Comprehensive Analysis of Simple Sequence Repeats in Pre-miRNAs

Ming Chen,[1,2,3] Zhongyang Tan,*[,3] Guangming Zeng,*[,1,2] and Jun Peng[3]

[1]College of Environmental Science and Engineering, Hunan University, Changsha, China

[2]Key Laboratory of Environmental Biology and Pollution Control (Hunan University), Ministry of Education, Changsha, China

[3]Institute of Life Sciences and Biotechnology, Hunan University, Changsha, China

***Corresponding author:** E-mail: zhongyang@hnu.cn; zgming@hnu.cn.

Associate editor: Takashi Gojobori

Abstract

Simple sequence repeats (SSRs) are tandem repeat units of 1–6 bp that are identified in various complete sequences. However, the distribution, nature, and origination of SSRs in pre-miRNAs, which are characteristic stem-loop sequences and are finally processed into ~22 nt functional miRNAs contributing to regulate several biological processes, are still not well studied. The availability of large numbers of pre-miRNAs makes it possible to analyze and compare the occurrences of SSRs, the relative count of SSRs, or the longest SSRs in pre-miRNAs. In this study, we analyzed SSRs in 8,619 pre-miRNAs from 87 species, including Arthropoda, Nematoda, Platyhelminthes, Urochordata, Vertebrata, Mycetozoa, Protistae, Viridiplantae, and Viruses. We find that SSRs widely exist in the pre-miRNAs analyzed. Our analysis shows that mononucleotide repeats are the most abundant repeats, followed by dinucleotide repeats, whereas tri-, tetra-, penta-, and hexanucleotide repeats rarely occurred in pre-miRNAs. The number of SSRs per pre-miRNA on average ranges from 4.1 for viruses to 13.5 for Mycetozoa. Our results confirm that the number of repeats correlates inversely to the length of repeats. Generally, in each taxonomic group, the occurrence and relative count of SSRs decrease with the increase of repeat unit. SSRs do not exhibit obvious preference for special location in pre-miRNAs. The repeats in pre-miRNAs are complementary to repeats in coding or noncoding regions of genomes, and no significant difference is observed between these two classes with respect to the occurrence of repeats. These data on SSRs may become a useful resource of pre-miRNAs, and their possible functions are discussed.

Key words: simple sequence repeat, pre-miRNA, miRNA, microsatellite.

Introduction

Simple sequence repeats (SSRs), or microsatellites, consist of repeated unit between 1 and 6 bp long (Chen et al. 2009). SSRs are believed to be originated from either de novo genesis or adoptive genesis (Kim et al. 2008). Errors of DNA replication and/or repair machinery and unequal recombination have been proposed to be responsible for the generation and instability of SSRs (Toth et al. 2000; Katti et al. 2001). Furthermore, the interplay between the repeat type, the genomic position of the SSR, and the genetic–biochemical background of the cell is believed to be important for the formation and fixation of SSRs (Toth et al. 2000). SSRs are found in both prokaryotic (Gur-Arie et al. 2000; Mrazek et al. 2007) and eukaryotic genomes (Toth et al. 2000; Katti et al. 2001; Rajendrakumar et al. 2007; Bacolla et al. 2008) and in protein-coding regions (Madsen et al. 2008) as well as noncoding regions (Riley and Krieger 2009a, 2009b) of a variety of different genome sequences. The distribution of SSRs in the genome is not random (Li et al. 2004; Kim et al. 2008). SSRs are inherently unstable and hence highly polymorphic (Heesacker et al. 2008). It is believed that the mutation rate of SSRs generally increases with the increase of the repeat unit and repeat tracts (Katti et al. 2001). There are accumulating evidences that SSR expansions or contractions within genome sequences can affect functions of these sequences, even lead to phenotypic changes (Fondon and Garner 2004; Kashi and King 2006). In protein-coding regions, SSR expansions and/or contractions can result in the generation of toxic or malfunctioning proteins (Usdin 2008). Furthermore, dynamic mutations including expansions/deletions in SSRs, especially in trinucleotide repeats, may cause diseases (Li et al. 2002; Usdin 2008).

