Abundance and Distribution of Microsatellites in The Entire Mosquito Genome*

YU Quan-You[1)**], LI Bin[1)** ***], LI Guan-Rong[2)], FANG Shou-Min[1)], YAN Hong[1)], TONG Xiao-Ling[1)], QIAN Ji-Feng[1)], XIA Qing-You[1)], LU Cheng[1)***]

([1)]*The Key Sericultural Laboratory of the Agricultural Ministry of China, College of Sericulture and Biotechnology, Southwest Agricultural University, Chongqing* 400716, China;

[2)]*College of Agronomy and Life Sciences, Southwest Agricultural University, Chongqing* 400716, China)

Abstract Microsatellite is a genetic marker, explored recently. In order to improve related studies on genetics of *Anopheles gambiae*, simple sequence repeats of the entire mosquito genome with 1~6 bp nucleotide motifs were analyzed. Abundance and distribution of microsatellites across the *A.gambiae* genome were analyzed and compared between various (exons, introns and intergenic) regions of all the chromosomes. About 2.14% of the mosquito genome was occupied by SSRs. Chromosome X had the maximum density of SSRs. Abundance of A repeats was similar to C repeats. AC was a little more than two times as abundant much as AG. However, AT and CG repeats were rare. For tri- and tetramer repeats, AGC, AAAC and AAAT predominated while ACG, ACT, AGG, CCG, ATGC, CCCG, ACTG, AACT, ACGT, AGAT, CCGG, ACCT and AGCT were rare. For some pentamer repeats, one was completely absent on a certain chromosome, even on several chromosomes. SSRs in exons of all chromosomes were less abundant than in introns and intergenic regions except for mono- and dimer repeats in exons of chromosome 2L. Abundance and distribution of SSRs on the two arms of each chromosome showed much in common.

Key words microsatellites, *Anopheles gambiae*, genome, abundance, distribution

Microsatellites or simple sequence repeats (SSRs) are tandemly arranged repeats of short DNA motifs (1~6 bp in length) and occur in eukaryotic and prokaryotic genomes, even in the smallest bacterial genomes [1,2]. SSRs are distributed throughout the genome, including coding and non-coding regions [3]. Genome-wide abundance and distribution of 1~6 bp repeat motifs were analyzed in *Escherichia coli* [4], several eukaryotic taxa [3], human [5], etc. Generally, SSRs in exons are no more abundant than in non-coding regions [5,6]. However, the disease-associated triplet repeats were mainly found in coding regions of the human genome [7]. Some SSR loci, especially trinucleotide repeats, are associated with neurological diseases[8, 9]. Dinucleotide repeats can also cause human diseases such as Norrie's disease [10]. In either coding or non-coding sequence, the expansion of unstable trinucleotide repeats might result in these neurological diseases[11,12].

Malaria infects 500 million people in the world and causes 2.7 million deaths annually, which, however, most of them occur in sub-Saharan Africa[13]. As such, the local economy is crippled. *Anopheles gambiae* is the most efficient vector of malaria parasite. People across the world have been fighting against malaria. In 2002, the entire *A.gambiae* genome was sequenced [14]. Population genetic structure of the mosquito was studied with microsatellites[15,16]. Vector population genetic studies are important tools for: predicting the spread of genes such as those that can confer resistance to insecticide or refractoriness to pathogens [17]. Ranson H *et al.* [18] used microsatellite markers to identify two quantitative trait loci (QTLs) affecting DDT resistance in the malaria vector *A.gambiae*. Although some methods to control malaria are available, unfortunately, the burden of malaria is on the increase globally [19]. But it is a truly effective method that we apply SSR to explore new control methods of malaria.

By analyzing distribution and abundance of SSR in mosquito, it can provide an outline for detailed research of SSRs. Comparing with other species,

*This work was supported by a grant from The National Natural Science Foundation of China (30300262).

