Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Biology and Chemistry

ELSEVIER

Computational Biology and Chemistry 27 (2003) 5–10

www.elsevier.com/locate/compbiolchem

Isochores merit the prefix ‘iso’

Wentian Li [a,*], Pedro Bernaola-Galván [b], Pedro Carpena [b], Jose L. Oliver [c]

[a] *Center for Genomics and Human Genetics, North Shore LIJ Research Institute, 350 Community Drive, Manhasset, NY 11030, USA*
[b] *Departamenta de Física Aplicada II, Universidad de Málaga, E-29071 Málaga, Spain*
[c] *Departamento de Genética, Instituto de Bioteconología, Universidad de Granada, E-18071 Granada, Spain*

Received 28 October 2002; received in revised form 7 November 2002; accepted 8 November 2002

Abstract

The isochore concept in the human genome sequence was challenged in an analysis by the *International Human Genome Sequencing Consortium* (IHGSC). We argue here that a statement in the IHGSC’s analysis concerning the existence of isochores is misleading, because the homogeneity was not examined at a large enough length scale and consequently an inappropriate statistical test was applied. A test of the existence of isochores should be equivalent to a test of homogeneity or equality of windowed GC%. The statistical test applied in the IHGSC’s analysis, the binomial test, is a test of whether individual bases are independent and identically-distributed (iid). For testing the existence of isochores, or homogeneity in windowed GC%, we propose to use another statistical test: the analysis of variance (ANOVA). It can be shown that DNA sequences that are rejected by the binomial test may not be rejected by the ANOVA test.

Keywords: Isochores; Human genome sequence; Analysis of variance; International Human Genome Sequencing Consortium

1. Background

The degree of regional homogeneity in base composition in the human genome is a fundamental property of the genome sequence. Not only does it characterize the organization and evolution of the genome, but it also provides a context for many practical sequence analyses. Statistical quantities such as GC%, used in sequence analyses for purposes such as computational gene recognition, should be sampled from a homogeneous region of the sequence. If these quantities are sampled from an inhomogeneous region, error is introduced and the quality of, e.g. the performance of gene prediction, could be affected.

It has been known for a long time from the work of Bernardi’s group that there are compositionally homogeneous regions in the human genome with sizes of at least 200–300 kb (Bernardi, 1993, 1995). These ‘relatively homogeneous’ or ‘fairly homogeneous’ regions (Bernardi, 2001) are called ‘isochores’ (Cuny et al., 1981), and the whole genome is a mosaic of isochores. Recently, however, this view of human genome was questioned in an initial analysis of the human genome draft sequence (Lander et al., 2001). The analysis apparently shows that no sequence of 300-kb length examined could be claimed to be ‘homogeneous’ (‘... the hypothesis of homogeneity could be rejected for each 300-kb window in the draft genome sequence’, page 877 of Lander et al. (2001)), and a rather remarkable statement was made that, essentially, the isochore concept does not hold (“... isochores do not appear to merit the prefix ‘iso’ ”, page 877 of Lander et al. (2001)).

Clearly, whether isochores exist or not depends on the definition of isochore, and that itself depends on the definition of being homogeneous. What was presented in Lander et al. (2001) was a test of whether bases are independent and identically distributed (iid) in a genome sequence, and this translated as being ‘homogeneous’. This is not how the isochore concept was originally developed and defined (Bernardi, 2001). For one thing,

Abbreviations: ANOVA, analysis of variance; IHGSC, International Human Genome Sequencing Consortium; iid, independent and identically distributed; MHC, major histocompatibility complex.

* Corresponding author. Tel.: +1-516-562-1076; fax: +1-516-562-1153.

E-mail address: wli@nslij-genetics.org (W. Li).

1476-9271/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S1476-9271(02)00090-7

there is distinction between ‘strictly homogeneous’ and ‘fairly homogeneous’ (Bernardi, 2001). For another, ‘identical’ distribution at the base level may not be translated to ‘homogeneous’ GC% at a much larger length scale.

