ARTICLE

Repeat Sequences and Base Correlations in Human Y Chromosome Palindromes

Neng-zhi Jin, Zi-xian Liu, Yan-jiao Qi, Wen-yuan Qiu*

Department of Chemistry, State Key Laboratory of Applied Organic Chemistry, Lanzhou University, Lanzhou 730000, China

(Dated: Received on December 28, 2008; Accepted on March 17, 2009)

On the basis of information theory and statistical methods, we use mutual information, n-tuple entropy and conditional entropy, combined with biological characteristics, to analyze the long range correlation and short range correlation in human Y chromosome palindromes. The magnitude distribution of the long range correlation which can be reflected by the mutual information is P5>P5a>P5b (P5a and P5b are the sequences that replace solely Alu repeats and all interspersed repeats with random uncorrelated sequences in human Y chromosome palindrome 5, respectively); and the magnitude distribution of the short range correlation which can be reflected by the n-tuple entropy and the conditional entropy is P5>P5a>P5b>random uncorrelated sequence. In other words, when the Alu repeats and all interspersed repeats replace with random uncorrelated sequence, the long range and short range correlation decrease gradually. However, the random uncorrelated sequence has no correlation. This research indicates that more repeat sequences result in stronger correlation between bases in human Y chromosome. The analyses may be helpful to understand the special structures of human Y chromosome palindromes profoundly.

Key words: Human Y chromosome, Palindrome, Mutual information, Long range correlation, Short range correlation

I. INTRODUCTION

In 2003, the sequencing of the human Y chromosome was completed. The most prominent features are eight palindromes (P1-P8), their arms ranging from 9 kb to 1.45 Mb in length. These arms are imperfect in that each contains a unique, non-duplicated spacer, 2-170 kb in length, at its centre, and have arm-to-arm nucleotide identities of 99.94%-99.997% [1]. A DNA palindrome is a sequence of duplex DNA that is the same when the strand and its complementary strand are read in the opposite directions. The perfect palindrome sequences are not stable; enhancing its stability usually requires sabotaging the central symmetry of palindrome sequences through the addition of asymmetry [2]. The palindromes in human Y chromosome are imperfect, and can form hairpin or cruciform structures through intramolecular base pairing because of the spacer in their center [1]. The palindromes in human Y chromosome carry recognized protein-coding genes, all of which seem to be expressed specifically in testes [3]. Accordingly, the palindromes play an important role in the long stability of the evolution of human males. The palindromes have some unknown structure and function, therefore, interpretation of its special structure and function, as well as how to describe and characterize them, are opportunities and challenges for the theoretical biologist.

The statistical analysis of DNA sequences is of importance for understanding the structure and function of genomes. Several statistic methods have been proposed to study DNA sequence, such as autocorrelation function [4-7], Fourier spectrum analysis [5,8], DNA walk [9,10], computational linguistics [11-14], and information theory [6,8,12,15-22]. As far as we know, the statistical results can reflect biologically significant features, for instance, the periodicity of 3 bp indicates the presence of coding sequence [6,18,23,24] and periodicities of 10-11 bp reflect DNA bendability [4,7]. In addition, statistical results can reflect the long range and short range correlation in DNA sequence [5,6,8-11,15-18,22].

The genetic information is mainly stored in base correlations which are the basis for the grammatical construction of genetic language [8]. Base correlations can be divided into the short range and long range correlation. Up to date, there are no reports about base correlations in palindromes of human Y chromosome. Therefore, in this study, we focus our attention on mutual information, n-tuple entropy H_n and conditional entropy h_n, on the basis of information theory and statistical methods, combining with biological characteristics, to analyze the long range correlation and short range cor-

*Author to whom correspondence should be addressed. E-mail: wyqiu@lzu.edu.cn

relation in human Y chromosome palindromes and the relationship between interspersed repeat sequences and base correlations.

II. DATA AND METHODS

A. Human Y chromosome palindromes

Human Y chromosome has eight palindromes (P1-P8), their arms ranging from 9 kb to 1.45 Mb in length, and their arm-to-arm nucleotide identities of 99.94%-99.997% [1]. We downloaded human Y chromosome palindromes (P1-P8) from the National Center for Biotechnology Information (NCBI) [25].

