Deposited research article

Isochores Merit the Prefix 'Iso'

Wentian Li[1], Pedro Bernaola-Galvan[2], Pedro Carpena[2] and Jose L. Oliver[3]

Addresses: [1] Center for Genomics and Human Genetics, North Shore - LIJ Research Institute, 350 Community Drive, Manhasset, NY 11030, USA. [2] Departmenta de Fisica Aplicada II, Universidad de Malaga, E-39071, Malaga, Spain. [3] Departmento de Genetica, Instituto de Biotechnologia, Universidad de Granada, E-18071 Granada, Spain.

Correspondence: Wentian Li. E-mail: wli@nslij-genetics.org

Isochores Merit the Prefix 'Iso'

Wentian Li[1,*], Pedro Bernaola-Galván[2], Pedro Carpena[2], Jose L. Oliver[3]

1. Center for Genomics and Human Genetics, North Shore - LIJ Research Institute

350 Community Drive, Manhasset, NY 11030, USA

2. Departmenta de Física Aplicada II, Universidad de Málaga, E-39071, Málaga, Spain

3. Departmento de Genética, Instituto de Biotechnología, Universidad de Granada,

E-18071 Granada, Spain

Email addresses: wli@nslij-genetics.org, rick@uma.es, pcarpena@ctima.uma.es, oliver@ugr.es.

* The corresponding author.

Running title: isochores merit iso

Abbreviations and acronyms:

ANOVA: analysis of variance; IHGSC: international human genome sequencing consortium; MHC: major histocompatibility complex

Abstract

The isochore concept in human genome sequence was challenged in an analysis by the *International Human Genome Sequencing Consortium* (IHGSC). We argue here that a statement in IGHSC analysis concerning the existence of isochore is incorrect, because it had applied an inappropriate statistical test. To test the existence of isochores should be equivalent to a test of homogeneity of windowed GC%. The statistical test applied in the IHGSC's analysis, the binomial test, is however a test of a sequence being random on the base level. For testing the existence of isochore, or homogeneity in GC%, we propose to use another statistical test: the analysis of variance (ANOVA). It can be shown that DNA sequences that are rejected by binomial test may not be rejected by the ANOVA test.

Background

The degree of homogeneity in base composition in human genome is a fundamental property of the genome sequence. Not only does it characterize the organization and evolution of the genome, but also it provides a context of many practical sequence analysis. Statistical quantities such as GC%, used for sequence analyses 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 a sequence analysis such as 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 compositional homogeneous regions in human genome with sizes of at least 200-300 kb [1, 2]. These homogeneous regions are called "isochores" [3], and the whole genome is a mosaic of isochores. Recently, however, this view of human genome is questioned in an initial analysis of human genome draft sequence [4]. The analysis presumably 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 [4], and a stunning statement was made that, essentially, isochore concept does not hold ("... isochores do not appear to merit the prefix 'iso'", page 877 of [4]).

The purpose of this Letter is to show that an incorrect statistical distribution for windowed GC% is assumed in [4], based on an unrealistic condition for DNA sequences. As a result, the statistical test used in [4] is invalid. We will present a correct statistical test, assuming a more reasonable statistical distribution of windowed GC%. Under the new test, the conclusion concerning the existence of isochore is drastically altered. Although our testing result may still depend on the window size at which GC% is sampled, and may possibly depend on the choice of GC% groups, it is clear that the test in [4] is too biased towards rejecting the homogeneity null hypothesis, and sequences that fail the test in [4] usually do not fail our new test.

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 on 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. Generally speaking, the mean and the variance are two independent parameters of a statistical distribution. However, for the homogeneity test in [4], the variance is assumed to be a function of the mean, thus it is not independently estimated.

In [4], the windowed GC% is assumed to follow a binomial distribution. For a binomial distribution to be true, bases within the window should be uncorrelated, similar to tossing a coin many times. Violating this assumption invalids the use of binomial application. The

more reasonable statistical distribution of GC% should be the normal distribution which, unlike the binomial distribution, has two independent parameters (mean and variance). Mean value can be estimated from a window, whereas variance can be estimated from a group of windows.