**These authors are contributed equally to this paper.

***Corresponding author.

LI Bin, Tel: 86-23-68250793, E-mail: libinlb@hotmail.com

LU Cheng, Tel: 86-23-68250346, Fax: 86-23-68251123, E-mail: lucheng@swau.cq.cn

Received: November 30, 2004 Accepted: January 31, 2005

similarity and difference of distribution and abundance of SSR would be obtained. The comparing analysis helps to understand the evolution and function of SSR. For *Anophelines* mosquito, distribution and abundance of microsatellites were studied partially [20,21]. But, on the entire genome level, such analysis in mosquito was not yet conducted. In this paper, overall SSRs of motifs of 1~6 nucleotides in the whole mosquito genome were screened to analyze their distribution and abundance on different chromosome and in its various regions.

1 Materials and methods

1.1 Mosquito genomic sequence

The entire mosquito genomic sequence was downloaded from http://www.ensembl.org/Anopheles_gambiae/, including chromosome arms (2L, 2R, 3L and 3R) and chromosome X and unknown. Ensembl shows the unassigned scaffolds, concatenated together in arbitrary order, as an artificial unknown "chromosome" (UNKN).

1.2 Analysis of dataset

The *A.gambiae* gene structure was annotated [14]. We developed a program in Perl 5 to extract exons, introns and intergenic regions from each chromosome which means 2L, 2R, 3L, 3R, X or UNKN in the present analysis. The SSRfinder program, with a series of Perl scripts, was obtained from a website (http://www.maizemap.org/) and modified. Using modified SSRfinder, the various regions were scanned in order to search for all perfect SSRs with motifs more than 12 bp. The results were reported in an out file in the form of motif, locus number and genomic locus. Another Perl 5 program was used to classify different repeat types. The SSR data presented here includes both strands of the DNA sequence. For instance, ACG also includes CGA, GAC and the reverse complements TCG, CGT and GTC. Theoretically, there are 501 repeat types with length of 1~6 bp.

Densities of each repeat type and class on a given chromosome and in its various regions were calculated. The density is defined as the ratio of the repeats length (bp) to chromosome length (Mb) or length of regions (Mb). Those data and the total repeat number of individual SSR locus in the mosquito genome was presented in the additional data file (http://www.swau.cq.cn/ssr/additional.data.file.xls) of this paper. Based on the data set, abundance of SSRs in the corresponding regions (exons, introns and intergenic regions) of different chromosomes and in different regions on a given chromosome was compared.

2 Results

2.1 Abundance of overall SSRs in the mosquito genome and on each chromosome

About 2.14% in the entire mosquito genome was occupied by SSRs, of which 0.11%, 0.59%, 0.37%, 0.18%, 0.05% and 0.85% occupied by mono- to hexamer repeats, respectively. Chromosome X contained the maximum density of SSR (35 142.75 bp/Mp), followed by 2L (24 195.38 bp/Mp) (Figure 1). It was interesting to note that 2R, 3L and 3R had nearly equal SSR density, though discrepancies in their length of DNA sequence were very notable.

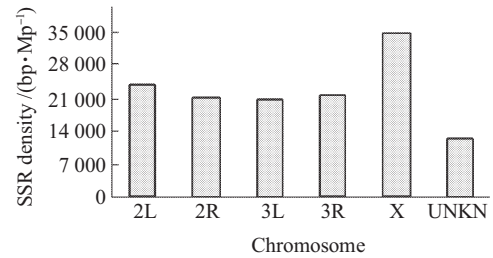

Fig.1 Overall SSR density in each chromosome of the mosquito

2.2 Abundance and distribution of each SSR class

2.2.1 Mononucleotide repeats. Mononucleotide repeat densities in exons of all the chromosomes were lower than those in intronic and intergenic regions, except for 2L which had the highest density (1 453.13 bp/Mp), while 3L had the lowest density of 88.17 bp/Mp (Figure 2a, see additional data file). Compared to other chromosomes, introns and intergenic regions of chromosome X showed the maximum density (3 343.63 bp/Mp and 2 808.79 bp/Mp, respectively). It was interesting that two arms of Chromosome 2 or 3 had nearly equal SSR densities in intergenic regions (Figure 2a).

The two monomer repeats poly A (or T) and poly C (or G) were comparable in abundance on 2L, 2R and 3R, while poly A (or T) was a slightly more abundant than poly C (or G) on the other chromosomes. Chromosome X had the highest repeat densities of poly A (or T) and poly C (or G) (Figure 3). For a given repeat, the number of repeats and length in different genomic regions of all chromosomes are included in the additional data file.

