J Mol Evol (1999) 48:313–316

JOURNAL OF MOLECULAR EVOLUTION

Evolution of Microsatellites in the Yeast *Saccharomyces cerevisiae:* Role of Length and Number of Repeated Units

Tal Pupko, Dan Graur

Department of Zoology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Ramat Aviv 69978, Israel

Received: 1 June 1998 / Accepted: 16 September 1998

Abstract. The observed and expected frequencies of occurrence of microsatellites in the yeast *Saccharomyces cerevisiae* were investigated. In all cases, the observed frequencies exceeded the expected ones. In contrast to predictions by Messier et al. (1996), there is no critical number of repeats beyond which the observed frequencies of microsatellites significantly exceed the frequencies expected in a random DNA sequence of the same size. Rather, the degree of deviation from expectation was found to be dependent on the length of the microsatellite. That is, a fourfold concatemeric repeat of 3 bp was found to deviate from expectation as much as threefold concatemeric repeat of 4 bp, unlike the deviation of a fourfold concatemeric repeat of 4 bp. These findings suggest that microsatellites evolve through strand-slippage events, rather than recombination events. This, in turn, suggests that the chances of erroneous hybridizations leading to strand-slippage are length dependent.

Key words: Microsatellites — Yeast — *Saccharomyces cerevisiae*

Introduction

Microsatellites are concatemeric repeats of short DNA sequences. Since the number of repeats in a microsatellite is not limited by its definition, any number of tandem repeats of a certain nucleotide combination may be regarded as a microsatellite. Therefore, microsatellites are expected to be found in all genomes. Microsatellites have attracted scientific attention because of (1) their use in the construction of genetic maps (e.g., Dib et al. 1996), (2) the relationship between instability in the number of repeats and genetic disease (e.g., Stallings 1994; Mahadevan et al. 1992; Kremer et al. 1991), and (3) their utility in the assessment of genetic diversity within populations (e.g., Tautz 1989).

How do microsatellites evolve? In a study by Messier et al. (1996), it was concluded that a critical number of repeat units is needed for a microsatellite to become hypervariable at a locus. This conclusion was based on the observation that a G→A mutation in the η-globin locus in the lineage leading to the common ancestor of the apes (gorilla, bonobo, chimpanzee, and humans) caused the sequence ATGTGTGT to change into the sequence ATGTATGT, thus creating an $(ATGT)_2$ microsatellite, which evolved into $(ATGT)_5$ in humans and $(ATGT)_4$ in the African apes. In the lineage leading to the owl monkey, an A→G mutation in the same gene caused the sequence GTATGTGTGT to change into the sequence GTGTGTGTGT, thus creating a $(GT)_5$ sequence that subsequently evolved into $(GT)_6$.

In this study we investigated the following questions: (1) Is there a critical number of repeats beyond which the observed frequencies of microsatellites significantly exceed the frequencies expected in a random DNA sequence of the same size, as anticipated by Messier et al. (1996)? (2) Is this critical point dependent on the number of repeats within the microsatellite or on the length of the microsatellite, and (3) which molecular mechanisms are likely to affect the dynamics of change in repeat number in microsatellites.

Correspondence to: D. Graur; *e-mail:* graur@post.tau.ac.il

Data and Methods

The complete sequence of the 16 chromosomes of the yeast were taken from the *Saccharomyces cerevisiae* database (Goffeau et al. 1997). The mononucleotides were counted for each yeast chromosome separately.

Each type of microsatellite was characterized by the following variables: (1) the sequence of the internal repetitive unit, (2) the length of the repetitive unit, (3) the number of concatemeric repeats, and (4) the total length of the microsatellite. We note that an ATATATAT microsatellite is considered as a fourfold repeat of AT rather than a twofold repeat of ATAT, i.e., the repetitive unit may not contain internal repetitive units. Microsatellites of the same type are distinguished from one another by their genomic position.