B. Mutual information

DNA is composed of 4 nucleotides (A, C, G, T), in which A refers to adenine, C refers to cytosine, G refers to guanine, and T refers to thymine. We attempt to determine the long range correlation by analyzing mutual information.

We denote by p_i the relative frequency of nucleotide i occurring in the sequence, and by $p_{ij}(k)$ the relative frequency of the pair of nucleotides i and j within a distance k. Two symbols within a distance k are statistically independent if $p_{ij}(k)$ factors to $p_{ij}=p_ip_j$ for all i and j. The base-base mutual information function [6,17-19,21,22] is

$$I(k) = \sum_{i,j=1}^{4} p_{ij}(k) \log_2 \frac{p_{ij}(k)}{p_i p_j} \tag{1}$$

Mutual information $I(k)$ can quantify the amount of information (in units of bits) that one can obtain about the identity of nucleotide i by learning the identity of nucleotide j located k nucleotides downstream. Clearly, $I(k)=0$ for random uncorrelated sequences, and $I(k)>0$ if $p_{ij}\neq p_ip_j$, so $I(k)$ measures any deviation from statistical independence. Due to the finite length, the bias of the mutual information for a sample of length N has been calculated to be $I=9/(2N\ln 2)$ [17], which is illustrated by a horizontal line in Fig.1 and Fig.3. The mutual information $I(k)$ is proportional to the sum over all 16 correlation function $C(k)$, so a power law decay of $C(k)\sim k^{-\gamma}$ is equivalently described by $I(k)\sim k^{-2\gamma}$ [17,21,22]. Hence, a scaling exponent γ of correlation function leads to an exponent 2γ for the mutual information. Furthermore, the mutual information $I(k)\sim k^{-2\gamma}$ is related to spectral function $S(f)\sim f^{-\beta}$ via $\gamma=1-\beta$ [22]. Therefore, we can obtain the spectral exponent β by calculating the mutual information exponent 2γ, and if β is close to 1, it shows that it has $1/f$ noise and long range correlation.

C. n-tuple Shannon entropy and conditional entropy

The relative frequency of an oligonucleotide (n-tuple) which is n consecutive nucleotides is denoted by $p_i{}^{(n)}$, then the n-tuple Shannon entropy H_n [15,16,20,21] is defined by

$$H_n = -\sum_{i=1}^{4^n} p_i^{(n)} \log_2 p_i^{(n)} \tag{2}$$

Short range correlations between bases are reflected in a sub-linear growth of n-tuple Shannon entropy H_n with length n, for example: $H_2<2H_1$.

The conditional entropy h_n [14,21] is defined by

$$h_n = H_{n+1} - H_n \tag{3}$$

h_n indicates the information contained in the $(n+1)$th letter, presuming the n previous letters are known.

The decay of series h_n (n=1, 2, ..., 7) reflects the statistical dependencies with oligonucleotides of length $n+1$. All logarithms are taken to base 2 and thus the entropies are measured in bits. Consequently, for a random uncorrelated sequence: $H_n=2n$, $h_1=h_2=\ldots=H_1$.

D. Curve fitting and regression analysis

Curve fitting is finding a curve which has the best fit to a series of data points, possibly considering other constraints. Regression analysis allows for an approximate fit by minimizing the difference between the data points and the curve. Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables).

Once a regression model has been constructed, it is important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit include R-squared and hypothesis testing. Statistical significance is checked by an F-test of overall fit, followed by t-tests of individual parameters.

III. CORRELATIONS IN HUMAN Y CHROMOSOME PALINDROME SEQUENCES

A. Long range correlation

In 1992, Peng *et al.* found the long range correlation in DNA sequence [9]. In subsequent study, several methods have been proposed to study long range correlation in DNA sequences: spectrum analysis [5], DNA walk [9], information theory [21,22], and so on. Consequently, in this study, we attempt to determine the long range correlation in the palindromes of human Y chromosome by analyzing mutual information.