We would like to show that as we shift the focus of attention from the base level to the window level, similar to the ‘coarse-graining’ concept in statistical physics, the nature of the testing of sequence homogeneity changes. A violation of iid property at the base level does not lead to a violation of equality of means of GC% averaged over larger length scales. The binomial distribution used in the old test, where the variance is completely determined by the mean, is replaced, in the new test, by a normal distribution where the variance and mean are two independent parameters. And rather than requiring hundreds of thousands of bases to be iid, the new test only requires tens or hundreds of windowed GC% to be equal on average. This change of the focus to a different length scale would also change the conclusion reached in Lander et al. (2001) concerning isochores.

2. Methods

2.1. Binomial test

Following the web supplementary material for Lander et al. (2001), a binomial test is applied to many GC% values measured from a fixed-sized window (e.g. 20 kb). For example, if the sequence length is 900 kb, there are $n = 45$ such 20-kb windows and 45 GC% values. If the GC% of window i is GC%$_i$, the variance of these GC% is $\sigma^2 = /(n-1)\ \Sigma^{\mathrm{n}}_{\mathrm{i}=1}(\mathrm{GC\%}_i - \overline{\mathrm{GC\%}})^2$. The variance expected from a binomial distribution is $\sigma_0^2 = p_{\mathrm{GC}}(1 - p_{\mathrm{GC}})/20000$, where the binomial probability p_{GC} can be estimated by $\overline{\mathrm{GC\%}}$. The test statistic is $c^2 = (n-1)\sigma^2/\sigma_0^2$. For the null hypothesis $\sigma^2 = \sigma_0^2$, which is true when all 45 base sequences are iid with the same p_{GC}, c^2 approximately follows the $\chi_{\mathrm{df2}=n-1}$ distribution (e.g. $\chi_{\mathrm{df2}=44}$ in our example). For any given c^2 value, the p-value can be determined by the corresponding χ^2 distribution. Other similar tests for the null $\sigma^2 = \sigma_0^2$ can also be used.

2.2. ANOVA test

The ANOVA test (analysis of variance) is applied to several groups of GC%’s (as a comparison, the binomial test is only applied to one group of GC%’s). The concept of ‘group’ and ‘member’ in ANOVA here becomes ‘super-window’ and ‘window’. The number of super-windows partitioned in a sequence is a, and the number of windows in super-window i is n_i. The two ‘sum of squares’ (SS) are defined: $\mathrm{SS}_w = \Sigma^a_{i=1}\ \Sigma^{n_i}_{j=1}(\mathrm{GC\%}_{ij} - \overline{\mathrm{GC\%}_i})^2$ (within a group), and $\mathrm{SS}_a = \Sigma^a_{i=1}\ n_i(\overline{\mathrm{GC\%}_i} - \overline{\mathrm{GC\%}})^2$ (among groups). The test statistic is $F = \mathrm{SS}_a/\mathrm{SS}_w \times \Sigma^a_{i=1}(n_i - 1)/(a-1)$. The distribution of F under the null hypothesis (i.e. $\overline{\mathrm{GC\%}_1} = \overline{\mathrm{GC\%}_2} = \ldots \overline{\mathrm{GC\%}_a}$) is known, and this distribution can be used to determine the p-value.

2.3. Kruskal–Wallis test

A test that is similar to ANOVA but nevertheless does not assume normal distribution for sample values, is the Kruskal–Wallis test, also known as the ‘analysis of variance by ranks’. Each sample is given an overall rank as all samples are pooled. The average rank of group i is R_i,. The ‘sum of squares of ranks’ among groups using the rank instead of GC% can be calculated: $\mathrm{SS}_{a(R)} = \Sigma^a_{i=1} n_i(\overline{R}_i - \overline{\overline{R}})^2$. The test statistic for Kruskal–Wallis test is defined as $H = \mathrm{SS}_{a(R)}/n(n+1/12)$, whose asymptotic distribution under the null hypothesis ($\overline{R}_1 = \overline{R}_2 = \ldots \overline{R}_a$) is the χ^2 distribution with df $= a-1$.

3. Results

For a sequence to be homogeneous in GC%, the mean/average of windowed GC% values sampled from one region of the sequence should be similar to that in another region, with a consideration of the amount of allowed variance. In other words, to claim that a sequence is homogeneous, not only do we need to calculate means of GC% along the sequence, but also we need to know the variance. Unless there is a good reason to suspect otherwise, the mean and the variance are two independent parameters of a statistical distribution. For the homogeneity test in Lander et al. (2001)), however, the variance is simply assumed to be a function of the mean, thus it is not independently estimated.