The expected number of occurrences for each type of microsatellite was calculated according to de Wachter (1981). For example, the expected probability of occurrence of the microsatellite ATAT or $(AT)_2$ is

$$p^2_{(\mathrm{ATAT})} = (p^2_{\mathrm{A}} \times p^2_{\mathrm{T}})(1 - p_{\mathrm{A}} \times p_{\mathrm{T}})^2 \qquad (1)$$

where p_A and p_T are the frequencies of nucleotides A and T, respectively. In other words, we look for the probability of occurrence of the 8-bp sequence XXATATXX, where XX is not AT. Unlike de Wachter (1981), we did not consider the case in which a microsatellite is located at the end of the sequence. Since in this study an entire genome was used, this source of error should be negligible.

The expected and observed number of occurrences of all types of yeast microsatellites with 1- to 5-bp-long repetitive units was calculated. Since by using a χ^2 test, we found that there is a significant difference ($p < 0.01$) among the relative mononucleotide frequencies among the 16 chromosomes, the expected microsatellite frequencies were calculated for each chromosome separately. We note that a sequence such as GATATATC will be counted both as a threefold AT repeat and as a twofold TA repeat. This bias, however, should be of the same magnitude for both the expected and observed frequencies of occurrence. The deviation from expected frequency was computed for each type of microsatellite.

Results

In order to check how the deviation depends on the length of the microsatellite, all the expected and observed numbers of occurrences for a specific microsatellite length were summarized. For example, observed occurrences of microsatellites containing a 2-bp repeated unit are calculated as the sum of occurrences of microsatellites containing the repetitive units: AC, AG, AT, CA, CG, CT, GA, GC, GT, TA, TC, and TG. The results in Fig. 1 indicate that marked deviations from expected frequencies exist. For example, a 10-fold concatemeric repeat of the sequence A (i.e., AAAAAAAAAA) is expected to occur approximately 46 times in the complete yeast genome, whereas the observed number of occurrences is 393, an excess of 754%. The deviation from expected frequencies increases with the number of repeats. At this stage two alternatives are suggested: (1) the deviation is dependent on the number of repeats, e.g., the deviation from the expected occurrence of two repeats of five nucleotides will be the same as that for two repeats of two nucleotides, and (2) the deviation is dependent on the length of the microsatellite, i.e., the deviation from expected frequencies of microsatellites containing five repeats of 2 bp will be the same as that for two repeats of 5 bp.

The logarithms of the positive deviations from expected frequencies are plotted against the number of concatemeric repeats of the microsatellite (Fig. 2a) and against the total length of the microsatellite (Fig. 2b). A regression analysis indicates that the log deviation is linearly correlated in a statistically significant manner with both the length and the number of repetitive units for each basic repetitive unit length ($p < 0.001$). The lines in Fig. 2a have slopes in the range of 0.32–0.45, whereas the slopes in Fig. 2b range from 0.32 to 2.00, indicating that for all repetitive unit lengths, there is an approximately constant increase in the log deviation from expectation and the total length of the microsatellite with each increase in the total length of the microsatellite causing approximately the same increase in the log deviation. In contrast, the increase in the log deviation due to the increase in the number of internal repeats varies with the length of the repetitive unit.

Since no significant difference was found in the behavior of microsatellites among the chromosomes, we present only the cumulative results for all the chromosomes together. For example, for all chromosomes the deviation from expectation for three repeats of 3 bp was 210 ± 21%.

Discussion

According to Messier et al. (1996) ‘‘a minimum number of repeats units may be necessary before initial expansion occurs.’’ Our results indicate that no critical point exists from which deviations of microsatellite frequencies from expectation are expected to occur. The deviation increases continuously as long as the length of the microsatellites increases.

A study on the *Saccharomyces cerevisiae* RAD5-encoded DNA repair protein (Johnson et al. 1992) indicates that in yeast, changes in microsatellite length arise mostly through polymerase slippage rather than recombination events. Our results imply that the chance of an error in the microsatellite replication from one generation to another is more length dependent than number-of-repeats dependent. This is reasonable if the error source involves an intermediate form due to an erroneous hybridization event. Hybridization strength is length dependent rather than number-of-repeats dependent. We, therefore, propose that the effect of an increase in the chance of slippage events is due to an increase in the frequency of occurrence of erroneous hybridizations.