DOI:10.1088/1674-0068/22/03/255-261

In order to compare with random uncorrelated sequences, we also analyze a random uncorrelated sequence whose length is 200 kbp. Taking into account the bias of statistics, we calculate mutual information $I(k)$ for $k \leq 10^3$ bp according to Eq.(1), and then fit the data by linear function ($y=ax+b$) on the double logarithmic scale, the slope 2γ, correlation coefficient R, spectral exponent β are summarized in Table I. The fitting is significant after F-test and t-test (significance level $\alpha=0.05$). Figure 1 shows mutual information $I(k)$ of palindrome P5 for k=1, 2, ..., 10^3 bp, and other palindromes are similar to it. The results show $I(k)$ of palindromes is dominated neither by sequence periodicity of k=3 bp nor by sequence periodicities of k=10-11 bp. Instead, they exhibit peaks at about k=135 bp and k=165 bp. If the spectral exponent β of palindromes is close to 1. The palindromes have long range correlation. On the other hand, the correlation coefficient R of a random sequence is approximately equal to 0, hence, the random sequence does not exhibit $1/f$ noise and long range correlation.

TABLE I The linear regression results of mutual information of human Y chromosome palindromes and a random uncorrelated sequence.

Palindrome	-2γ	R	β
P1	−0.6025	−0.9561	0.6988
P2	−0.5304	−0.7890	0.7348
P3	−0.5946	−0.9302	0.7027
P4	−0.6911	−0.8632	0.6545
P5	−0.5228	−0.9001	0.7386
P6	−0.7101	−0.8614	0.6450
P7	−0.3110	−0.4784	0.8445
P8	−0.6648	−0.8470	0.6676
Random sequence	−0.0035	−0.0068	

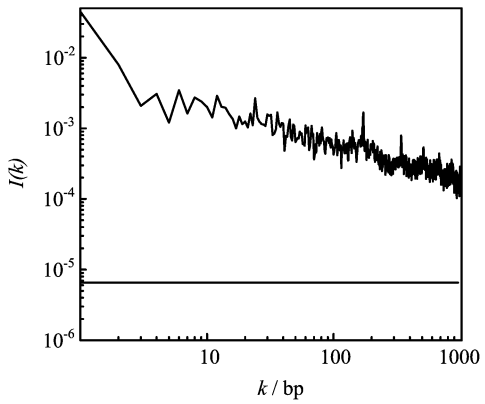

FIG. 1 Mutual information $I(k)$ of human Y chromosome palindrome P5 *vs.* k, the horizontal line is the bias of the mutual information.

B. Short range correlation

In this section we analyze statistical dependencies within oligonucleotides of length n=1, 2, ..., 7, by computing H_n according to Eq.(2) and h_n according to Eq.(3).

Due to the finite length, the estimates of H_n are biased and the bias of entropy increases with the oligonucleotide length proportionally to $(\lambda^n-1)/(2N\ln 2)$ [21], hence, weakly biased estimates of H_n and h_n can be obtained for $n \leq 7$.

Figure 2 shows that H_n of a random uncorrelated sequence is linearly increasing with n and H_n of palindromes are sublinearly increasing with increasing n, which indicates that there are weak, but nonvanishing, short range correlations with n-tuples for n=1, 2, ..., 7. We find that (i) h_n of a random uncorrelated sequence is approximately equal to 2; (ii) h_n of palindromes decrease with the increasing of n, and the decay of H_n and h_n reflects that weak statistical dependencies exist within the n-tuple.

IV. EFFECTS OF REPEATS ON CORRELATIONS

Base correlations can be divided into the long range and short range correlation. In human Y chromosome palindromes, both the long range and short range cor-

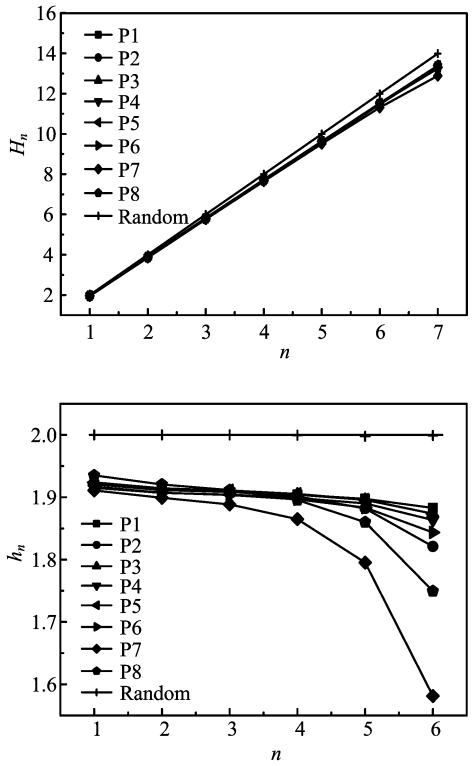

FIG. 2 n-tuple Shannon entropy H_n and conditional entropy h_n of palindromes and a random uncorrelated sequence *vs.* n.