2.2.2 Dinucleotide repeats. Dinucleotide repeat density in introns was a somewhat higher than that in intergenic regions among all the chromosomes except

3R (Figure 2b). Like the monomer repeats, chromosome X also had the maximum dinucleotide repeat density in introns and intergenic regions among all the chromosomes. In exons, 2L contained the highest dimer repeat density (8 056.53 bp/Mp), while 3L had the lowest dimer repeat density (31.54 bp/Mp) (see additional data file). Exons showed lower density of dinucleotide repeats than introns and intergenic regions, with the exception of 2 L.

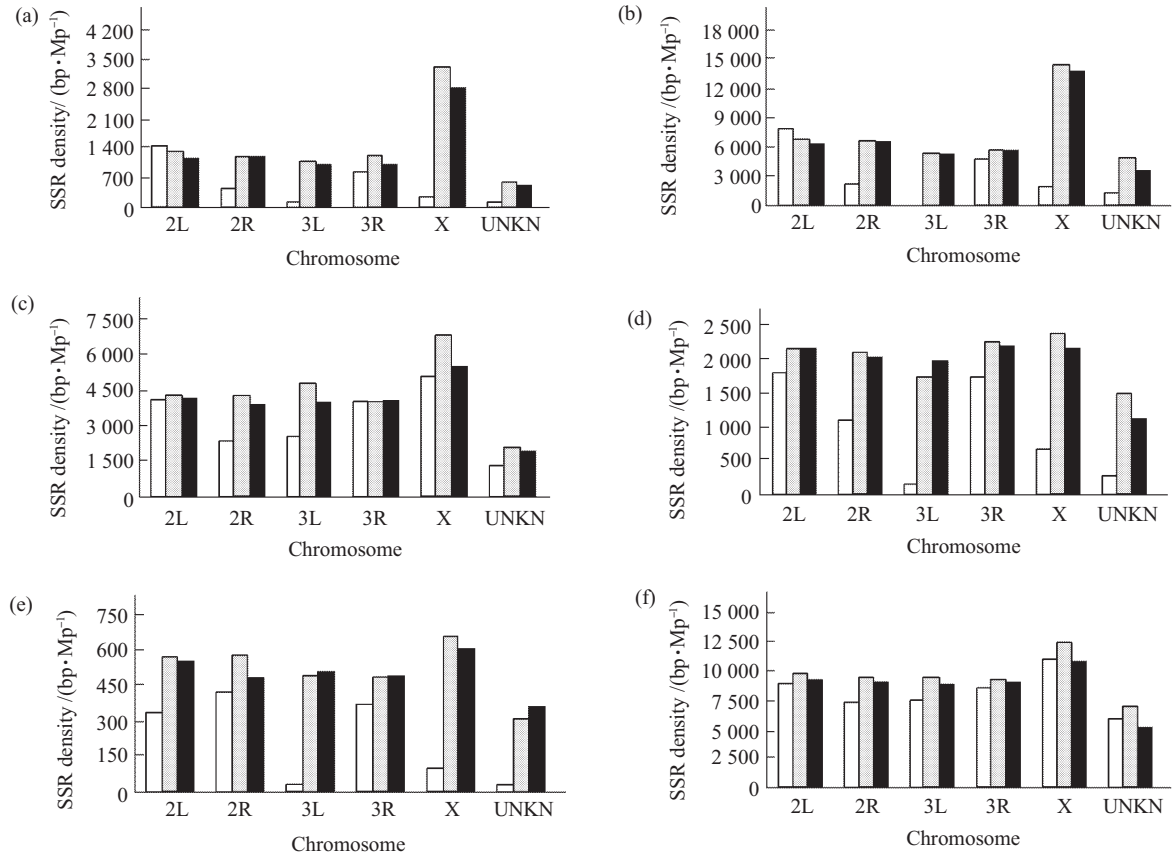

Fig.2 SSR density in exonic, intronic and intergenic regions of each chromosome of the mosquito

(a) monomers; (b) dimers; (c) trimers; (d) tetramers; (e) pentamers; (f) hexamers. □: exons, ▨ introns, ■: intergenic regions.