In Lander et al. (2001), the windowed GC% is assumed to follow a binomial distribution. For a binomial distribution to be true, bases within the window should be iid, similar to tossing the same coin independently many times. Violating this assumption invalidates the use of the binomial distribution. If we remove the iid condition at the base level, the statistical distribution of windowed GC% can be simply considered to be a normal distribution, which, unlike the binomial distribution, has two independent parameters (mean and variance). The mean value can be estimated from a window, whereas the variance should be estimated from a group of windows. Even if the observed GC% does not follow a normal distribution (for example, a skewed distribution, see e.g. Clay et al. (2001)), the key point is that the mean and variance are no longer necessarily dependent as in the case of binomial distributions.

To illustrate our point, we analyze two well known isochore sequences, the Major Histocompatibility Complex (MHC) class III and class II sequences on human chromosome 6 (Fukagawa et al., 1995, 1996; Stephens et al., 1999; Beck et al., 1999), with lengths 642.1 and 900.9 kb, respectively. The exact borders of the two isochore sequences are determined by a segmentation procedure (Oliver et al., 2001; Li, 2001; and an online resource on isochore mapping: http://bioinfo2.ugr.es/isochores/). We first repeat the test in Lander et al. (2001) that these two sequences, when viewed as a collection of many 20 kb windows, are sampled from a binomial distribution. According to Lander et al. (2001), a rejection of this test is considered to be evidence for heterogeneity, thus a rejection of the sequence being an isochore. The test results are included in Table 1, which clearly shows that the variances of GC% values sampled from 20-kb windows are much larger than expected from a binomial distribution, with p-value equal to 0 within the limit of machine precision ($<10^{-50}$).

This result, that the variance of GC% sampled from windows is much larger than expected by binomial distribution, has been known for a long time (Sueoka 1959, 1962; Cuny et al., 1981; Li et al., 1998; Clay et al., 2001 and references therein). It is not surprising that the binomial distribution assumption is rejected even for isochore sequences as shown in Table 1. Nevertheless, this rejection only shows that a 900-kb sequence is not a series of 900 000 identically distributed and uncorrelated bases, as examined 20 000 bases at a time; it is not, as suggested by Lander et al. (2001), a rejection of homogeneity of windowed GC% along the sequence.

To reaffirm our assertion that the binomial test used in Lander et al. (2001) is a test of iid or randomness of the sequence at the base level instead of homogeneity at the window level, we applied the test to one bacterial genome sequence (*Borrelia burgdorferi*, 910.7 kb) and two randomly generated sequences (with same length and base composition as the MHC class III and class II sequences). Table 1 shows that the null hypothesis cannot be rejected by the binomial test for the two random sequences, but it is rejected for the *B. burgdorferi* genome sequence that is one of the most homogeneous in GC%, as shown in a recent survey of archaeal and bacterial genome heterogeneity (Bernaola-Galván et al., 2002) using traditional (standard deviation based) and other (entropic segmentation) criterion.

The statistical test to be used in testing equality (homogeneity) for windowed GC% when the variance and the mean of these GC levels are independent, should be the analysis of variance (ANOVA), or the similar distribution-free Kruskal–Wallis test (‘analysis of variance by ranks’). ANOVA focuses directly on the obtained windowed GC% at a large length scale without assuming any iid property at the base level. ANOVA was previously applied to the study of inter-chromosomal homogeneity in the yeast genome (Li et al., 1998; Oliver and Li, 1998).

To apply the ANOVA or Kruskal–Wallis test to test homogeneity, we split a sequence into several super-windows, and several windows per super-window. GC% from each window is calculated. The null hypothesis is that the mean of windowed GC levels (for ANOVA), or the mean of the ranks of windowed GC levels (for Kruskal–Wallis test) is the same in each super-window. The simplest selection of super-windows and windows is when all windows have the same length. To match the discussions in Lander et al. (2001), we choose the window size close to 20 kb and the super-window size close to 300 kb. This corresponds to two super-windows, 16 windows per super-window for the MHC class III sequence, and three super-windows, 15 windows per super-window for the MHC class II sequence. ANOVA test results of these two isochores are listed in Table 2. The p-values are 0.192, 0.323 (for ANOVA), and 0.386, 0.210 (for Kruskal–Wallis test), respectively, for the MHC class III and class II sequence. The null hypothesis, that means of GC% (or their ranks) in different super-windows are the same, is not rejected.

