尾蚺属于蚺科而原矛头蝮属于蝰科。说明了基因组中微卫星组成特征差异在一定程度上可以反映物种间的亲缘关系。

蛇亚目的红尾蚺和原矛头蝮、哺乳纲灵长目的人类和小鼠(本研究)以及哺乳纲食肉目的大熊猫和北极熊(李午佼等，2014)、节肢动物门的亚利桑那州树皮蝎和马氏正钳蝎(Wang *et al.*，2016)都是以单碱基数量最多。但在一些其他物种的基因组中，如中国对虾*Fenneropenaeus chinensis*(高焕等，2004)、蜜蜂*Apis mellifera*(魏朝明等，2007)等出现了二碱基为主要重复类型的情况，而酿酒酵母*Saccharomyces cerevisiae* (Katti *et al.*，2001)、粗糙脉孢菌*Neurospora crassa*(李成云等，2004)等基因组中占主导地位的SSR重复类型是三碱基，说明不同物种中不同重复类型的丰度差异较大。不同物种的优势SSR的重复类型不一样，这反映了不同物种基因组特征。有研究认为$(A)_n$类型微卫星的高频出现是由高密度散在分布的逆转录转座子，如Alu和LINE，以及经加工的假基因的Poly A尾所产生的(Tóth *et al.*，2000)。红尾蚺中重复序列含量最高的类型为LINE，占基因组的13.03%(Yin *et al.*，2016)，与这一推测相符合。

我们比较了红尾蚺、原矛头蝮、人和小鼠4个物种中编码区，外显子区，内含子区和基因间区中SSR的数量和丰度。4个物种中，非翻译区(包括5’UTR和3‘UTR)SSR的丰度都比编码区要高很多，说明了SSR在非翻译区聚集，推测其可能影响基因的转录活性。红尾蚺和原矛头蝮2个蛇类物种与人类和小鼠2个哺乳类物种相比，编码区SSR的数量和丰度相差很小，而在基因的内含子、外显子和基因间区相差很大(表4)。这表明了蛇类与哺乳类相比基因中编码区微卫星的数量和丰度相差较小，可能是因为编码序列在不同物种中比较保守，受到的选择压力大。

微卫星对扩张和收缩非常敏感，编码区单个单碱基重复、二碱基重复、四碱基重复和五碱基重复单元的插入或缺失都会导致移码。我们发现红尾蚺基因编码区中三碱基重复类型SSR占编码区SSR总数的84.07%，红尾蚺基因编码区中三碱基重复类型SSR占编码区SSR总数的95.11%，在编码区6种重复类型的SSR中占有绝对优势。原矛头蝮基因编码区三碱基类型的SSR的比例比红尾蚺基因编码区三碱基类型的SSR的比例要高，可能的解释是红尾蚺是一种比原矛头蝮更古老的蛇类(Reyes-Velasco *et al.*，2015)，单碱基、二碱基、四碱基和五碱基类型SSR在进化的过程中发生插入或缺失突变，导致蛋白功能的改变，从而很可能在进化的过程中被淘汰；另一种可能的解释是编码区三碱基微卫星的增加可以增加性状的多样性，有利于物种在进化过程中的适应性改变，从而在进化的过程中被保留。有研究表明在对人、大猩猩、红毛猩猩、猕猴4个高等哺乳动物中SSR的比较分析，发现编码区6种重复类型的SSR的进化速度超过非编码区6种重复类型的SSR的进化速度的两倍多(Loire *et al.*，2013)。由此可见，编码区SSR所受到的选择压力要比非编码区大，进化的速度更快。

红尾蚺和原矛头蝮两种蛇类物种基因区微卫星的密度分布位置特征相似，并且与人和小鼠2种哺乳动物中微卫星的密度分布的位置特征也相似，都呈现出在基因上下游500 bp密度最高，在内含子区SSR的密度次之，而外显子区SSR的密度最低(图3，图4)。有研究报道拟南芥和水稻2种植物中SSR的密度沿着基因区5’到3’方向呈现出递减的趋势(Fujimori *et al.*，2003)。说明了动物和植物基因区SSR的密度分布特征存在差异。有研究对42个已经测序的原核生物的基因组编码区SSR的差异和密度进行分析，发现编码区SSR的密度呈现出一个“U型”分布，即基因左右末端SSR的密度较高，中间区域SSR的密度较低(Lin & Kussell，2012)。这说明了真核生物和原核生物基因区SSR的密度分布存在差异。