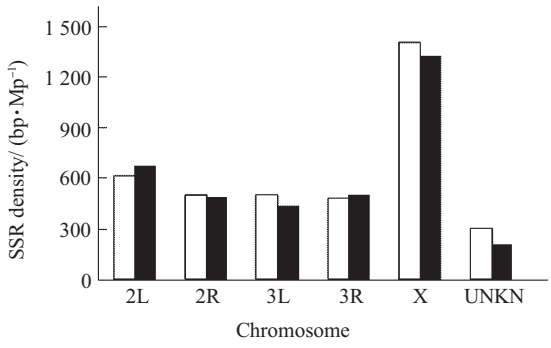

Fig.3 Density of each monomer repeat on different chromosomes

□: A, ■: C.

Among the dimer repeats, AC and AG were predominant across all chromosomes and the former density was over two times higher than the latter, whereas CG repeats were the rarest (Figure 4). Chromosome X contained the maximum density of AC (8 385.10 bp/Mp) and AG (3 988.19 bp/Mp), followed by 2 L contained AC (4 893.97.36 bp/Mp) and AG (1 986.15 bp/Mp), respectively (see additional data file).

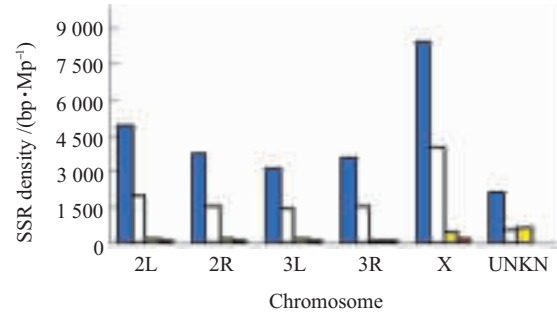

Fig.4 Density of each dimeric repeat across all chromosomes

Blue: AC, white: AG, yellow: AT, red: CG.

2.2.3 Trinucleotide repeats. Trimer repeats occurred abundantly in exons, while the density of trimer repeats in exons in all chromosomes were lower than that in non-coding regions. Except for 3R, the density in exons was slightly higher than that in non-coding regions (Figure 2c). Chromosome X had the maximum densities in various regions than other chromosomes. Densities of trimer repeats in non-coding regions of 2L, 2R, 3L, 3R were comparable, however, it was interesting that the distribution of density of trimer repeats on 2R and 3L was very similar (Figure 2c).

AGC was the predominant repeat on all chromosomes, except chromosome unknown. ACG, ACT, AGG and CCG were relatively rare on all chromosomes except for CCG on chromosome X and ACG on chromosome 2R and X (Figure 5). In chromosome unknown, AAT, AAG and AGC were more abundant than other trinucleotide repeats (see additional files).

2.2.4 Tetranucleotide repeats. All the chromosomes showed comparable tetranucleotide repeat densities in intronic and intergenic regions (Figure 2d). 2L and 3R had nearly equal densities of tetranucleotide repeats in exons, however, exons of 3 L contained the least density, 159.14 bp/Mp (see additional data file).

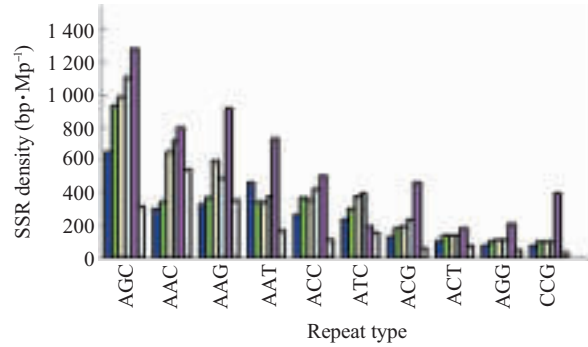

Fig.5 Density of different trimer repeats on each
■: 2L; ■: 2R; □: 3L; □: 3R; ■: X; □: UNKN.

Among the tetramer repeats, AAAC and AAAT were the most abundant across all chromosomes, in contrast, ATGC, CCCG, ACTG, AACT, ACGT, AGAT, CCGG, ACCT and AGCT repeats were apparently rare on all chromosomes with the exception of ACGT and AGCT on chromosome 2R (Figure 6). As for those with moderate densities, AAGT was an exception because its density on 3L, 3R and unknown chromosome was far higher than that on other chromosomes (Figure 6). In addition, AAAG, ACGC, AGCG, ACAT, ACCC, AAGT and AATG were also abundant on chromosome X (see additional files).