To check that the ANOVA or Kruskal–Wallis test does reject the null hypothesis when the sequence is not

Table 1
Testing the hypothesis that iid bases sampled from 20-kb windows follow a distribution that is independent of the window (the binomial test)

Sequence	# win (n)	Mean	Var σ^2	Binomial var σ_0^2	σ^2/σ_0^2	$c^2=(n-1)\sigma^2/\sigma_0^2$	p-value
MHC class III	32	0.5188	0.0005345	0.00001248	42.8215	1327.47	0
MHC class II	45	0.4105	0.0007268	0.00001210	60.0709	2703.19	0
Random (class III)	32	0.5185	0.00001137	0.00001248	0.9110	28.2402	0.609
Random (class II)	45	0.4106	0.00001255	0.00001210	1.0369	45.6244	0.404
B. burgdorferi	45	0.2859	0.0001515	0.00001021	14.8432	653.099	0

Five sequences are tested: MHC class III and MHC class II isochore sequences, two random sequences corresponding to these two MHC sequences (same length and same base composition), and bacterium *B. burgdorferi* genome sequence. Detailed explanation of column headers: (1) sequence name; (2) total number of windows in the sequence (n), each contributing a GC% value; (3) mean of the GC% (m); (4) variance of the GC% (σ^2); (5) variance of GC% expected from a binomial distribution ($\sigma^2=m(1-m)/20\,000$); (6) ratio of the two variances σ^2/σ_0^2; (7) test statistic $c^2=(n-1)\,\sigma^2/\sigma_0^2$; (8) p-value from the binomial distribution test.

Table 2
ANOVA and Kruskal–Wallis test results of the seven sequences (two MHC isochore sequences, their combined sequence, three randomized sequences, and bacterium *B. burgdorferi* sequence)

ANOVA						Kruskal–Wallis		
	df	SS	MS	*F*-value	*P*-value	χ^2	df	*P*-value
MHC class III (*sw* = 2, *w* = 16)								
Between windows	1	0.0009159	0.0009159	1.781	0.192	0.751	1	0.386
Within windows	30	0.01543	0.0005143					
MHC class II (*sw* = 3, *w* = 15)								
Between windows	2	0.001658	0.0008288	1.162	0.323	3.122	2	0.210
Within windows	42	0.02997	0.0007137					
MHC class III and II (*sw* = 5, *w* = 15)								
Between windows	4	0.2080	0.05201	70.710	0	53.386	4	0
Within windows	70	0.05148	0.0007355					
A random seq corresponding to class III (*sw* = 2, *w* = 16)								
Between windows	1	0.00000288	0.00000288	0.247	0.623	0.0089	1	0.925
Within windows	30	0.0003496	0.00001165					
A random seq corresponding to class II (*sw* = 3, *w* = 15)								
Between windows	2	0.00004546	0.00002273	1.884	0.165	4.646	2	0.098
Within windows	42	0.0005066	0.00001206					
Another random seq corresponding to class II (*sw* = 3, *w* = 15)								
Between windows	2	0.00002285	0.00001143	1.134	0.331	2.444	2	0.295
Within windows	42	0.0004232	0.00001008					
B. burgdorferi (*sw* = 3, *w* = 15)								
Between windows	2	0.0002064	0.0001032	0.671	0.517	0.193	2	0.908
Within windows	42	0.006461	0.0001538					

df, degrees of freedom; SS, sum of squares; MS, mean squares; *F*-value, test statistic value for ANOVA; χ^2, test statistic value for Kruskal–Wallis test; *P*-value, *P*-value from the ANOVA or Kruskal–Wallis tests; sw and w are the number of super-windows and windows.

considered to be homogeneous, we apply the same test on a sequence that covers both the MHC class III and the MHC class II segment. Since this sequence contains two isochores with different GC%, intuitively this sequence should be heterogeneous. Indeed, the p-values from the ANOVA and Kruskal–Wallis test are both 0 within the limit of machine precision.