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 [5, 6, 7, 8]), with lengths 642.1 kb and 900.9 kb, respectively. The exact borders of the two isochore sequences are determined by a segmentation procedure [9, 10] and an online resource on isochore mapping [11]). We first repeat the test in [4] that these two sequences, when viewed as a collection of many 20 kb windows, are sampled from a binomial distribution. According to [4], a rejection of this test is considered to be an evidence for heterogeneity. 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 close to be 0 ($< 10^{-50}$).

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 [12, 13, 3, 14], [15] (and the 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 20-kb window is not a series of 20000 uncorrelated bases; it is not a rejection of homogeneity of windowed GC% along the sequence.

To reaffirm our belief that the binomial test used in [4] is a test of randomness of the sequence instead of homogeneity, one bacterial 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) are used for test. 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 *Borrelia burgdorferi*, a particularly homogeneous genome, as shown in a

recent survey of archaeal and bacterial genome heterogeneity [16].

We would like to suggest that the more reasonable statistical distribution of windowed GC% is the normal/Gaussian distribution, and the more appropriate test of homogeneity of these GC% values along a sequence is the analysis of variance (ANOVA). There are at least two reasons to believe that ANOVA is the more appropriate test. First, it is a test of equality between means, which is identical to the intuitive meaning of homogeneity, i.e., GC% are the same along the sequence. Second, ANOVA and normal distribution reflects the real situation of DNA sequences that these are not random sequences, and windowed GC%'s exhibit higher values of variances. ANOVA allows the variance to be estimated from the data, rather than being fixed by the mean value as in binomial distribution. ANOVA was previously applied to the study of inter-chromosomal homogeneity of yeast genome [14, 17].

To apply ANOVA 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%'s in each super-window is the same. The simplest selection of super-windows and windows is to assume all windows to have the same length. To match the discussions in [4], we choose 20-kb windows and 300-kb super-windows. This corresponds to roughly 2 super-windows, 16 windows per super-window for the MHC class III sequence, and 3 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 and 0.323, respectively, for MHC class III and class II sequence. The null hypothesis, that means of GC% in different super-windows are the same, is not rejected.

When the ANOVA test is applied to the *Borrelia burgdorferi* genome sequence and two randomly generated sequences, null hypothesis cannot be rejected, indicating that all three sequences are homogeneous at the respective window and super-window sizes (20 kb

and 300 kb). This is a more satisfactory situation than the binomial test because now a homogeneous bacterial sequence is indeed confirmed to be homogeneous by the test.

Discussions

Due to the "domains within domains" phenomenon in DNA sequences [18, 19, 20], we should not assume automatically that a homogeneity test result obtained at 20-kb window and 300-kb super-window will hold true for other window and super-window sizes. To check this, we carry 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 $(-\log_{10}(p-\text{value}))$ for window sizes of around 20 kb, 10 kb, 5 kb, and 2.5 kb, and the sequence is partitioned into 2, 3, 5, 8 (2,3,5,9) super-windows for MHC class III (II) sequence.

Several observations could be made from Fig.1. First, when GC%'s 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 GC% values (taken from a fixed window size) having the similar value, regrouping these values does not make an insignificant result to be significant.

Second, the ANOVA test becomes more significant when the window size decreases. This observation is understandable because at smaller length scales, GC% fluctuations are no longer averaged out. These smaller-length-scale fluctuations could be due to repeats, insertions, foreign elements, etc. For MHC class II sequence, as the subwindow size is reduced to around 2.5 kb, the ANOVA test result is typically significant (Fig.1). This is consistent to the definition of isochores as "fairly homogeneous" (as versus "strictly homogeneous") segments above a size of 3 kb [21, 22], and justifies the "coarse graining" procedure to locate isochore boundaries in [9].

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 MHC class II sequence is more heterogeneous than MHC class III sequence when viewed at the 2-10 kb length scales. It is known that GC-poor sequences are generally considered to be 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% [13, 3, 15]. Since the GC% of MHC class III and II sequence is 51.9% and 41.1%, respectively, we might expect MHC class II sequence to be more homogeneous than class III sequence. Interestingly, Fig.1 shows the contrary.