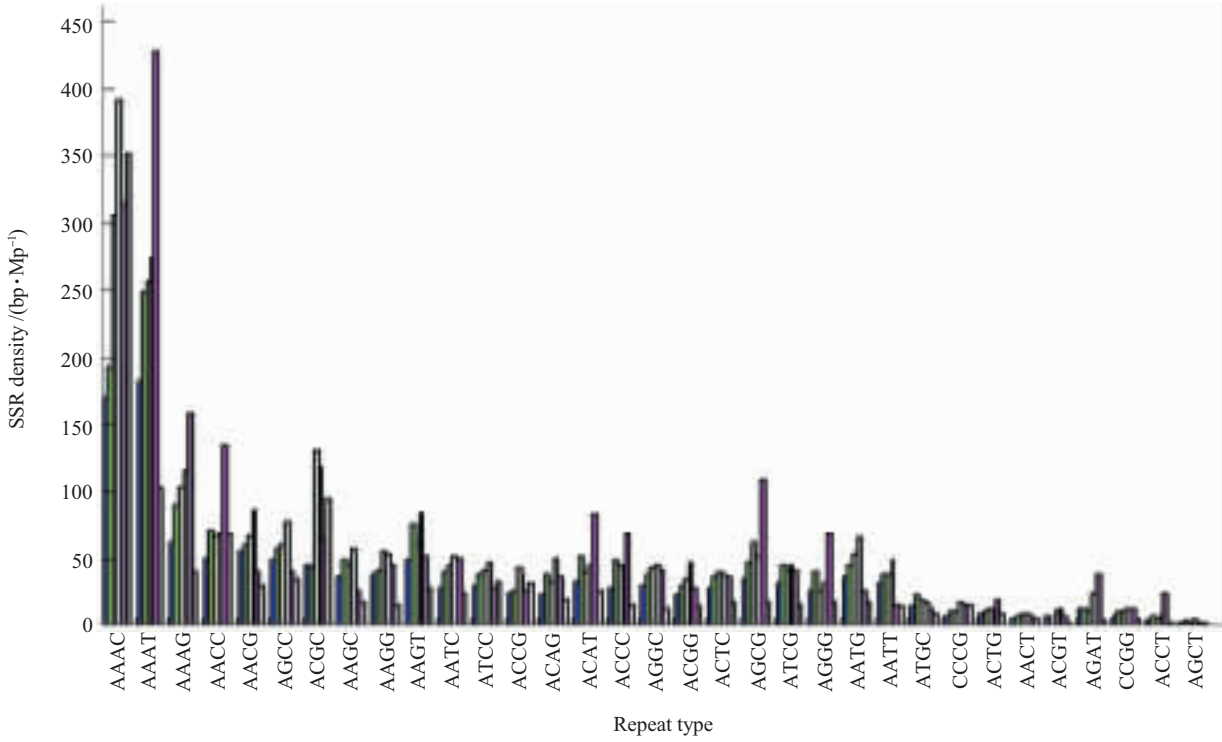

Fig.6 Density of each tetramer repeat across the entire chromosome set
■: 2L; ■: 2R; □: 3L; □: 3R; ■: X; □: UNKN.

2.2.5 Pentanucleotide repeats. Densities of pentanucleotide repeats in exons of 3L, X and unknown were very low, only 30.47 bp/Mp, 90.15 bp/Mp and 22.37bp/Mp, respectively (see additional files). AAAAC and AAAAT, which were the most abundant in the entire mosquito chromosome set, whereas most of other pentanucleotide repeats were rare (density on a given chromosome < 5 bp/Mp). In addition, for some pentamer repeats, one was completely absent on a certain chromosome, even on several chromosomes. CCGCG did not occur on 2L, 2R and 3L, for instance (see additional data file). The density of each pentanucleotide repeat and the total number of occurrences in various regions were given in additional data file.