When the ANOVA test is applied to the *B. burgdorferi* genome sequence and three randomly generated sequences, the null hypothesis is not rejected, indicating that all these sequences are homogeneous at the respective window and super-window sizes (20 and 300 kb). One random sequence that corresponds to the MHC class II sequence is actually not as homogeneous as one would expect (with p-value of 0.098 in the Kruskal–Wallis test). We re-generated another random sequence that also corresponds to the MHC class II sequence; and this time, the p-value was much larger. These results are more consistent with knowledge of DNA than that those from the binomial test because now a bacterial DNA sequence, as well as random sequences, that would be considered homogeneous by traditional standards (Sueoka, 1962), are indeed confirmed to be homogeneous by statistical tests.

4. Discussions

Due to the 'domains within domains' phenomenon in DNA sequences (Li et al., 1994; Bernaola-Galván et al., 1996; Li, 1997; and an online bibliography on this topic: http://www.nslij-genetic.org/dnacorr/), we should not assume automatically that a homogeneity test result obtained at 20-kb window and 300-kb super-window sizes will hold true for other window and super-window sizes. To check this, we carried out ANOVA tests on the MHC class III and class II sequences at other window and super-window sizes. Fig. 1 shows the result for the ANOVA test result ($-\log_{10}$(p-value)) for window sizes of around 20, 10, 5 and 2.5 kb, and the sequence is partitioned into 2, 3, 5, 8 (2, 3, 5. 9) super-windows for the MHC class III (II) sequence.

Several observations could be made from Fig. 1. First, when GC levels are sampled from (e.g.) 20-kb windows, changing the number of super-windows (i.e. number of partitions of the sequence) does not greatly influence the ANOVA test result. This change corresponds to a regrouping of windowed GC%'s. Generally speaking, if the sequence is homogeneous, with all (super) windows of a fixed size having similar GC% values,