SSRs have been extensively used as genetic markers (Temnykh et al. 2001; Dereeper et al. 2007), such as for construction of genetic maps (Katti et al. 2001). SSRs are also used in the studies of linkage association (Abdurakhmonov et al. 2005), phylogenetics (Flannery et al. 2006), and population genetics (Rosenberg et al. 2002). Moreover, SSRs have been developed to target disease genes (Mein et al. 1998), to test for paternity (Foster et al. 1998), or to study the evolutionary history of some species (Pritchard et al. 1999).

However, despite widespread study, little is known about SSRs in very short functional sequences or other specific well-defined contexts. Pre-miRNA is very short in length and is finally processed into a functional miRNA (Bartel 2004). Many characteristics and functions of pre-miRNA and miRNA have been demonstrated in recent years (Lim et al. 2003; Yousef et al. 2007). The rapid accumulation of pre-miRNAs brings the opportunities and

Table 1. List of Analyzed Pre-miRNAs.

Taxa	Number of Species[a]	Number of Pre-miRNAs	Total Length of Analyzed Pre-miRNAs (bp)	Average GC Content (%)[b]
Arthropoda	16	1,194	103,335	42.72
Nematoda	2	249	23,786	43.92
Platyhelminthes	1	63	5,552	34.00
Urochordata	3	127	12,791	43.46
Vertebrata	28	5,157	452,667	47.33
Mycetozoa	1	2	289	36.23
Protistae	1	49	12,284	59.68
Viridiplantae	21	1,638	238,433	44.92
Viruses	14	140	10,763	54.43

[a] Indicates number of species that is shown to contain pre-miRNAs in miRBase 12.0. For more detailed information about these pre-miRNAs, please browse the miRBase (release 12.0).
[b] Average GC content = total GC content of analyzed pre-miRNAs/the number of analyzed pre-miRNAs.

needs to explore and understand SSRs in them. In this study, we analyzed SSRs in 8,619 pre-miRNAs for their occurrence, nature, organization, and distribution. We reasoned that characterizing the SSR distribution in pre-miRNAs would provide an opportunity to: 1) observe the organization of pre-miRNAs in terms of SSRs, 2) lay a good foundation for further understanding the roles of SSRs in pre-miRNAs, and 3) compare with SSRs in other genome sequences such as in genomes of *Escherichia coli*.

Methods

Pre-miRNAs Sequences

A total of 8,619 pre-miRNAs, release 12.0 (September 2008), were downloaded in FASTA format from ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/12.0/.

SSR Analysis

A new tool based on simple "replace" method was developed with Microsoft visual basic 6.0 for extracting SSRs from DNA or RNA sequence. The software, Mononucleotide Repeats Identifier, was used to get the total number of SSRs from single pre-miRNA or multiple pre-miRNAs. Di-, tri-, tetra-, penta-, and hexanucleotide repeats were identified and localized by the software SSRIT (http://www.gramene.org/db/searches/ssrtool). In this study, the two tools were employed to search SSRs repeating three times or more. Total SSRs of each taxonomic group have been averaged to allow comparison among different taxa:

$$E = \frac{\mathrm{Total}_1 + \mathrm{Total}_2 + \cdots + \mathrm{Total}_T}{T} = \frac{1}{T}\sum_{i=1}^{T}\mathrm{Total}_i,$$

where E, named as relative count in this study, is the mean of total repeats in each taxonomic group; T is the number of pre-miRNAs in each taxonomic group; and "Total" is the total repeats in each taxonomic group.

Results

When identifying SSRs in a given sequence, definition of the minimum repeat number is an important empirical criterion. Minimum repeat number of a repeat tract is the number of repeat units in a tract to be considered as a valid SSR tract. For detection of various repeats in pre-miRNAs, we selected three as the minimum repeat number that was used for the survey of SSRs in *E. coli* (Gur-Arie et al. 2000).

In this study, all pre-miRNAs (miRBase 12.0) were classified into nine taxonomic groups (Arthropoda, Nematoda, Platyhelminthes, Urochordata, Vertebrata, Mycetozoa, Protistae, Viridiplantae, and Viruses) according to the taxonomic criterion of miRBase 12.0 (table 1).

Comprehensive analysis of SSRs confirmed that SSRs in pre-miRNAs were diverse in terms of motif and repeat number, and SSRs were widely distributed in most of the sequences analyzed.

