IDENTIFYING NONRANDOM OCCURRENCES OF SIMPLE SEQUENCE REPEATS IN GENOMIC DNA SEQUENCES

Wilfred Ndifon; Asamoah Nkwanta, PhD; Dwayne Hill, PhD

Numerous disorders, including prostate cancer and muscular dystrophy, have been associated with nonrandom occurrences of certain simple sequence repeats (SSRs) found in genomic DNA. In a previous paper, we introduced probabilistic methods for identifying such SSRs that possess nonrandom distribution profiles. Here, we apply these methods to the distribution profiles of SSRs of monomers, dimers, trimers, and tetramers occurring in the human genomic sequence data. In particular, we show that the nonrandomness of SSRs is an exponential function of SSR length. We also demonstrate the existence of threshold SSR lengths for the significant nonrandomness (specifically, under/over-representation) of SSRs. These results are consistent with previous findings and demonstrate the potential use of the previously derived probabilistic methods in the identification of (putative disease associated) SSRs that exhibit significant deviations from random expectations. (*Ethn Dis*. 2005;15 [suppl 5]:S5-67–S5-70)

Key Words: DNA, Simple Sequence Repeats

From the Departments of Biology (WN, DH) and Mathematics (WN, AN) and The Richard N. Dixon Science Research Center (DH), Morgan State University; Baltimore, Maryland.

Address correspondence and reprint requests to Asamoah Nkwanta, PhD; Morgan State University, Department of Mathematics; Carnegie Hall, 258; 1700 E Coldspring Lane; Baltimore, MD 21251; 443-885-4652; 443-885-8216 (fax); nkwanta@jewel.morgan.edu

INTRODUCTION

Microsatellites or simple sequence repeats (SSRs) are tandemly repeated sequences of nucleotides, 1 to 6 base-pairs in length that are ubiquitous in both prokaryotic and eukaryotic genomes.[1] SSRs are highly polymorphic and are considered to be good genetic markers. They have been used in linkage analysis, DNA fingerprinting, genome sequencing, paternity and forensic tests, and population genetic studies.[2–4] The main genetic events responsible for the high polymorphism of SSRs is DNA polymerase slippage.[5] During replication, DNA polymerase may slip, resulting in one or more repeat units present in the parent strand, which becomes expanded or contracted in the nascent strand.[6] In addition to DNA polymerase slippage, unequal recombination, mutation, and selection all contribute to the high polymorphism of SSRs.[6]

The distribution of SSRs in DNA is very nonrandom. Such nonrandomness is associated with numerous disorders including prostate cancer, sickle cell anemia, rheumatoid arthritis, Huntington's disease, muscular dystrophy, fragile X syndrome, Friedrich's ataxia, spinal and bulbar muscular atrophy, spinocerebellar ataxia type 1, 2, 3, 6, and 7.[7–11] Hence, quantifying nonrandomness in genomic DNA sequence data may assist the identification of putative disease-associated SSRs with nonrandom distribution profiles. Several approaches have been developed for quantifying nonrandomness in DNA including complexity studies[12] and word frequency counting, using Bernoulli and Markov models of DNA.[13,14] The methods applied here are based on the latter approach.

Some probabilistic methods, derived in our previous research (Ndifon W, Nkwanta A, Hill D, unpublished data, August 2005), are used in this paper to analyze the nonrandomness of SSRs of monomers, dimers, trimers, and tetramers that occur in the human genomic sequence data. The results from these analyses reveal an exponential relationship between SSR length and nonrandomness. The results also demonstrate the existence of threshold SSR lengths for the significant nonrandomness (specifically, under/over-representation) of each class of SSRs. Moreover, these results are consistent with previous findings and demonstrate the potential use of the previously derived probabilistic methods in identifying putative disease-associated SSRs that possess nonrandom distribution profiles. Note that since SSR expansion is thought to be caused by the presence of unusual DNA structures, the findings reported here lend support to the hypothesis that the probability of forming such unusual structures is an exponential function of DNA sequence length.