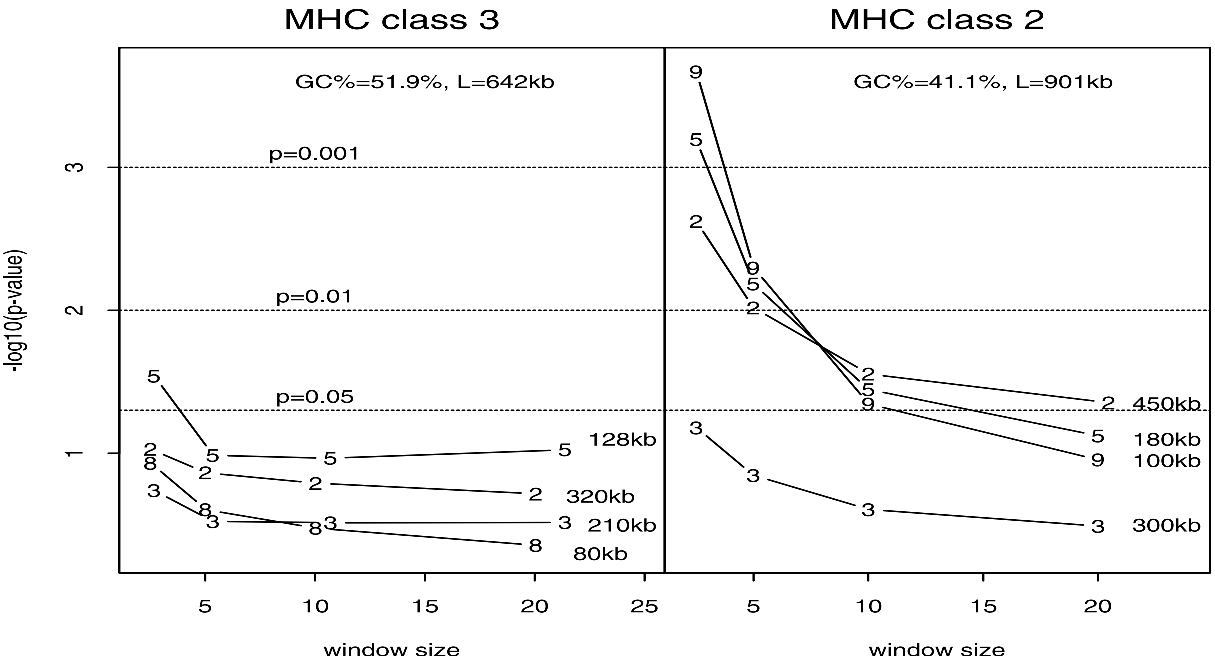

Fig. 1. The $-\log_{10}(P\text{-value})$ of ANOVA tests as a function of the window sizes, for MHC class III (left) and MHC class II (right) sequences. These tests with the same number of super-windows are connected in a line. The size of the super-window and the number of super-windows in the sequence is indicated for each line.

regrouping these values does not make an insignificant result significant.

Second, the ANOVA test becomes more significant when the window size decreases. This observation is understandable because at smaller length scales, with more number of subwindows and larger sample sizes, GC% fluctuations are no longer averaged out. These smaller-length-scale fluctuations could be due to repeats, insertions, foreign elements, etc. For the MHC class II sequence, when the subwindow size is reduced to around 2.5 kb, the ANOVA test result is typically significant (Fig. 1). This observation is consistent with the definition of isochores as ‘fairly homogeneous’ (as vs. ‘strictly homogeneous’) segments above a size of 3 kb (Bettecken et al., 1992; Bernardi, 2000, 2001), and justifies the ‘coarse graining’ procedure to locate isochore boundaries in Oliver et al. (2001).

Third, two isochore sequences may look similar at one length scale (e.g. 20 kb), but quite different at another length scale. Fig. 1 shows that the MHC class II sequence is more heterogeneous than the MHC class III sequence when viewed at the 2–10 kb length scales. GC-poor vertebrate (and random) sequences are generally more homogeneous than GC-rich sequences; or more accurately, a sequence with, a GC% closer to 50% is more heterogeneous than a sequence whose GC% is far away from 50% (Sueoka, 1962; Cuny et al., 1981; Clay et al., 2001). Since the GC%’s of MHC class III and II sequences is 51.9 and 41.1%, respectively, we might expect the MHC class II sequence to be more homogeneous than the class III sequence. Interestingly, Fig. 1 shows the contrary.

To summarize, as we change from focusing on independence and identical distribution at the base level to ‘equality’ or ‘homogeneity’ of GC% at a window level, i.e., by changing from binomial test to ANOVA test, the test result can be completely altered. Since the first draft of this paper, we have gathered reactions to our conclusion, to which we respond as follows.

1) ‘In the binomial test, $GC\%_1$, $GC\%_2$,. . .$GC\%_i$, . . .are more or less independent if the window size is large enough (e.g. 20kb)’. We do not doubt the independence of these GC levels, we only question the expected variance of these GC%’s obtained by assuming each GC% resulted from (e.g.) 20 000 iid bases. The correlation between bases (not windowed GC%’s) invalidates the expected value of variance as used in the binomial test.
2) ‘The sample size of the ANOVA is much smaller than that in the binomial test, and this makes it more difficult for ANOVA to reject null hypothesis than the binomial test’. Indeed, in the example of a 900-kb sequence partitioned into 45 20-kb windows, the binomial test described in the supplementary material of (Lander et al., 2001) is equivalent to dropping 900 000 samples into a 45-by-2 table and testing the correlation between the column and the row by a χ^2 test. In an ANOVA test, since each window is a sample, the number of samples is much smaller. But it is exactly the point we are trying to

address that a change of the length scale is needed in describing homogeneity of windowed GC%. Requiring all 900 000 bases to be sampled from the same iid source is a much more stringent condition than requiring 45 windowed GC levels to be sampled from the same distribution.

3) ‘Whether isochores exist or not depends on the definition of homogeneity’. This is certainly a legitimate reaction, and we agree that any definition of homogeneity is relative (Li, 2001). The issue is that if homogeneity is defined as iid on the base level (‘strictly homogeneous’), then there are perhaps no long DNA sequences (with the exception of highly repetitive DNA sequences) that could ever pass the test. To show that DNA sequences of some genomic regions are relatively more homogeneous than other regions, one has to relax the definition of homogeneity (‘fairly homogeneous’). The ANOVA test accommodates the high variance in real DNA sequences by allowing it to be an independent parameter, which could be estimated from the sequence data. By doing so, we focus on the true meaning of isochores—‘iso’ or equality of windowed GC%.