To conclude, the binomial test used in [4] should not be a test of homogeneity if the expected variance does not reflect the true variance in the sequence. The reason that the expected variance in a binomial test (which is derived from the mean GC% instead of being an independent parameter) is unrealistic is because the underlying base sequence is not random/uncorrelated. We are naturally led to the ANOVA test if we actually estimate the variance from the data. With ANOVA tests, it is clear that homogeneous regions of GC% in human genome do exist; in other words, isochores exist.

Methods

Binomial test: Following [23], 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. The variance of these GC%'s (σ^2) is calculated, and the variance as expected from a binomial distribution is $\sigma_0^2 = m(1-m)/20000$, where m is probability of G or C. The value of m can be estimated by the actual GC% of the sequence. The test statistic is $c^2 = (n-1)\sigma^2/\sigma_0^2$. For null hypothesis (that windowed GC% measurements do follow binomial distribution, which is true when the underlying base sequence is random/uncorrelated within the window), c^2

follows the $\chi^2_{df=n-1}$ distribution (e.g. $\chi^2_{df=44}$ in our example). For any given c^2 value, the p-value can be determined by the corresponding χ^2 distribution.

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

Acknowledgments

References

[1] Bernardi G: **The isochore organization of the human genome and its evolutionary history–a review**, Gene 1993, 135:57-66.

[2] Bernardi G: **The human genome: organization and evolutionary history**, Annual Review of Genetics 1995, 23:637-661.

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

[4] Lander ES, Waterston RH, Sulston J, Collins FS, et al. (International human genome sequencing consortium): **Initial sequencing and analysis of the human genome**, Nature 2001, 409:860-921.

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

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

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

[8] Beck S, Geraghty D, Inoko H, Rowen L, et al. (The MHC sequencing consortium): **Complete sequence and gene map of a human major histocompatibility complex**, Nature 1999, 401:921-923.

[9] Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R: **Isochore chromosome maps of eukaryotic genomes**, Gene 2001, 276:47-56.

[10] Li W: **Delineating relative homogeneous G+C domains in DNA sequences**, Gene 2001, 276:57-72.

[11] **An online resource on isochore mapping** : [http://bioinfo2.ugr.es/isochores/]

[12] Sueoka N: **A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation**, Proceedings of the National Academy of Sciences 1959, 45:1480-1490.

[13] Sueoka N: **On the genetic basis of variation and heterogeneity of DNA base composition**, Proceedings of the National Academy of Sciences 1962, 48(4):582-592.

[14] Li W, Stolovitzky G, Bernaola-Galván P, Oliver JL: **Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes**, Genome Research 1998, 8:916-928.

[15] Clay O, Carel N, Douady C, Macaya G, Bernardi G: **Compositional heterogeneity within and among isochores in mammalian genomes. I. CsCl and sequence analyses**, Gene 2001, 276:15-24.

[16] Bernaola-Galván, P, Oliver JL, Carpena P, Clay O, Bernardi G: **Intragenomic heterogeneity in prokaryotic genomes**, Gene 2002, submitted.

[17] Oliver JL, Li W: **Quantitative analysis of compositional heterogeneity in long DNA sequences: the two-level segmentation test** (abstract), Genome Mapping, Sequencing & Biology (Cold Spring Harbor Laboratory) 1998, page 163.

[18] Li W, Marr T, Kaneko K: **Understanding long-range correlations in DNA sequences**, Physica D 1994, 75:392-416.

[19] Bernaola-Galván P, Román-Roldán R, Oliver JL : **Compositional segmentation and long-range fractal correlations in DNA sequences**, Physical Review E 1996, 53:5181-5189.

[20] Li W: **The study of correlation structures of DNA sequences – a critical review**, Computer & Chemistry (special issue on open problems of computational molecular biology) 1997, 21:257-271.

[21] Bettecken Th, Aissani B, Müller CR, Bernardi G: **Compositional mapping of the human dystrophin-encoding gene**, Gene 1992, 122:329-335.

[22] Bernardi G: **Isochores and the evolutionary genomics of vertebrates**, Gene 2000, 241:3-17.