MATERIALS AND METHODS

SSR distribution profiles

The distribution profiles of SSRs of monomers, dimers, trimers, and tetramers occurring in the human genomic sequence data were obtained from Dierenger and Schlotterer.[15] Although Dierenger and Schlotterer analyzed the SSR profile data for deviations from random expectations, their approach was different in that they used computer simulations to predict random expectations. The results reported here are based on probabilistic methods derived in our previous research (Ndifon W,

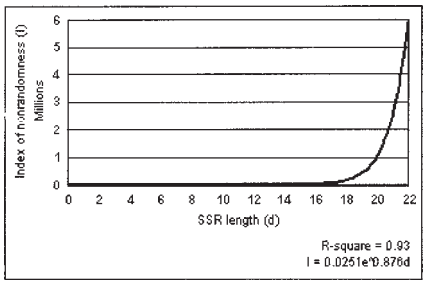

Fig 1. The index of nonrandomness of SSRs of monomers found in the human genome. The index of nonrandomness, *I*, is (approximately) an exponential function of SSR length, *d*: $I_Y = 0.0251e^{(0.876d)}$

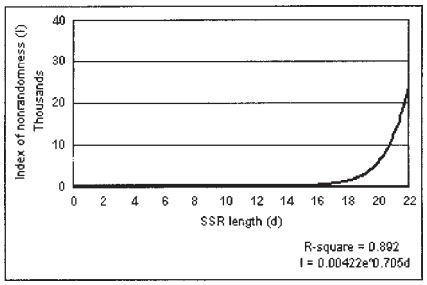

Fig 3. The index of nonrandomness of SSRs of trimers found in the human genome. The index of nonrandomness, I, is (approximately) an exponential function of SSR length, *d*: $I_Y = 0.0042e^{(0.705d)}$

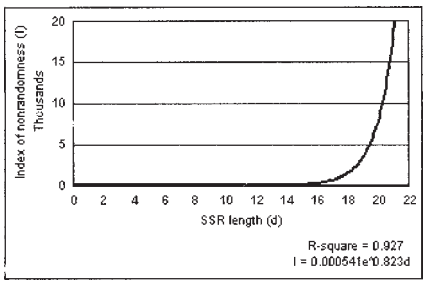

Fig 4. The index of nonrandomness of SSRs of tetramers found in the human genome. The index of nonrandomness, I, is (approximately) an exponential function of SSR length, *d*: $I_Y = 0.0005e^{(0.823d)}$

Nkwanta A, Hill D, unpublished data, August 2005). The methods are briefly discussed below.

Probabilistic methods

We begin with introducing some SSR-related definitions. A DNA sequence X of length n can be thought of as a random word $X = x_1x_2\ldots x_n$ defined over the conventional nucleotide alphabet, $x_i \in \{A, C, G, T\}$. The letters A, C, G, and T denote the bases (or nucleotides) adenine, cytosine, guanine, and thymine, respectively. A k-mer Y is a DNA sequence of length k, where $1 \leq k \leq 6$. A kt-linked SSR of a k-mer Y is a subsequence of X of length kt that consists of t tandem copies of Y.

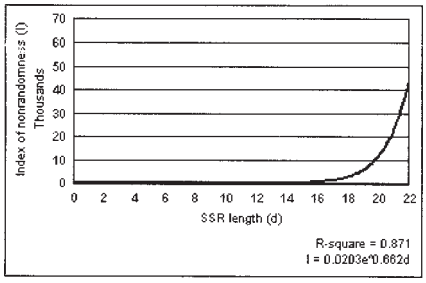

Fig 2. The index of nonrandomness of SSRs of dimers found in the human genome. The index of nonrandomness, I, is (approximately) an exponential function of SSR length, *d*: $I_Y = 0.203e^{(0.662d)}$

The probability $P(U_i = t)$ that an occurrence of a kt-linked SSR of Y starts at position i, $1 \leq i \leq n\text{-}kt\text{+}1$, in X is given by

$$P(U_i = t) = \frac{(N_Y)^t (n + 1 - kN_Y)}{(n + 1 + N_Y(1-k))^{t+1}} \qquad (1.1)$$

Where N_y is the number of occurrences of Y in X, and U_i is the random variable representing the number of tandem copies of Y occurring at position i in X.