2.2.6 Hexanucleotide repeats. Hexanucleotide repeat density in introns was slightly higher than that in exons and intergenic regions on each chromosome (Figure 2f). Interestingly, exons of chromosome 2L and 3R had almost equal densities of hexmer repeats, and exons of chromosome 2R and 3L was like this (Figure 2f). AAAAAC, AAAAAG, AAAAAT and AAACAC were most abundant on all chromosomes and their densities ranged between 111.20 bp/Mp and 549.51 bp/Mp, followed by AAAATT, AACAGC, AAAGAG and ACACGC with densities about 100 bp/Mp. For most of hexamer repeats, the densities were about 10 bp/Mp (see additional data file).

3 Discussion

The present analysis is a study of SSRs which were divided into six classes of repeats and 501 repeat types. The abundance and distribution of perfect SSRs over 12-bp long in whole genome of mosquito were analyzed, including the overall density of SSRs in each chromosome (Figure 1), the density of each class in various genomic regions of individual chromosome (Figure 2) and the density of each repeat type on each of the chromosomes (Figure 3~6).

About 2.14% of the mosquito genome was occupied by SSRs. This is the same as the previous conclusion, that SSR is abundant in the genomes of *Anophelines* mosquitoes[21,22]. SSR is not abundant in all mosquitoes, for microsatellites are rare in two *Aedes* species, *I. scapularis* and *Aedes aegypti*[20].

The relative abundance of six classes of repeats across all the chromosomes was analyzed (Figure 7). Mono-, tri-, tetra-, penta- and hexamer repeats on chromosome arms 2L, 2R, 3L, 3R had almost equal densities, respectively. Chromosome X had the highest densities for mono-, di-, tri- and hexamer repeats, with hexamer repeats being most abundant in the mosquito genome, followed by dimer repeats.

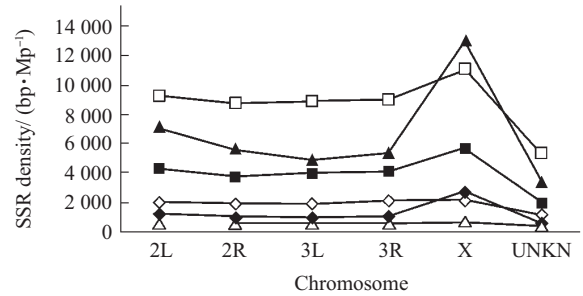

Fig.7 Distribution of microsatellites from mono- to hexamer repeats on each chromosome

◆—◆: monomers, ▲—▲: dimers, ■—■: trimers, ◇—◇: tetramers, △—△: pentamers, □—□: hexamers.

For chromosome arms 2L and 2R, the overall densities of SSRs were comparable (Figure 1). Chromosome arms 2L and 2R showed a more-or-less similar density of each class in non-coding regions (Figure 2), although different repeat types often showed tremendous variation in density in different genomic regions. The overall density of a given repeat type on chromosome arms 2L and 2R was very similar (see additional data file), however, the density of each class in exons of chromosome arms 2L and 2R show a big difference (Figure 2). In general, two arms of a given chromosome, distribution and abundance of SSRs had much in common.

The six classes of repeats were relatively rare in exons of chromosome X and 3L, except for tri- and hexamer repeats (Figure 2). Due to evolution and selection, SSRs were abundant in chromosome X and 3L, especially in its non-coding regions. The possibility remains that chromosome X and 3L played an important role in gene regulation, for SSRs are relevant to regulation of gene activity[23].

All of the tri- and hexanucleotide repeats were more abundant in exonic regions, compared to the ones in intronic and intergenic regions, on all chromosomes except Y[5], and dominated in exons of several taxa[3,24,25]. Although, tri- and hexanucleotide repeats were abundant in the mosquito, both of them were less abundant in exons than in introns and intergenic regions across the chromosome set, except for trimer repeats in 3R. However, of particular interest, the mono- and dinucleotide repeats were dominant in exons of 2L, and we have no valid the reason for this anomaly.

AC was the most common, with AAC, AAG and

AAT noted to be very abundant. AAAC and AAAT, among tetramer repeats, were predominant in the mosquito genome, with AAAAC, AAAAT and AAATG among pentamer repeats and AAAAAC, AAAAAG and AAAAAT in the case of hexamers. This result was the same as repeats with <50% of G+C are generally more abundant[3]. During the evolution of SSR loci, many loci changed into ones containing more A, so it is possible that these loci were relevant to some function, however, for further comprehension we must experiment in order to scientifically testify.