Mononucleotide Repeats

Mononucleotide repeats were the most common repeats in all pre-miRNAs (table 2). In Arthropoda, Nematoda, Platyhelminthes, and Urochordata, poly (A) and poly (U) were predominant in their pre-miRNAs, whereas in Vertebrata, the numbers of poly (A), poly (U), poly (G), and poly (C) were comparable (table 3). In contrast, in Virus, poly (G/C) outnumbered poly (A/U) in 10 of 14 viral species (supplementary table S2 and fig. S1, Supplementary Material online). Pre-miRNA contained approximately 3.8–7.2 mononucleotide repeats on average in each taxonomic group (table 2). However, two pre-miRNAs from Mycetozoa contained more than 13 mononucleotide repeats (table 2).

Dinucleotide Repeats

Dinucleotides were the second most common repeats in all of surveyed pre-miRNAs (table 2). We found that GU/UG repeats were predominant in Arthropoda, Urochordata, Vertebrata, and Viruses, whereas AU/UA repeats were common in Platyhelminthes (supplementary fig. S2, Supplementary Material online). Interestingly, in Nematoda, the numbers of AG/GA, GU/UG, AC/CA, CU/UC, and AU/UA repeats were nearly equal, but GC/CG repeats were relatively rare (table 4; supplementary table S3, Supplementary Material online). Although the pre-miRNAs from Protistae and Viridiplantae contained on average 0.6 dinucleotide repeats, no dinucleotide repeats were found in the two pre-miRNAs from Mycetozoa.

Trinucleotide Repeats

It has been reported that the trinucleotide repeats were the most abundant one in *Neurospora crassa* genome (Kim

Table 2. Occurrence and Relative Count[a] of SSRs in Pre-miRNAs.

Taxa	Number of Pre-miRNAs	Repeat Type						
		Mono	Di	Tri	Tetra	Penta	Hexa	Total
Arthropoda	1,194	5,490 (4.6)	375 (0.3)	14 (0.01)	0 (0)	0 (0)	1 (0.0008)	5,880 (4.9)
Nematoda	249	1,186 (4.8)	44 (0.2)	6 (0.02)	0 (0)	0 (0)	0 (0)	1,236 (5.0)
Platyhelminthes	63	326 (5.2)	9 (0.1)	2 (0.03)	0 (0)	0 (0)	0 (0)	337 (5.3)
Urochordata	127	622 (4.9)	26 (0.2)	7 (0.06)	0 (0)	0 (0)	0 (0)	655 (5.2)
Vertebrata	5,157	23,593 (4.6)	1,551 (0.3)	143 (0.03)	28 (0.01)	3 (0.0006)	1 (0.0002)	25,319 (4.9)
Mycetozoa	2	27 (13.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	27 (13.5)
Protistae	49	550 (11.2)	29 (0.6)	17 (0.35)	2 (0.04)	1 (0.02)	1 (0.0204)	600 (12.2)
Viridiplantae	1,638	11,765 (7.2)	1,024 (0.6)	173 (0.11)	42 (0.03)	6 (0.003)	3 (0.0018)	13,013 (7.9)
Viruses	140	529 (3.8)	36 (0.3)	0 (0)	1 (0.01)	0 (0)	1 (0.0071)	567 (4.1)

[a] SSR relative count (in parentheses) is the total repeats per pre-miRNA on average in every taxonomic group. For example: Arthropoda taxonomic group has 1,194 pre-miRNAs in which 5,490 mononucleotide repeats, 375 dinucleotide repeats, 14 trinucleotide repeats, 0 tetranucleotide repeats, 0 pentanucleotide repeats, and 1 hexanucleotide repeats were found. Thus, relative count of mononucleotide repeats = 5,490/1,194 ≈ 4.6; relative count of dinucleotide repeats = 375/1,194 ≈ 0.3; relative count of trinucleotide repeats = 14/1,194 ≈ 0.01; relative count of tetranucleotide repeats = 0/1,194 = 0; relative count of pentanucleotide repeats = 0/1,194 = 0; relative count of hexanucleotide repeats = 1/1,194 ≈ 0.0008; and relative count of total repeats = 5,880/1,194 ≈ 4.9.

et al. 2008). However, in pre-miRNAs, the trinucleotide repeats were a minor class SSR type. Protistae contained the highest average number of trinucleotide repeats (0.35), followed by Viridiplantae (0.11), and pre-miRNAs of Mycetozoa did not possess any trinucleotide repeat (table 2). Interestingly, most, but not all, of trinucleotide repeats contained base "U." The longest trinucleotide repeat, $(CCG)_8$ with a length of 24 bp, was found in the pre-miRNAs of *Oryza sativa* (supplementary table S4, Supplementary Material online).