The expected number of SSRs of Y is denoted by E_Y and given by

$$E_Y = P(U_i = t)(n - kN_Y + 1) \qquad (1.2)$$

Using Equation 1.2, two metrics for quantifying deviations from expected frequencies of SSRs of Y can be defined. These metrics are the representation and the index of nonrandomness, denoted respectively by R_Y and I_Y. In particular,

$$R_Y = \frac{O_Y}{E_Y},\ E_Y \neq 0, \qquad (1.3)$$

where O_Y is the observed number of SSRs of Y, and

$$I_Y = \frac{1 + (R_Y)^2}{R_Y},\ R_Y \neq 0. \qquad (1.4)$$

Note that SSRs of Y are said to be over-represented or under-represented if $R_Y > 1$ or $R_Y < 1$, respectively. I_Y is defined such that when k-mers Y_1 and Y_2 with similar magnitudes of representation in opposite directions (ie, $R_{Y_1}.R_{Y_2} = 1$), we have $I_{Y_1} = I_{Y_2}$ (ie, SSRs of Y_1 and Y_2 exhibit similarly degrees of nonrandomness).

Using the above probabilistic methods, the nonrandomness of SSRs of monomers, dimers, trimers, and tetramers occurring in the human genomic

Table 1. Threshold SSR lengths for the under- and over-representation of SSRs in the human genome

Repeat Type/Threshold	Under-Representation (bp)	Over-Representation (bp)
Monomer	≤5	≥12
Dimer	≤10	≥15
Trimer	≤10	≥16
Tetramer	≤11	≥15

Note: Observed SSR frequencies, *O*, are close to random expectations, *E*, for relatively short SSRs but begin to deviate significantly for SSRs of length greater than or equal to 12 (monomers), 15 (dimers and tetramers), and 16 bp (trimers)

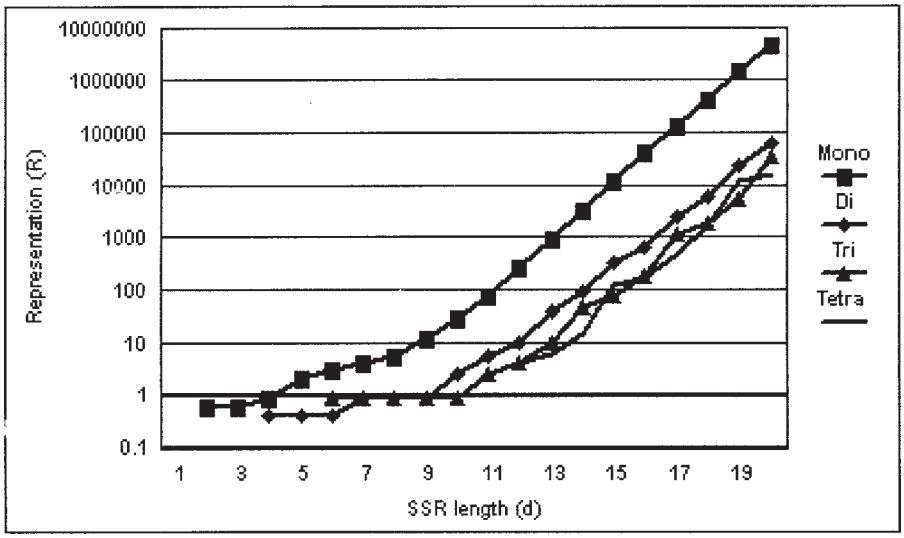

Fig 5. Representation of SSRs of monomers, dimers, trimers, and tetramers found in the human genome. Note the under-representation of SSRs of length less than or equal to 5, 10, 10, and 11 for monomers, dimers, trimers, and tetramers, respectively, and the corresponding over-representation of longer SSRs

sequence data were quantified and subsequently analyzed. Note that only SSRs of *k*-mers of length greater than or equal to *2k* and less than or equal to 20 base pairs were included in the analyses, as was the case in research from Dierenger and Schloterre.[16]