A and C repeats on all chromosomes showed in almost equal densities, while A repeats are predominant and C repeats are rare in many other species [3,5,26]. AGC of trinucleotide repeats were the most abundant in the mosquito, as in some mammalian an arthropoda[3,27], while AAT is the most abundant in the human species [5]. The mosquito had many pentamer repeats that did not occur in a certain chromosome, nor in several chromosomes, such as AGGCG, ACCGG, ACGCG and CCGCG. The repeats of other classes were found that the ones did not be absent on a special chromosome, although some of them may not occur in one or two of three regions of the chromosome (see additional data file).

Controversies over the microsatellites roles are never-ending, but there are some presumptions that microsatellites have many functions [23,27]. We all are familiar with some neurological diseases resulting from expansion of trimers[11,12]. What is the relationship of functions and evolutional mechanisms of SSRs? Why did abundance and distribution of SSRs have much in common in different species? There are many questions that need to be resolved and explored.

In short, distribution and abundance of SSRs in many species were analyzed. These analyses, particularly, the genome-wide analysis, help us find some laws of distribution of microsatellites and also to profoundly understand its evolution and function. The new mosquito control methods are urgently called for and desperately needed. With further comprehension and study of SSRs, perhaps there will be a scientific breakthrough.

Acknowledgements We wish to express our thanks to YE Chen, LI Guo-Qing, LI Rui-Qiang, LI Heng and TONG Wei for trimming, cleaning and managing the set of sequences.

References

1 Tautz D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acid Res, 1989, **17** (16): 6463~6471

2 Hancock J M. Simple sequences in a ‘minimal’ genome. Nat Genet, 1996, **14** (1):14~15

3 Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res, 2000, **10** (7):967~981

4 Gur-Arie R, Cohen C J, Eitan Y, *et al*. Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. Genome Res, 2000, **10** (1):62~71

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

6 Hancock J M. The contribution of slippage-like processes to genome evolution. J Mol Evol, 1995, **41** (6): 1038~1047

7 Nadir E, Margalit H, Gallily T, *et al*. Microsatellite spreading in the human genome: Evolutionary mechanisms and structural implications. Proc Natl Acad Sci USA, 1996, **93** (13): 6470~6475

8 Ashley C T, Warren S T. Trinucleotide repeat expansion and human disease. Annu Rev Genet, 1995, **29**: 703~728

9 Richards R I, Sutherland G R. Dynamic mutation: possible mechanisms and significance in human disease. Trends Biochem Sci, 1997, **22** (11): 432~436

10 Kenyon J R, Craig I W. Analysis of the 5′ regulatory region of the human Norrie's disease gene: evidence that a non-translated CT dinucleotide repeat in exon one has a role in controlling expression. Gene, 1999, **227** (2): 181~188

11 Rubinsztein D C. Trinucleotide expansion mutations cause diseases which do not conform to classical Mendelian expectations. In: Goldstein D B, Schlotterer C, eds. Microsatellites: Evolution and Applications, Oxford: Oxford University Press, 1999. 81~97

12 Masino L, Pastore A. A structural approach to trinucleotide expansion diseases. Brain Research Bull, 2001, **56** (3~4): 183~189

13 Breman J G, Egan A, Keusch G T. The intolerable burden of malaria: a new look at the numbers. Am J Trop Med Hyg, 2001, **64** (1-2 suppl.): iv~vii

14 Holt R A, Subramanian G M, Halpern A, *et al*. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science, 2002, **298** (5591):129~149

15 Ravel S, Herve J P, Diarrassouba S, *et al*. Microsatellite markers for population genetic studies in *Aedes aegypti* (Diptera: Culicidae) from Cote d'Ivoire: evidence for a microgeographic genetic differentiation of mosquitoes from Bouake. Acta Trop, 2002, **82** (1): 39~49

16 Nyanjom S R, Chen H, Gebre-Michael T, *et al*. Population genetic structure of *Anopheles arabiensis* mosquitoes in eEthiopia and eritrea. J Hered, 2003, **94** (6): 457~463

