

# ARE ISOCHORE SEQUENCES HOMOGENEOUS?

5th Anton Dohrn Workshop on Natural Selection and the Neutral Theory  
October 24-27, 2001.

*Wentian Li, Ph.D*

Center for Genomics and Human Genetics  
North Shore University Hospital  
Manhasset, NY 11030

web: <http://linkage.rockefeller.edu/wli/>  
email: [wli@linkage.rockefeller.edu](mailto:wli@linkage.rockefeller.edu), [wli@nshs.edu](mailto:wli@nshs.edu)

From *International Human Genome Sequencing Consortium* (Nature'01)

“We studied the draft genome sequence to see whether strict isochores could be identified... sequence was divided into 300-kb windows, and each window was subdivided into **20-kb sub-windows**, and investigated how much of the variance in the GC content of subwindows across the genome can be statistically ‘explained’ by the average GC content in each window. About 3/4 of the genome-wide variance among 20-kb windows can be statistically explained by the average GC content of 300-kb windows that contain them, but the residual variance

among subwindows is still far too large to be consistent with a homogeneous distribution. In fact, **the hypothesis of homogeneity could be rejected** for each 300-kb window in the draft genome sequence...These results rule out a strict notion of isochore as compositionally homogeneous. Instead, **there is substantial variation at many different scales...** Although isochores do not appear to merit the prefix 'iso', the genome clearly does contain large regions of distinctive GC content and it is likely to be worth redefining the concept so that **it becomes possible rigorously to partition the genome into regions.**"

#1. It seems to refer to an analysis of variance (ANOVA) where the variance within a group and variance among groups are compared. But **the definition of “groups”, i.e. 20-kb subwindows, is arbitrary.** There is also no reason to insist on **using bases as “members”.**

What if a 300-kb window is divided into only three 100-kb subwindows? And each 50-kb sequence as a member ?

**#2. The null hypothesis of homogeneous random sequence can be compared to the alternative hypothesis of a “change” of the base composition at some point.** The null can be easily rejected for a typical long DNA sequence at the usual significance levels (e.g. 0.01, 0.001,...).

#3. It had been known for some time that there are statistical variations at different length scales in DNA sequences.

We called it **“domains within domains”, “hierarchical patterns”, “fractal-like structure”, “1/f noise”, “long-range correlations”, “long-memory process”...**

#4. We knew how to rigorously partition a DNA sequence into relatively homogeneous domains.

These methods are under the names of “**segmentation**”, “**change-point analysis**”,...

Three “sure” isochores are used to illustrate these comments.

\*class III of MHC (human ch6, 642 kb)

\*class II of MHC (human ch6, 901 kb)

\* contig on human ch21 (7.104 Mb)

# OUTLINE

1. How these three isochore sequences are delineated?

→ **segmentation algorithm**

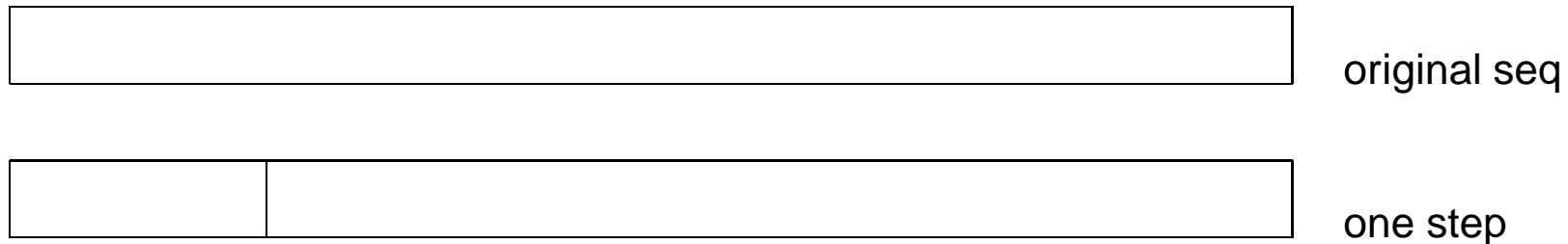
2. Do these isochore sequences have internal structures? →