Tetranucleotide Repeats

The relative count of tetranucleotide repeats was found to be much less than dinucleotide and trinucleotide repeats in the majority of taxonomic groups. However, there were some exceptions, for example, in Viruses and Protistae taxonomic groups, the relative count of tetranucleotide repeats was more than that of trinucleotide repeats. Most tetranucleotide repeats showed a general dependence on base "G," and only a small number of tetranucleotide repeats did not contain base "G." For example, $(UUUC)_3$, $(UUCA)_3$, and $(AAAC)_4$ repeats did not harbor base "G." Protistae pre-miRNA showed the highest relative count of tetranucleotide repeats (0.04 on average), whereas no tetranucleotide repeats were found in the two pre-miRNAs from Mycetozoa (table 2). The longest tetranucleotide repeat was $(UAUG)_5$ in *Mus musculus* (supplementary table S5, Supplementary Material online).

Table 3. Occurrence and Relative Count[a] of Mononucleotide Repeats in Pre-miRNAs.

Taxa	Repeat Type			
	A	U	G	C
Arthropoda	1,647 (1.38)	2,673 (2.24)	694 (0.58)	476 (0.40)
Nematoda	414 (1.66)	480 (1.93)	140 (0.56)	152 (0.61)
Platyhelminthes	129 (2.05)	167 (2.65)	20 (0.32)	10 (0.16)
Urochordata	206 (1.62)	262 (2.06)	72 (0.57)	82 (0.65)
Vertebrata	5,312 (1.03)	7,442 (1.44)	6,127 (1.19)	4,712 (0.91)
Mycetozoa	10 (5)	10 (5)	2 (1)	5 (2.5)
Protistae	76 (1.55)	87 (1.78)	194 (3.96)	193 (3.94)
Viridiplantae	3,243 (1.98)	4,642 (2.83)	2,149 (1.31)	1,731 (1.06)
Viruses	86 (0.61)	142 (1.01)	169 (1.21)	132 (0.94)

[a] SSR relative count (in parentheses) is the total repeats per pre-miRNA on average in every taxonomic group. For example: Arthropoda taxonomic group has 1,194 pre-miRNAs in which 1,647 A repeats; 2,673 U repeats; 694 G repeats; and 476 C repeats were found. Thus, relative count of A repeats = 1,647/1,194 ≈ 1.38; relative count of U repeats = 2,673/1,194 ≈ 2.24; relative count of G repeats = 694/1,194 ≈ 0.58; and relative count of C repeats = 476/1,194 ≈ 0.40.

Pentanucleotide Repeats

Pentanucleotide repeats were relatively rare in pre-miRNAs. The highest relative count of pentanucleotide repeats (0.02 on average) was found in the pre-miRNAs of Protistae. No pentanucleotide repeats were identified in Arthropoda, Nematoda, Platyhelminthes, Urochordata, Mycetozoa, and Viruses (table 2). Only nine classes of pentanucleotide repeat motifs were found in all of pre-miRNAs: CGGCU, GGGCU, GGAGU, AGAAU, UCGCC, AACCC, UUUUA, GGAGA, and CUUGG. Protistae group owned the longest pentanucleoide repeats $(GGAGU)_7$ in pre-miRNAs of *Chlamydomonas reinhardtii* (supplementary table S6, Supplementary Material online).

Hexanucleotide Repeats

Hexanucleotide repeats are less frequent than pentanucleotide repeats in pre-miRNAs. The highest relative count of hexanucleotide repeats (0.0204) was found in Protistae, and no hexanucleotide repeats were detected in Nematoda, Platyhelminthes, Urochordata, and Mycetozoa (table 2). We identified the longest hexanucleotide repeat in pre-miRNAs of *Drosophila ananassae* of the Arthropoda group, which is $(ACUGCC)_5$ with a length of 30 bp (supplementary table S7, Supplementary Material online).