DOI:10.1088/1674-0068/22/03/255-261

relation exist. There is a great need to understand the origin of this phenomenon, making this is one of extremely urgent tasks that we are now addressing.

Repetitive sequences account for at least 50% of the human genome [26,27]. They can be classified into several categories, such as LINE, SINE, and LTR [26]. A human-specific and very abundant type of SINE is the Alu repeat, which constitutes about 10% of the human genome [26]. The Alu element, whose total length is about 300 bp, is a dimer consisting of two monomeric units, with an insertion sequence (about 31 bp) in the second unit [28,29]. Accordingly, the left monomer is about 135 bp long and the right monomer is about 165 bp long. The reason for peaks of mutual information at about k=135 bp and k=165 bp is the Alu element [22].

We utilize Repeatmasker [30] to find the type and content of interspersed repeats in palindromes of human Y chromosome; the results are summarized in Table II. The content of Alu element is lower than the average level of human genome which is about 10% [26]. The GC content of palindromes is lower than the average level of human genome which is about 41% [26]. In order to determine the relationship between the base correlations and repeat sequences, we generate random uncorrelated sequence and substitute repeats by simulated sequence in each palindrome. Take palindrome P5 as an example. In P5, the number of repeat elements is 1129 with a percentage of 66.05%, which range from 12 b to 9.9 kb including SINE, LINE, LTR, satellites, simple repeats, and low complexity elements. P5a is the sequence that replaces solely Alu repeats with random uncorrelated sequences, and P5b is the sequence that replaces all interspersed repeats with random uncorrelated sequences. Figure 3 shows a plot of mutual information $I(k)$ of P5, P5a, and P5b *vs.* k. The horizontal line is the bias of mutual information. Figure 3 demonstrates that $I(k)$ is larger in P5 than in P5a and P5b: in other words, $I(k)$ in P5a and P5b shows significantly less correlation than in P5. The peaks in P5a at about k=135 bp and k=165 bp still exist, and hence the origin of the peaks is not Alu repeats. Furthermore, the decay of $I(k)$ of P5b is close to the bias of mutual

TABLE II Selected features of human Y chromosome palindromes.

Feature	Length/bp	GC%	Repeats/%	Alu repeats/%
P1	3030439	39.30	58.16	7.26
P2	297329	38.04	47.02	4.25
P3	735621	39.78	39.12	6.46
P4	419972	37.94	51.26	8.73
P5	994398	39.65	66.05	8.73
P6	266246	38.47	64.92	8.77
P7	30087	36.81	61.87	4.68
P8	78457	40.25	53.48	7.92

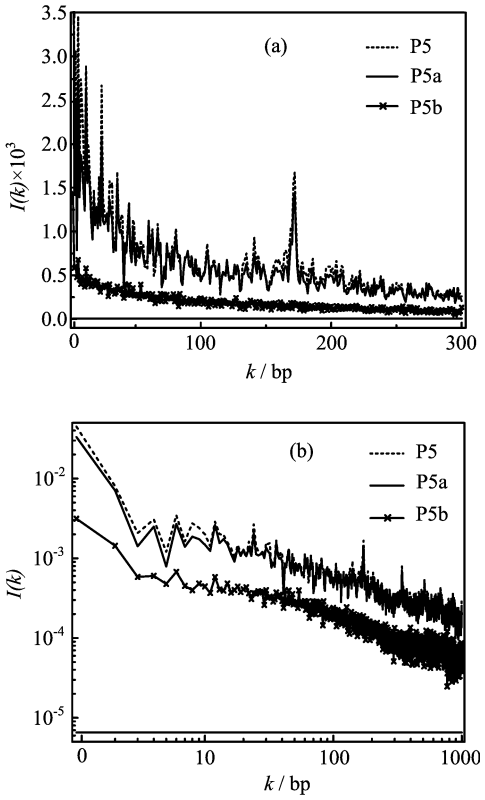

FIG. 3 Mutual information $I(k)$ *vs.* k. The horizontal line is the bias of the mutual information.

information which is shown in the plot of normal scale of Fig.3, nevertheless the decay of $I(k)$ of P5b is greater than the bias of mutual information which is shown in the plot of double logarithmic scale of Fig.3. Therefore, long range correlation of P5b still exists for the scale of k=10^3 bp. Generally, the magnitude distribution of mutual information is P5>P5a>P5b for k=1, 2, ..., 10^3 bp. In other words, the distribution of long range correlation is P5>P5a>P5b. Therefore, the long range correlation is mainly dominated by all interspersed repeats, and the long range correlation is given by repeat sequences.