17 Toure Y T, Oduola A M, Morel C M. The *Anopheles gambiae* genome: next steps for malaria vector control. Trends Parasitol, 2004, **20** (3): 142~149

18 Ranson H, Jensen B, Wang X, *et al*. Genetic mapping of two loci affecting DDT resistance in the malaria vector *Anopheles gambiae*. Insect Mol Biol, 2000, **9** (5): 499~507

19 Maitland K, Bejon P, Newton C R. Malaria. Curr Opin Infect Dis, 2003, **16** (5): 389~395

20 Fagerberg A J, Fulton R E, Black W C. Microsatellite loci are not abundant in all arthropod genomes: analyses in the hard tick, Ixodes scapularis and the yellow fever mosquito, *Aedes aegypti*. Insect Mol Biol, 2001, **10** (3): 225~236

21 Rongnoparut P, Yaicharoen S, Sirichotpakorn N, *et al*. Microsatellite polymorphism in *Anopheles maculatus*, a malaria vector in Thailand. Am J Trop Med Hyg, 1996, **55** (6): 589~594

22 Sinkins S P, Hackett B J, Costantini C, *et al*. Isolation of polymorphic microsatellite loci from the malaria vector *Anopheles funestus*. Mol Ecol, 2000, **9** (4): 490~492

23 Li Y C, Korol A B, Fahima T, *et al*. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol, 2002, **11** (12): 2453~2465

24 Li B, Xia Q Y, Lu C, *et al*. Analysis on frequency and density of microsatellites in coding sequence of several eukaryotic genomes. Genomics Protemics & Bioinformatics, 2004, **2** (1): 24~31

25 Li B, Xia Q Y, Lu C, *et al*. Analysis of microsatellites derived from Bee ESTs. Acta Genetica Sinica, 2004, **31** (10): 1089~1094

26 Katti M V, Ranjekar P K, Gupta V S. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol, 2001, **18** (7):1161~1167

27 Li Y C, Korol A B, Fahima T, *et al*. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol, 2004, **21** (6): 991~1007

蚊子全基因组中微卫星的丰度及其分布 *

余泉友[1)**] 李 斌[1)** ***] 李关荣[2)] 房守敏[1)] 颜 虹[1)]
童晓玲[1)] 钱吉凤[1)] 夏庆友[1)] 鲁 成[1)***]

([1)]西南农业大学蚕学与生物技术学院，蚕桑学重点实验室，重庆 400716；
[2)]西南农业大学农学与生命科学学院，重庆 400716)

摘要 微卫星是近年大力开发的一种遗传标记，为推进按蚊遗传学相关研究，对按蚊全基因组中由 1~6 个碱基重复单元组成的简单序列重复 (微卫星) 进行了分析. 进而对其微卫星的丰度和分布进行了比较分析，也比较了染色体各个区域 (外显子、内含子和基因间隔区) 之间的分布差异. 微卫星在按蚊基因组中的比例约占 2.14%，其中 X 染色体拥有微卫星的密度最大. 对按蚊基因组中微卫星丰度而言，A 碱基和 C 碱基重复在基因组中丰度相似，AC 单元的丰度是 AG 单元的两倍多，然而 AT 和 CG 单元非常稀少；对于三四碱基而言，AGC，AAAC 和 AAAT 单元最为丰富，ACG，ACT，AGG，CCG，ATGC，CCCG，ACTG，AACT，ACGT，AGAT，CCGG，ACCT 和 AGCT 单元等均很稀少，而一些五碱基重复，在某条甚至某几条染色体中均未分布. 除两碱基重复单元在 2L 的外显子区域丰度较高外，其他重复单元均在内含子和基因间隔区丰富. 进一步分析显示，微卫星在每条染色体两臂的丰度和分布存在着很多的相似性.

关键词 微卫星，按蚊，基因组，丰度，分布

学科分类号 Q71

---

*国家自然科学基金资助项目(30300262).

同等贡献作者. * 通讯联系人.

Tel: 023-68250793, E-mail: libinlb@hotmail.com,

Tel: 023-68250346, Fax: 023-68251123, E-mail: lucheng@swau.cq.cn

收稿日期: 2004-11-30, 接受日期: 2005-01-31