**more segmentations**

3. ANOVA → **two-level test**

4. “Long-range correlation” → **spectral analysis and 1/f noise**

# Introduction of One-Step Segmentation



It is a hypothesis test (likelihood-ratio test) that compares two-subsequence likelihood ( $L_2$ ) and the one-sequence likelihood ( $L_1$ ).

$$L_1 = p_{G+C}^{N_{G+C}} p_{A+T}^{N_{A+T}} \quad L_2 = p_{G+C,l}^{N_{G+C,l}} p_{A+T,l}^{N_{A+T,l}} p_{G+C,r}^{N_{G+C,r}} p_{A+T,r}^{N_{A+T,r}}$$

Maximizing two likelihoods over parameters at a fixed segmentation point (change-point, partition point):

$$\hat{L}_1 \equiv \max_{p_{A+T}, p_{G+T}} L_1(p_{A+T}, p_{G+T}) = \left( \frac{N_{A+T}}{N} \right)^{N_{A+T}} \left( \frac{N_{G+C}}{N} \right)^{N_{G+C}}$$

$$\hat{L}_2 \equiv \max_{p_{A+T,l}, p_{G+T,l}, \dots} L_2(p_{A+T,l}, p_{G+T,l}, p_{A+T,r}, p_{G+T,r}) = \dots$$

It can be shown that

$$LLR(i) \equiv 2 \log \frac{\hat{L}_2}{\hat{L}_1} = 2N \left( \hat{H} - \frac{N_l}{N} \hat{H}_l - \frac{N_r}{N} \hat{H}_r \right) \equiv 2N \cdot \hat{D}_{JS}$$

where  $H$ ,  $H_l$ ,  $H_r$  are the entropies of the whole, left, and right sequences;  $D_{JS}$  is termed Jensen-Shannon distance.

**For a fixed change-point,  $2 \log \hat{L}_2 / \hat{L}_1$  follows  $\chi^2$  distribution with  $2 - 1 = 1$  degrees of freedom under null hypothesis.**

**For all possible change-points, we must maximize the position also:**

$$LLR \equiv \max_i 2 \log \hat{L}_2 / \hat{L}_1$$

LLR is an **extreme value**, whose mean increases with the sequence length  $N$  ( $\sqrt{N}$ ?,  $\log \log(N)$ ?  $\log(N)$ ?).

- $\sqrt{N}$ : random walk
- $\log \log(N)$ : Horvath, J Multivariate Analysis, 31:148-159 (1989)
- $\log(N)$ : Bayesian information criterion in model selection

## **Model Selection: An Alternative to Hypothesis Testing**

Considering **two statistical models** of a DNA sequence: 1. it's a homogeneous random sequence; 2. it contains two homogeneous subsequences each with a different (C+G)% (either at a fixed border, or at an unknown border)

Model selection methods uses some criterion that balances the **data-fitting performance** and the **model complexity**. It avoids both **over-fitting** and **under-fitting**.

## One Model Selection Technique: Bayesian Information Criterion (BIC)

**data-fitting performance:** maximized likelihood  $\hat{L}$

**model complexity:** num of free parameters in the model  $K$

**definition of BIC:**  $BIC = -2 \log(\hat{L}) + \log(N) \cdot K$

- the smaller the BIC, the better the model
- difference of two BIC's is an approximation of the (-2log)

**Bayes factor**

**BIC** of one homogeneous random sequence (GC vs AT binary sequence) is:  $BIC_1 = -2 \log(\hat{L}_1) + \log(N) \cdot 1$

**BIC** of two homogeneous subsequence with a fixed border is:  $BIC_2 = -2 \log(\hat{L}_2) + \log(N) \cdot 2$

**BIC** of two homogeneous subsequence with an unknown border is:  $BIC'_2 = -2 \log(\hat{L}_2) + \log(N) \cdot 3(?)$

Footnote: The border position is not exactly the same as other free parameters (e.g. (C+G)%) because it's discrete, and whose range increases with  $N$ . This is more like a index for a group of models, i.e. a parameter in model of models. The value 3 in  $BIC'_2$  might be altered to 4 or something else.