5. Web references

http://bioinfo2.ugr.es/isochores/: an online resource on isochore mapping.

http://www.nature.com/nature/journal/v409/n6822/suppinfo/409860a0.html: web supplement material for Lander et al. (2001).

http://www.nslij-genetics.org/dnacorr/: an online bibliography on features, patterns, correlations in DNA and protein texts.

Acknowledgements

We would like to acknowledge the financial support from the 5th Anton Dohrn Workshop at Ischia (2001) where some of the ideas presented here were discussed. W.L. acknowledges partial support from NIH contract N01-AR12256, P.B.G., P.C. and J.L.O. acknowledge the grant support BIO99-0651-CO2-01 and BFM2002-00183 from the Spanish Government. W.L. thanks Ivo Grosse for extensive discussions on the topic.

References

Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Physical Review E 53, 5181–5189.

Bernaola-Galván, P, Oliver, JL, Carpena, P, Clay, O, Bernardi, G, 2003. Intragenomic heterogeneity in prokaryotic genomes, Gene, submitted for publication.

Bernardi, G., 1993. The isochore organization of the human genome and its evolutionary history—a review. Gene 135, 57–66.

Bernardi, G., 1995. The human genome: organization and evolutionary history. Annual Review of Genetics 23, 637–661.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Bernardi, G., 2001. Misunderstandings about isochores. Part 1. Gene 276, 3–13.

Bettecken, T., Aissani, B., Müller, C.R., Bernardi, G., 1992. Compositional mapping of the human dystrophin-encoding gene. Gene 122, 329–335.

Clay, O., Carels, N., Douady, C., Macaya, G., Bernardi, G., 2001. Compositional heterogeneity within and among isochores in mammalian genomes. I. CsCl and sequence analyses. Gene 276, 15–24.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes. I. preparation, basic properties and compositional heterogeneity. European Journal of Biochemistry 115, 227–233.

Fukagawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H., Ikemura, T., 1995. A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoau-tosomal boundary-like sequence exists near the boundary. Genomics 25, 184–191.

Fukagawa, T., Nakamura, Y., Okumura, K., Nogami, M., Ando, A., Inoko, H., Saito, N., Ikemura, T., 1996. Human pseudoautosomal boundary-like sequences: expression and involvement in evolutionary formation of the present-day pseudoautosomal boundary of human sex chromosomes. Human Molecular Genetics 5, 23–32.

International Human Genome Sequencing Consortium, Lander, E.S., Waterston, R.H., Sulston, J., Collins, F.S., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, W., 1997. The study of correlation structures of DNA sequences—a critical review. Computer. and Chemistry 21, 257–271 (special issue on open problems of computational molecular biology).

Li, W., 2001. Delineating relative homogeneous G+C domains in DNA sequences. Gene 276, 57–72.

Li, W., Marr, T., Kaneko, K., 1994. Understanding long-range correlations in DNA sequences. Physica D 75, 392–416.

Li, W., Stolovitzky, G., Bernaola-Galván, P., Oliver, J.L., 1998. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Genome Research 8, 916–928.

Oliver, J.L., Li, W., 1998. Quantitative analysis of compositional heterogeneity in long DNA sequences: the two-level segmentation test (abstract), Genome Mapping, Sequencing and Biology, Cold Spring Harbor Laboratory, pp. 163.

Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.

Stephens, R., Horton, R., Humphray, S., Rowen, L., Trwosdale, J., Beck, S., 1999. Gene organisation, sequence variation and isochore structure at the centrometric boundary of the human MHC. Journal of Molecular Biology 291, 789–799.

Sueoka, N., 1959. A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proceedings of the National Academy of Sciences 45, 1480–1490.

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proceedings of the National Academy of Sciences 48 (4), 582–592.

The MHC sequencing consortium, Beck, S, Geraghty, D, Inoko, H, Rowen, L, et al., 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature 401, 921–923.