对红尾蚺和原矛头蝮基因组含有了微卫星的编码序列进行 GO 注释分析，可以看出这 2 个物种含有微卫星的编码区注释出的功能分类基本一致(图 4)，但是与人和小鼠 2 种哺乳动物以及 2 种蝎子的结果(Wang *et al.*，2016)相比差异较大。对红尾蚺、原矛头蝮、人和小鼠 4 个物种中包含微卫星的 CDS 进行直系同源分析，发现只有 3 个基因家族被这 4 个物种共享，2 个蛇类物种之间共享的基因家族比各自特有的基因家族多，2 个哺乳类物种之间共享的基因家族也比各自特有的基因家族多。这说明了含有微卫星的编码序列的功能在不同门类间存在种系差异。微卫星的收缩或扩张为物种适应性进化过程中的遗传变异提供了丰富的原材料(Kashi & King，2006)。对群体之间、近源物种之间、种系之间基因组层面微卫星的挖掘和比较分析，将有助于更进一步了解微卫星在基因组中的功能。

参考文献：

高焕, 刘萍, 孟宪红, 等. 2004. 中国对虾(*Fenneropenaeus chinensis*)基因组微卫星特征分析[J]. 海洋与湖沼, 35(5): 249-254.

李成云, 李进斌, 周晓罡, 等. 2004. 粗糙脉孢菌基因组中的微卫星序列的组成和分布[J]. 中国农业科学, 37(6): 851-858.

李午佼, 李玉芝, 杜联明, 等. 2014. 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, 33(6): 874-878.

魏朝明, 孔光耀, 廉振民, 等. 2007. 蜜蜂全基因组中微卫星的丰度及其分布[J]. 昆虫知识, 44(4): 501-504.

Conesa A, Götz S, García-Gómez JM, *et al*. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research[J]. Bioinformatics, 21(18): 3674-3676.

Fujimori S, Washio T, Higo K, *et al*. 2003. A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription[J]. FEBS Letters, 554(1): 17-22.

Jurka J, Pethiyagoda C. 1995. Simple repetitive DNA sequences from primates: compilation and analysis[J]. Journal of Molecular Evolution, 40(2): 120-126.

Kajitani R, Toshimoto K, Noguchi H, *et al.* 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads[J]. Genome Research, 24(8): 1384-1395.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution[J]. Trends in Genetics, 22(5): 253-259.

Katti MV, Ranjekar PK, Gupta VS. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences[J]. Molecular Biology and Evolution, 18(7): 1161-1167.

Kerkkamp HM, Kini RM, Pospelov AS, *et al*. 2016. Snake genome sequencing: results and future prospects[J]. Toxins, 8(12): 360-375.

Laurie JV, Janalee PC. 2009. Herpetology: an introduction biology of amphibians and reptiles (third edition)[M]. London: Academic Press: 551-578.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes[J]. Genome Research, 13(9): 2178-2189.

Li YC, Korol AB, Fahima T, *et al*. 2004. Microsatellites within genes: structure, function, and evolution[J]. Molecular Biology and Evolution, 21(6): 991-1007.

Lin WH, Kussell E. 2012. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions[J]. Nucleic Acids Research, 40(6): 2399-2413.

Loire E, Higuet D, Netter P, *et al.* 2013. Evolution of coding microsatellites in primate genomes[J]. Genome Biology and Evolution, 5(2): 283-295.

Reyes-Velasco J, Card DC, Andrew AL, *et al.* 2015. Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom[J]. Molecular Biology and Evolution[J], 32(1): 173-183.

Thiel T, Michalek W, Varshney R, *et al*. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.)[J]. Theoretical and Applied Genetics, 106(3): 411-422.

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

Wang C, Kubiak L, Du L, *et al*. 2016. Comparison of microsatellite distribution in genomes of *Centruroides exilicauda* and *Mesobuthus martensii*[J]. Gene, 594(1): 41-46.

Yin W, Wang Z, Li Q, *et al*. 2016. Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper[J]. Nature Communications, 13107(7): 1-11.