Discussion

We investigated the SSR distribution in pre-miRNAs of Arthropoda, Nematoda, Platyhelminthes, Urochordata, Vertebrata, Mycetozoa, Protistae, Viridiplantae, and Viruses. Our data revealed both similarities and uniqueness in composition and distribution patterns of SSRs in pre-miRNAs (tables 2–4; supplementary tables S1–S9, Supplementary Material online). Due to the fact that pre-miRNAs derive from introns, exons, or intergenic regions (Bartel 2004), the repeats in pre-miRNAs can be complementary

Table 4. Occurrence and Relative Count[a] of Dinucleotide Repeats in Pre-miRNAs.

	Repeat Type					
Taxa	AG/GA	GU/UG	AC/CA	CU/UC	AU/UA	CG/GC
Arthropoda	56 (0.05)	104 (0.09)	26 (0.02)	72 (0.06)	75 (0.06)	42 (0.04)
Nematoda	7 (0.03)	8 (0.03)	9 (0.04)	9 (0.04)	9 (0.04)	2 (0.008)
Platyhelminthes	0 (0)	1 (0.02)	1 (0.02)	0 (0)	7 (0.11)	0 (0)
Urochordata	4 (0.03)	8 (0.06)	3 (0.02)	4 (0.03)	6 (0.04)	1 (0.008)
Vertebrata	206 (0.04)	546 (0.11)	231 (0.04)	316 (0.06)	215 (0.04)	37 (0.007)
Mycetozoa	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Protistae	0 (0)	3 (0.06)	7 (0.14)	0 (0)	3 (0.06)	16 (0.33)
Viridiplantae	252 (0.15)	151 (0.09)	90 (0.05)	269 (0.16)	229 (0.14)	33 (0.02)
Viruses	5 (0.04)	12 (0.09)	3 (0.02)	10 (0.07)	0 (0)	6 (0.04)

[a] SSR relative count (in parentheses) is the total repeats per pre-miRNA on average in every taxonomic group. For example: Arthropoda taxonomic group has 1,194 pre-miRNA in which 56 AG/GA repeats were found. Thus, relative count of AG/GA repeats = 56/1,194 ≈ 0.05.

to repeats in coding or noncoding regions. Although previous studies showed a distinct distribution of SSRs in protein-coding and intergenic regions of genomes because of their different selective constrains, we found that there is no significant difference between these two classes with respect to the occurrence of repeats in pre-miRNAs based on the present data. In most instances, the number of the SSRs (di–hexa) across pre-miRNAs with similar sizes is also similar (supplementary table S9, Supplementary Material online). For example, in a comparison of same-sized pre-miRNAs, hsa-mir-7-3 from protein-coding gene of *Homo sapiens* has the same number of SSRs (di–hexa) with has-mir-181b-1 from intergenic regions of *H. sapiens* (Rodriguez et al. 2004). SSRs in pre-miRNAs may derive from backward slippage and resynthesis of the same RNA sequence that result from transient pausing of the RNA polymerase complex during transcription (Jacques and Kolakofsky 1991). Our results also demonstrated that SSR distribution in pre-miRNAs varied among species and/or taxonomic groups, and there was an excess of repeats shorter than those traditionally considered to be SSRs.

Among most of surveyed pre-miRNAs, poly (A/U) repeats were more frequent than poly (G/C) repeats. When we considered "U" as "T," this observation was similar to that of *N. crassa* genome (Kim et al. 2008) or primate genome (Subramanian et al. 2003). However, poly (G/C) repeats were more abundant in most of viral pre-miRNAs. These could be explained by their distinct GC content (the GC content of pre-miRNAs in most of species is below 50%, whereas that of viruses is mostly above 50%) (table 1; supplementary tables S10 and S11, Supplementary Material online). Mononucleotide and dinucleotide repeats were significantly predominant, which was similar to that of introns in which a majority of SSRs were also mononucleotides and dinucleotides (Li et al. 2004), whereas tri-, tetra-, penta-, and hexanucleotide repeats were relatively rare. It has been reported that $(GT)_n$ is the most predominant dinucleotide repeat motif in animal and invertebrates (Stallings et al. 1991). Interestingly, we found that the most abundant repeats in some of their pre-miRNAs were also $(GU)_n/(UG)_n$ (table 4). However, the most abundant repeats appeared to be different in the genome of plants from their pre-miRNAs. The most common repeats was $(AT)_n$ in plant (Lagercrantz et al. 1993), whereas $(CU)_n/(UC)_n$ predominated in their pre-miRNAs (table 4).