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

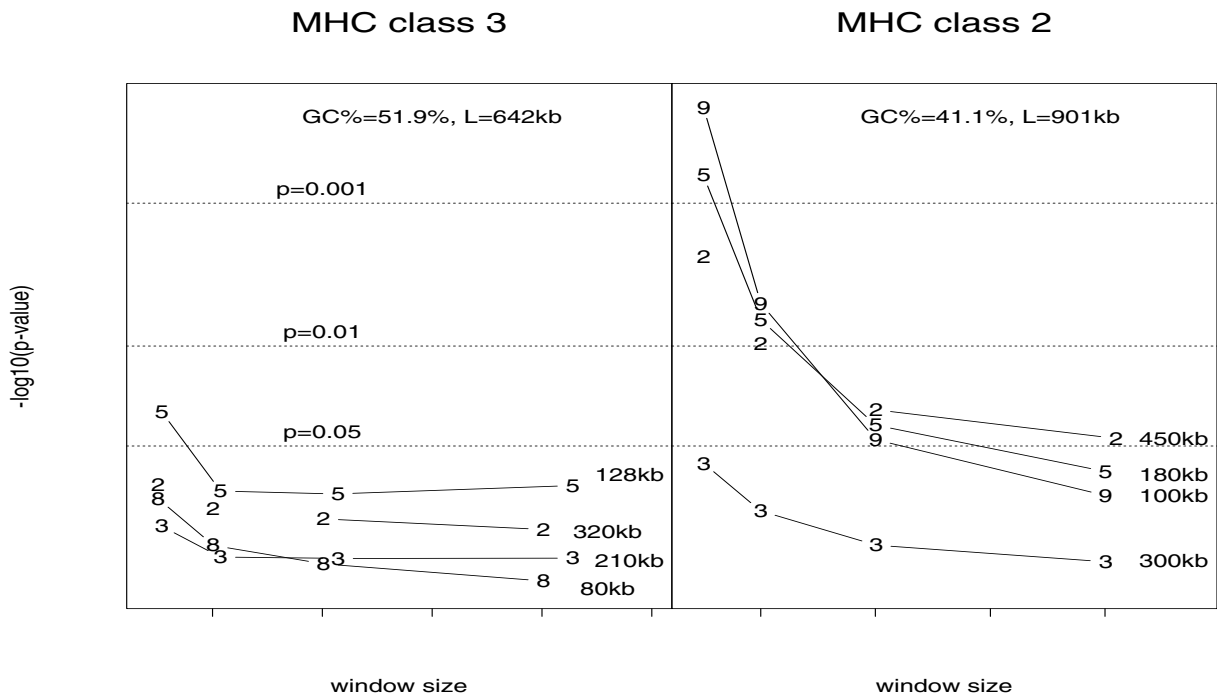## MHC class 3

MHC class 2

Figure 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.

seq	# 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**

Table 1: **Testing the hypothesis that GC% values sampled from 20-kb windows follow a binomial distribution**. Five sequences are tested: MHC class III and MHC class II isochore sequences, two random sequences similar these two MHC sequences (same length and same base composition), and bacterium *Borrelia burgdorferi* genome sequence. Detailed explanation of column headers: 1. Sequence name. 2. Total number of windows in the sequence (n), with 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 ($\sigma_0^2 = m(1-m)/20000$). 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.

	df	SS	MS	F-value	*p*-value
MHC class III (sw=2, w=16)					
between windows	1	0.0009159	0.0009159	1.781	*0.192*
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*
within windows	42	0.02997	0.0007137		
random seq similar to class III (sw=2, w=16)					
between windows	1	0.00000288	0.00000288	0.247	*0.623*
within windows	30	0.0003496	0.00001165		
random seq similar to class II (sw=3, w=15)					
between windows	2	0.00004546	0.00002273	1.884	*0.165*
within windows	42	0.0005066	0.00001206		
B. burgdorferi (sw=3, w=15)					
between windows	2	0.0002064	0.0001032	0.671	*0.517*
within windows	42	0.006461	0.0001538		

Table 2: **ANOVA test results of the five sequences** (two MHC isochore sequences and their randomized sequences, and bacterium *Borrelia burgdorferi* sequence). df: degrees of freedom; SS: sum of squares. MS: mean squares. F-value: test statistic value; p-value: p-value from the ANOVA test. sw and w are the number of super-windows and windows.