RESULTS AND DISCUSSION

Analysis of the index of nonrandomness of SSRs of monomers, dimers, trimers, and tetramers, found in the human genomic sequence data, showed an exponential relationship between SSR length and the nonrandomness (specifically, over-representation) of SSRs (Figures 1–4). In particular, $I_Y = 0.0251e^{(0.876d)}$, $I_Y = 0.203e^{(0.662d)}$, $I_Y = 0.0042e^{(0.705d)}$, and $I_Y = 0.0005e^{(0.823d)}$ for monomers, dimers, trimers, and tetramers, respectively. This exponential length-dependence of nonrandomness is consistent with previous results[14] and may be due to the fact that the rate of DNA polymerase slippage, the main mechanism responsible for SSR expansion, is also length-dependent. Also, it is known from thermodynamics that the probability of forming a DNA structure is an exponential function of DNA sequence length.[14] This, together with the observation that over-represented SSRs form unusual and stable DNA structures such as H-DNA and cruciforms,[14] suggests that the observed exponential length-dependence of nonrandomness is simply a reflection of the potential of long SSRs to form such unusual DNA structures.

Furthermore, our results support previous hypotheses regarding the existence of a threshold SSR length above which significant over-representation of SSRs becomes evident.[17] Although the actual threshold lengths differed among the various repeat types studied (Table 1), there was a consensus threshold SSR length of approximately 15 bp (Figures 1–4). Interestingly, this corresponds to "zone" three, which is also characterized by significant over-representation of SSRs, in the representation plots of Dierenger and Schlotterrer.[15]

Another interesting result from the analyses was the under-representation of short SSRs. SSRs of monomers, dimers, trimers, and tetramers of length less than or equal to 5, 10, 10, and 10 bp, respectively, were found to be under-represented (Table 1 and Figure 5). Similar under-representation of short SSRs has previously been observed in the genomes of yeast, rice, mouse, and several prokaryotes.[15–16] The under-representation could be explained by the slower rate of DNA polymerase

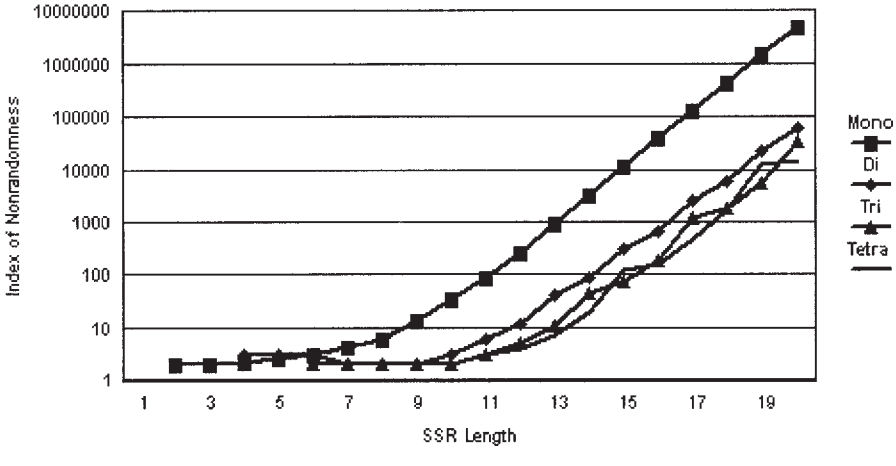

Fig 6. Log plot of the index of nonrandomness of SSRs monomers, dimers, trimers, and tetramers found in the human genomes. There is a stronger correlation between SSR length and nonrandomness for monomers, followed by dimers. For trimers and tetramers, the correlation is very similar and is less than that of monomers and dimers

slippage occurring at genomic regions containing short SSRs.

A comparison of the index of nonrandomness of SSRs of each repeat type showed very similar correlations between SSR length and nonrandomness (Figure 6). SSRs of monomers showed the strongest correlation between length and nonrandomness, followed by SSRs of dimers. The patterns of nonrandomness exhibited by SSRs of trimers and tetramers were very similar.