Generally, in pre-miRNAs, the numbers of SSRs decreased as the increase of repeat unit and number (table 2; supplementary table S1, Supplementary Material online). This observation is nearly consistent with that in eukaryotes (Toth et al. 2000; Katti et al. 2001), prokaryotes (Gur-Arie et al. 2000), and HIV-1 (Chen et al. 2009). This may be explained by the fact that SSRs with a greater repeat number may be more instable due to the increased probability of slippage (Ellegren 2004). Furthermore, pre-miRNAs of each species had their own longest SSRs (supplementary table S8, Supplementary Material online). The length distributions of all SSRs indicated that the frequency of repeats decreased gradually as length of SSRs increased. This might be due to longer repeats being less stable (Wierdl et al. 1997; Kruglyak et al. 1998). It could also be due to their downward mutation bias and short persistence time (Harr and Schlotterer 2000). SSRs are generally divided into three categories, and the length of the shortest type is 6–12 nt (Rajendrakumar et al. 2007). However, very few SSRs longer than 6 nt were identified in pre-miRNAs. Instead, most of SSRs were only 3 or 4 bp long. The absence of long SSRs and the limited number of repeat types in these pre-miRNAs might be attributable to their smaller size, a relatively stable nature, and low mutation; similar reasoning was used to explain the absence of long SSRs and the limited number of repeat types in mitochondrial, chloroplast (Rajendrakumar et al. 2007), and fungal genomes (Karaoglu et al. 2005). Pre-miRNAs are known to be rather short, but we found several very long SSRs, such as "$(UA)_{11}$," "$(UG)_{11}$," and "$(A)_{19}$" (supplementary table S8, Supplementary Material online). This is especially surprising because such long SSRs are rare even though in some of larger prokaryotic genomes (Field and Wills 1998; Mrazek et al. 2007).

The SSRs of pre-miRNAs in each species did not show obvious preference for a special location. Instead, these SSRs were found anywhere in pre-miRNAs, suggesting that SSRs are an important component of pre-miRNAs (supplementary table S9, Supplementary Material online). Some experiments demonstrated that SSRs located in UTRs and introns can regulate gene expression (Li et al. 2002).

Thus, it is naturally assumed that SSRs, in pre-miRNAs derived from independent transcription units (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001) and the introns of pre-mRNAs (Aravin et al. 2001; Lagos-Quintana et al. 2003), might relate to functions. These SSRs may provide a molecular basis for organization of pre-miRNAs in vivo or fast formation of miRNAs. SSRs variation within pre-miRNAs might be very critical for normal miRNA regular activity because expansion or contraction of SSRs in pre-miRNAs might directly affect the corresponding miRNA products and even cause unexpected changes.

The use of SSRs as genetic markers has become more and more popular because of their abundance and polymorphisms (Karaoglu et al. 2005). Moreover, SSRs also have been used for the study of linkage, association (Abdurakhmonov et al. 2005), and population (Rosenberg et al. 2002). Thus, we believed that characterizing the SSR distribution would foster proper use of SSRs in the future study of pre-miRNA.

Supplementary Material

Supplementary tables S1–S11 and figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors sincerely thank Dr Takashi Gojobori and two anonymous reviewers for suggestions on improving the paper. The authors also thank Dr Jianbo Chen for critically reviewing the manuscript. The study was financially supported by the National Natural Science Foundation of China (50608029, 50978088, 50808073); Hunan Provincial Innovation Foundation for Postgraduate; National Basic Research Program (973 Program) (2005CB724203); Program for Changjiang Scholars and Innovative Research Team in University (IRT0719); Hunan Key Scientific Research Project (2009FJ1010).

References

Abdurakhmonov IY, Abdullaev AA, Saha S, et al. (12 co-authors). 2005. Simple sequence repeat marker associated with a natural leaf defoliation trait in tetraploid cotton. *J Hered*. 96:644–653.

Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, Gvozdev VA. 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr Biol*. 11:1017–1027.

Bacolla A, Larson JE, Collins JR, Li J, Milosavljevic A, Stenson PD, Cooper DN, Wells RD. 2008. Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. *Genome Res*. 18:1545–1553.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297.