We also select palindromes P5, P5a, P5b to study the short range correlation. Figure 4 is H_n and h_n of P5, P5a, P5b and random uncorrelated sequence versus n. According to Fig.4, which demonstrate that P5 is close to P5a, the distribution of short range correlation is P5>P5a>P5b>random uncorrelated sequence. Therefore, we propose that the short range correlation is mainly decided by interspersed repeats.

V. RESULTS AND DISCUSSION

The palindromes in human Y chromosome which carry recognized protein-coding genes which seem to be expressed specifically in testes play an important role in the long stability of the evolution of the hu-

DOI:10.1088/1674-0068/22/03/255-261

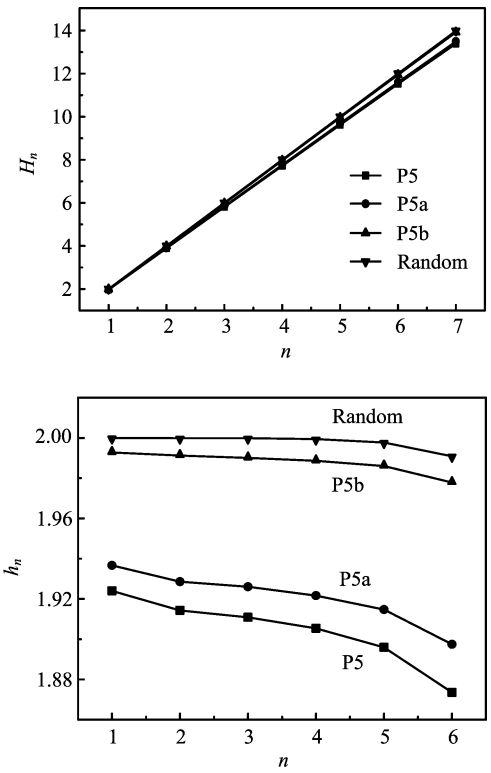

FIG. 4 The n-tuple entropy H_n and the conditional entropy h_n of the human Y chromosome palindrome P5, the repeat-modified versions P5a and P5b, and a random uncorrelated sequence *vs.* n.

man male. Furthermore, the palindromes have some unknown structure and function. Hence, interpretation of the genetic information which is stored in the palindromes in human Y chromosome is an urgent task for bioinformatics.

The genetic information is mainly stored in base correlations which are the basis for the grammatical construction of genetic language. In order to exploit genetic information on genetic language one should first study the base correlations of the language.

As the basis for grammatical construction of genetic language, base correlations which are very important for genetic language in DNA sequence can be divided into the long range and short range correlation. Several statistical methods have been proposed to study long range and short range correlation in DNA sequences, but until the current research there were no reports about the base correlations in palindromes of human Y chromosome. Therefore, in this work, we utilize the methods of statistics and information theory to study the long range and short range correlation of palindromes of human Y chromosome. We find that the palindromes of human Y chromosome have both the long range and short range correlation, whereas the random uncorrelated sequence has no correlation between bases.

The plots of mutual information $I(k)$ of P5 show that $I(k)$ is dominated neither by sequence periodicity of k=3 bp nor by sequence periodicities of k=10-11 bp. However, $I(k)$ exhibits peaks at about k=135 bp and k=165 bp and this can not be explained by Alu repeats because the peaks still exist in repeat modified P5a. The reason for this phenomenon requires further study. The decay of $I(k)$ of P5b is greater than the bias of mutual information although $I(k)$ of P5b is close to the bias of mutual information for k=1, 2, ..., 10^3 bp. In other words, if all interspersed repeats are substituted by random uncorrelated sequences, the repeat modified P5b still has long range correlation for the scale of k=10^3 bp. According to Fig.3, the magnitude distribution of mutual information is P5>P5a>P5b for k=1, 2, ..., 10^3 bp. Therefore, we can conclude that the long range correlation is related to repeat sequences, and along with the content of repeat sequence increasing, the long range correlation will increase. On the other hand, in the light of Fig.4, the magnitude distribution of the short range correlation which can be reflected by H_n and h_n is P5>P5a>P5b>random uncorrelated sequence.