Weber and Wond (1993) tried to estimate the average mutation rate (mutation = change in the number of repetitive units) in humans and estimated that the rate for 4-bp repeats is four times greater then the rate for 2-bp

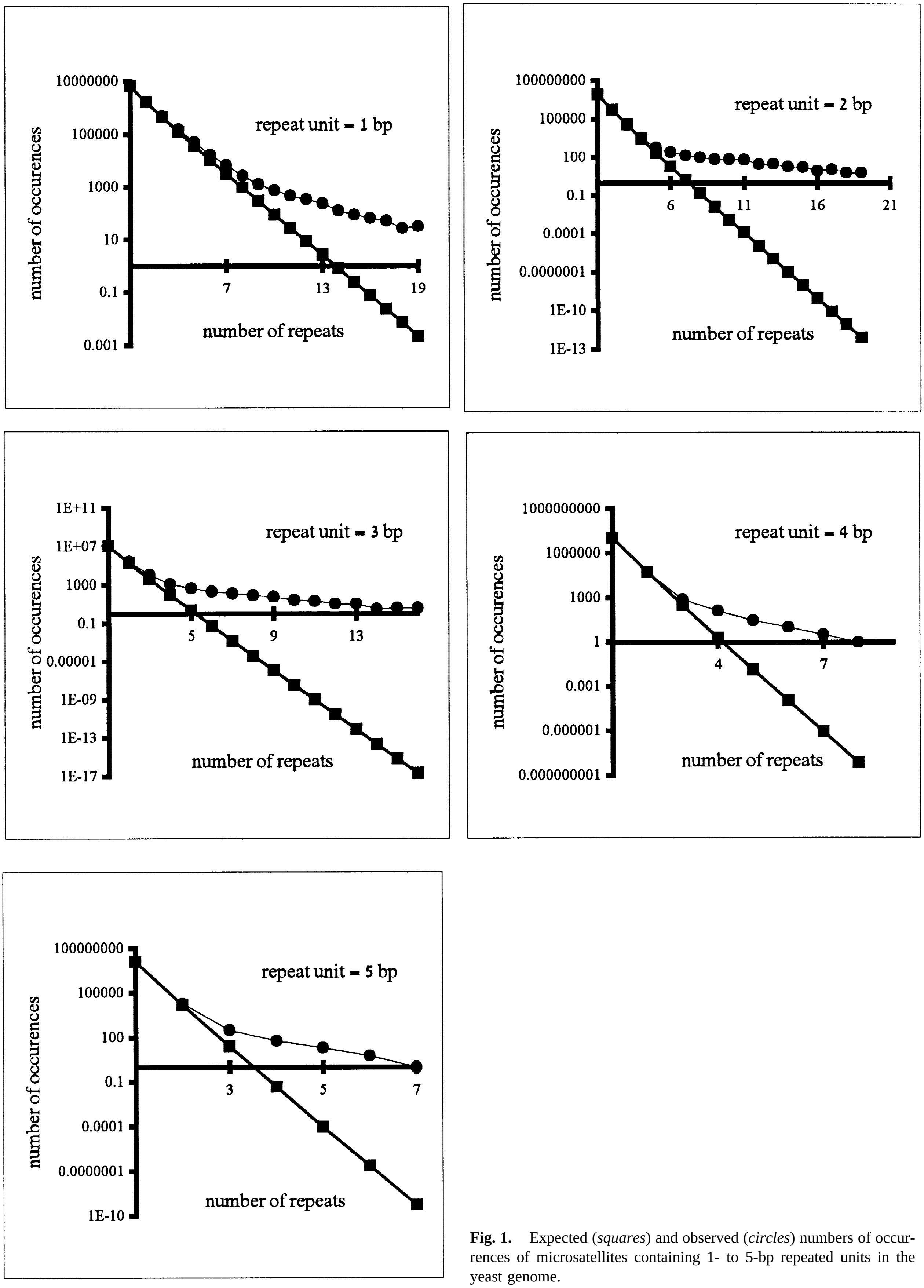

Fig. 1. Expected (*squares*) and observed (*circles*) numbers of occurrences of microsatellites containing 1- to 5-bp repeated units in the yeast genome.