Chen M, Tan Z, Jiang J, Li M, Chen H, Shen G, Yu R. 2009. Similar distribution of simple sequence repeats in diverse completed Human Immunodeficiency Virus Type 1 genomes. *FEBS Lett*. 583: 2959–2963.

Dereeper A, Argout X, Billot C, Rami JF, Ruiz M. 2007. SAT, a flexible and optimized Web application for SSR marker development. *BMC Bioinformatics*. 8:465.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*. 5:435–445.

Field D, Wills C. 1998. Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. *Proc Natl Acad Sci U S A*. 95:1647–1652.

Flannery ML, Mitchell FJ, Coyne S, Kavanagh TA, Burke JI, Salamin N, Dowding P, Hodkinson TR. 2006. Plastid genome characterisation in Brassica and Brassicaceae using a new set of nine SSRs. *Theor Appl Genet*. 113:1221–1231.

Fondon JW 3rd, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*. 101:18058–18063.

Foster EA, Jobling MA, Taylor PG, Donnelly P, de Knijff P, Mieremet R, Zerjal T, Tyler-Smith C. 1998. Jefferson fathered slave's last child. *Nature* 396:27–28.

Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y. 2000. Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. *Genome Res*. 10:62–71.

Harr B, Schlotterer C. 2000. Long microsatellite alleles in Drosophila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics* 155:1213–1220.

Heesacker A, Kishore VK, Gao W, et al. (12 co-authors). 2008. SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. *Theor Appl Genet*. 117: 1021–1029.

Jacques JP, Kolakofsky D. 1991. Pseudo-templated transcription in prokaryotic and eukaryotic organisms. *Genes Dev*. 5:707–713.

Karaoglu H, Lee CM, Meyer W. 2005. Survey of simple sequence repeats in completed fungal genomes. *Mol Biol Evol*. 22:639–649.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet*. 22:253–259.

Katti MV, Ranjekar PK, Gupta VS. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evol*. 18:1161–1167.

Kim TS, Booth JG, Gauch HG Jr., Sun Q, Park J, Lee YH, Lee K. 2008. Simple sequence repeats in Neurospora crassa: distribution, polymorphism and evolutionary inference. *BMC Genomics*. 9:31.

Kruglyak S, Durrett RT, Schug MD, Aquadro CF. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A*. 95:10774–10778.

Lagercrantz U, Ellegren H, Andersson L. 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nucleic Acids Res*. 21:1111–1115.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* 294:853–858.

Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T. 2003. New microRNAs from mouse and human. *RNA* 9:175–179.

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* 294:858–862.

Lee RC, Ambros V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science* 294:862–864.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol*. 11:2453–2465.

Li YC, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol*. 21: 991–1007.

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. 2003. The microRNAs of Caenorhabditis elegans. *Genes Dev*. 17:991–1008.

Madsen BE, Villesen P, Wiuf C. 2008. Short tandem repeats in human exons: a target for disease mutations. *BMC Genomics*. 9:410.

Mein CA, Esposito L, Dunn MG, et al. (15 co-authors). 1998. A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nat Genet*. 19:297–300.

Mrazek J, Guo X, Shah A. 2007. Simple sequence repeats in prokaryotic genomes. *Proc Natl Acad Sci U S A*. 104: 8472–8477.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*. 16:1791–1798.

Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM. 2007. Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. *Bioinformatics* 23:1–4.

Riley DE, Krieger JN. 2009a. Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences. *Gene* 429:74–79.

Riley DE, Krieger JN. 2009b. UTR dinucleotide simple sequence repeat evolution exhibits recurring patterns including regulatory sequence motif replacements. *Gene* 429:80–86.

Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res*. 14:1902–1910.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* 298:2381–2385.

Stallings RL, Ford AF, Nelson D, Torney DC, Hildebrand CE, Moyzis RK. 1991. Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. *Genomics* 10:807–815.

Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, Singh L. 2003. Triplet repeats in human genome: distribution and their association with genes and other genomic regions. *Bioinformatics* 19:549–552.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res*. 11:1441–1452.

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

Usdin K. 2008. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res*. 18:1011–1019.

Wierdl M, Dominska M, Petes TD. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* 146:769–779.

Yousef M, Jung S, Kossenkov AV, Showe LC, Showe MK. 2007. Naive Bayes for microRNA target predictions—machine learning for microRNA targets. *Bioinformatics* 23:2987–2992.