IMPLICATIONS FOR IMPROVING HEALTH DISPARITIES

The ability to identify SSRs with nonrandom distribution profiles, as demonstrated in this article, is of interest since it may assist the elucidation of disease associated SSRs and the development of novel indicators of adverse health conditions with high prevalence among minority groups. Examples of such SSR-associated adverse health conditions include prostate cancer, sickle cell anemia, arthritis, pulmonary disorders, idiopathic thrombocytopenic purpura, and primary myelodysplastic syndrome.[9–11] The development of indicators of these health conditions will assist the diagnosis, treatment, and management of affected individuals and will contribute to the alleviation of extant health disparities.

ACKNOWLEDGMENTS

This work was funded in part by NSF-HBCU Grant 0236753, by DOE Grant 63580, and by RCMI Grant RR017581.

REFERENCES

1. Schlotterer C. Evolutionary dynamics of microsatellite DNA. *Chromosoma*. 2000;109: 365–371.
2. Jeffreys AJ, Wilson V, Thein SL. Individual-specific 'fingerprints' of human DNA. *Nature*. 1985;316:76–79.
3. Nakamura Y, Leppert M, O'Connell P, et al. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science*. 1987;235:1616–1622.
4. Butler JM. *Forensic DNA Typing: Biology and Technology Behind STR -Markers*. London, UK: Academic Press; 2001.
5. Pearson CE, Sinden RR. Trinucleotide repeat DNA structures: Dynamic mutations from dynamic DNA. *Curr Opin Struct Biol*. 1998;8:321–330.
6. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Gen Res*. 2000;10:967–981.
7. Margolis R L, McInnis MG, Rosenblatt A, Ross CA. Trinucleotide repeat expansion and neuropsychiatric disease. *Arch Gen Psychiatry*. 1999;56:1019–1031.
8. Arzimanoglou II, Gilbert F, Barber HK. Microsatellite instability in human solid tumors. *Cancer*. 1998;82:1808–1820.
9. Panz VR, Joffe BI, Spitz I, et al. Tandem CAG repeats of the androgen receptor gene and prostate cancer risk in Black and White men. *Endocrine*. 2001;15:213–216.
10. Perischon B, Ragusa A, Lapoumeroulie C, et al. Inter-ethnic polymorphism of the beta-globin gene locus control region (LCR) in sickle-cell anemia patients. *Hum Genet*. 1993;91:464–468.
11. Martinez A, Fernandez-Arquero M, Pascual-Salcedo D, et al. Primary association of tumor necrosis factor-region genetic markers with susceptibility to rheumatoid arthritis. *Arthritis Rheum*. 2000;43:1366–1370.
12. Gusev VD, Nemytikova LA, Chuzhanova NA. On the complexity measures of genetic sequences. *Bioinformatics*. 1999;15:994–999.
13. Lippert RA, Huang H, Waterman MS. Distributional regimes for the number of k-word matches between two random sequences. *Proc Natl Acad Sci*. 2002;99:13980–13989.
14. Cox R, Mirkin MS. Characteristic enrichment of DNA repeats in different genomes. *Proc Nat Acad Sci*. 1997;94:5237–5242.
15. Dierenger D, Schloterrer C. Two distinct modes of microsatellite mutation processes: Evidence from the completed genomic sequences of nine species. *Genome Biol*. 2003;13:2242–2251.
16. Kruglyak S, Durrett R, Schug MD, Aquadro CF. Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Mol Biol Evol*. 2000;17:1210–1219.
17. Field D, Wills S. Abundant microsatellite polymorphism in Sacharomyces cereviseae, and the different distributions of microsatellites in eight prokaryotes and S. cereviseae, resulting from strong mutation pressures and a variety of selective forces. *Proc Natl Acad Sci*. 1998;95:1647–1652.
18. Kuraoda J, Kimura S, Kobayashi Y, Wada K, Uoshima N, Yoshikawa T. Unusual myelodysplastic syndrome with the initial presentation mimicking idiopathic thrombocytopenic purpura. *Acta Hematol*. 2002; 108:139–143.
19. Sashida G, Ohyashiki JH, Ito Y, Ohyashiki K. Monoclonal constitution of neutrophils detected by PCR-based human androgen receptor gene assay in a subset of idiopathic thrombocytopenic purpura patients. *Leuk Res*. 2002;26:825–830.