VI. CONCLUSION

In summary, the palindromes of human Y chromosome have both long range and short range correlation, whereas the random uncorrelated sequence has no correlation between bases. Furthermore, the long range and short range correlation of palindromes of human Y chromosome are mainly dominated by repeat sequences, and the base correlations become stronger when the repeat sequence content increases.

VII. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.20173023 and No.90203012) and the Specialized Research Fund for the Doctoral Program of Higher Education of China (No.20020730006).

[1] H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J.r Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S. Yang, R. H. Waterston, R. K. Wilson, S. Rozen, and D. C. Page, Nature **423**, 825 (2003).

DOI:10.1088/1674-0068/22/03/255-261

[2] E. Akgn, J. Zahn, S. Baumes, G. Brown, F. Liang, P. J. Romanienko, S. Lewis, and M. Jasin, Mol. Cell. Biol. **17**, 5559 (1997).
[3] R. Reijo, T. Lee, P. Salo, R. I. Alagappan, L. G. Brown, M. Rosenberg, S. Rozen, T. Jaffe, D. Straus, O. Hovatta, A. Chapelle, S. Silber, and D. C. Page, Nat. Genet. **10**, 383 (1995).
[4] E. N. Trifonov and J. L. Sussman, Proc. Natl. Acad. Sci. **77**, 3816 (1980).
[5] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[6] H. Herzel, E. N. Trifonov, O. Weiss, and I. Groβe, Physica A **249**, 449 (1997).
[7] H. Herzel, O. Weiss, and E. N. Trifonov, Bioinformatics **15**, 187 (1999).
[8] L. Luo, W. Lee, L. Jia, F. Ji, and L. Tsai, Phys. Rev. E **58**, 861 (1998).
[9] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992).
[10] S. Karlin and V. Brendel, Science **259**, 677 (1993).
[11] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).
[12] L. Frappat, C. Minichini, A. Sciarrino, and P. Sorba, Phys. Rev. E **68**, 061910 (2003).
[13] L. Frappat and A. Sciarrino, Physica A **369**, 699 (2006).
[14] N. Z. Jin, Z. X. Liu, and W. Y. Qiu, Chin. J. Chem. Phys. **22**, 27 (2009).
[15] H. Herzel, W. Ebeling, and A. O. Schmitt, Phys. Rev. E **50**, 5061 (1994).
[16] A. O. Schmitt and H. Herzel, J. Theor. Biol. **188**, 369 (1997).
[17] H. Herzel and I. Groβe, Phys. Rev. E **55**, 800 (1997).
[18] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, Phys. Rev. E **61**, 5624 (2000).
[19] I. Grosse, S. V. Buldyrev, and H. E. Stanley, Pac. Symp. Biocomput. **5**, 611 (2000).
[20] O. V. Kirillova, Phys. Lett. A **274**, 247 (2000).
[21] D. Holste, I. Grosse, and H. Herzel, Phys. Rev. E **64**, 041917 (2001).
[22] D. Holste, I. Grosse, S. Beirer, P. Schieq, and H. Herzel, Phys. Rev. E **67**, 061913 (2003).
[23] R. Staden and A. D. McLachlan, Nucleic Acids Res. **10**, 141 (1982).
[24] D. Holste, I. Grosse, S. V. Buldyrev, H. E. Stanley, and H. Herzel, J. Theor. Biol. **206**, 525 (2000).
[25] http://www.ncbi.nlm.nih.gov/mapview/seqreg.cgi?taxid=9606&chr=Y&from=1&to=57772954
[26] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki, Nature **409**, 860 (2001).
[27] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. Russo Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W.

DOI:10.1088/1674-0068/22/03/255-261

Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, Rui-Ru Ji, Z. Ke, K. A. Ketchum, Z.u Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Yuan Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Lai Cheng, L. Curry, S.e Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Jo$h_n$son, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guig, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, Science **291**, 1304 (2001).

[28] M. A. Batzer and P. L. Deininger, Nat. Rev. Genet. **3**, 370 (2002).

[29] Y. Quentin, Nucleic. Acids Res. **20**, 3397 (1992).

[30] http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker

DOI:10.1088/1674-0068/22/03/255-261