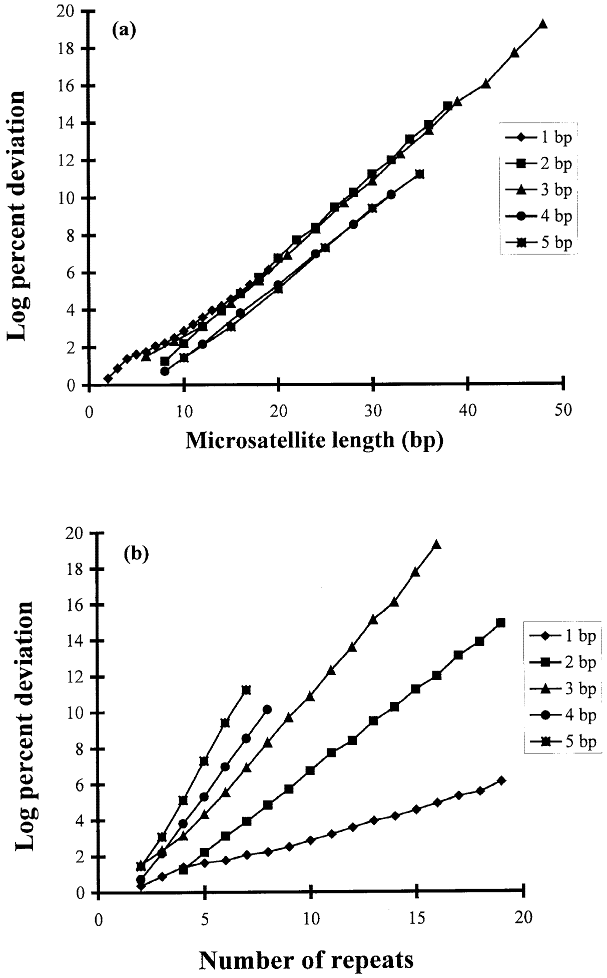

Fig. 2. Observed deviations from expected frequencies of occurrence of microsatellites as a function of **(A)** microsatellite length and **(B)** number of internal repeats.

repeats. Our results imply that the mutation rate should not be estimated on the basis of the length of the repetitive unit but on the basis of the overall length of the microsatellite in question.

The fact that the degree of deviation from expected frequencies is not chromosome dependent implies that the mechanism involved in creating the deviation is working on all the chromosomes at more or less the same level.

Acknowledgments. We thank Yoad Lustig for his help with the computer program. This study was supported by the Magnet Da'at Consortium of the Israel Ministry of Industry and Trade and a grant from Tel Aviv University (689/96).

References

de Wachter R (1981) The number of repeats expected in random nucleic acid sequences and found in genes. J Theor Biol 91:71–98

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissete J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5264 microsatellites. Nature 380:152–154

Goffeau A, et al. (1997) The yeast genome directory. Nature 387S:1–105

Johnson RE, Henderson ST, Petes TD, Prakash S, Bankmann M, Prakash L (1992) Saccharomyces cerevisiae RAD5-encoded DNA repair protein contains DNA helicase and zinc-binding sequence motifs and affects the stability of simple repetitive sequences in the genome. Mol Cell Biol 12:3807–3818

Kremer EJ, Pritchard M, Lynch M, Yu S, Holman K, Baker E, Warren ST, Sclessinger D, Sutherlnd GR, Richards RI (1991) Mapping of DNA instability at the fragile-X to a trinucleotide repeat sequence p(CCG)n. Science 252:1711–1714

Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, Jansen G, Neville C, Narang M, Barcelo J, O'Hoy K, Leblond S, Earle-Macdonald J, De Jong PJ, Wieringa B (1992) Myotonic dystrophy mutation: An unstable CTG repeat in the 3′ untranslated region of the gene. Science 255:1253–1258

Messier M, Li SH, Stewart CB (1996) The birth of microsatellites. Nature 381:483

Stallings RL (1994) Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: Implication for human genetic diseases. Genomics 21:116–121

Tautz D (1989) Hypervariability of simple sequences as a general source of polymorphic DNA markers. Nucleic Acids Res 17:6463–6471

Weber JL, Wond C (1993) Mutation of short tandem repeats. Hum Mol Genet 2:1123–1128