---

(A) If  $BIC_1 < BIC_2$  (or  $BIC'_2$ ), the sequence is homogeneous

(B) If  $BIC_1 > BIC_2$  (or  $BIC'_2$ ), the sequence contains two domains

---

(A) is equivalent to  $2 \log \hat{L}_2 / \hat{L}_1 < \log(N)$  (or  $2 \log(N)$ ); or  $2N\hat{D}_{JS} < \log(N)$  (or  $2 \log(N)$ )

(B) is equivalent to  $2 \log \hat{L}_2 / \hat{L}_1 > \log(N)$  (or  $2 \log(N)$ ); or  $2N\hat{D}_{JS} > \log(N)$  (or  $2 \log(N)$ )

## How inhomogeneous?

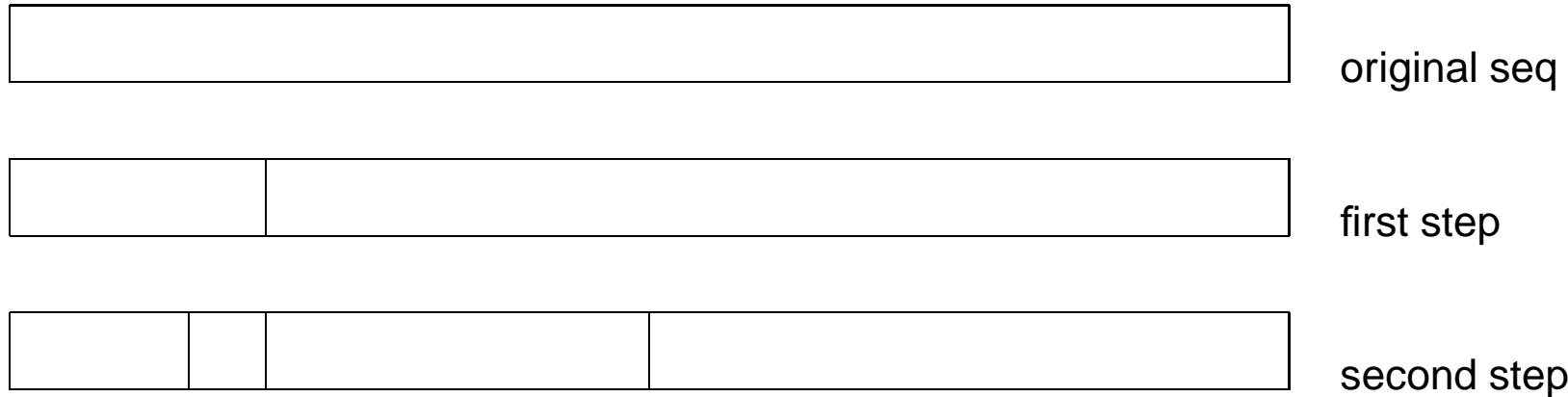
If  $2N\hat{D}_{JS}$  is larger than  $2\log(N)$ , how much larger? We can use the percentage increase from the threshold as a “segmentation strength”:

$$s = \frac{2N\hat{D}_{JS} - 2\log(N)}{2\log(N)}$$

(for binary sequences).

Footnotes: 1. the 2 in  $\log(N)$  might be changed to 3 or larger; 2. the  $\log(N)$  penalty is larger than the  $\log\log(N)$  from the work by Horvath. so this condition is more stringent than the work by Bernaola-Galvan et al.

# Recursive Segmentation



Repeating the one-step segmentation on subsequences, until it could not be segmented anymore.

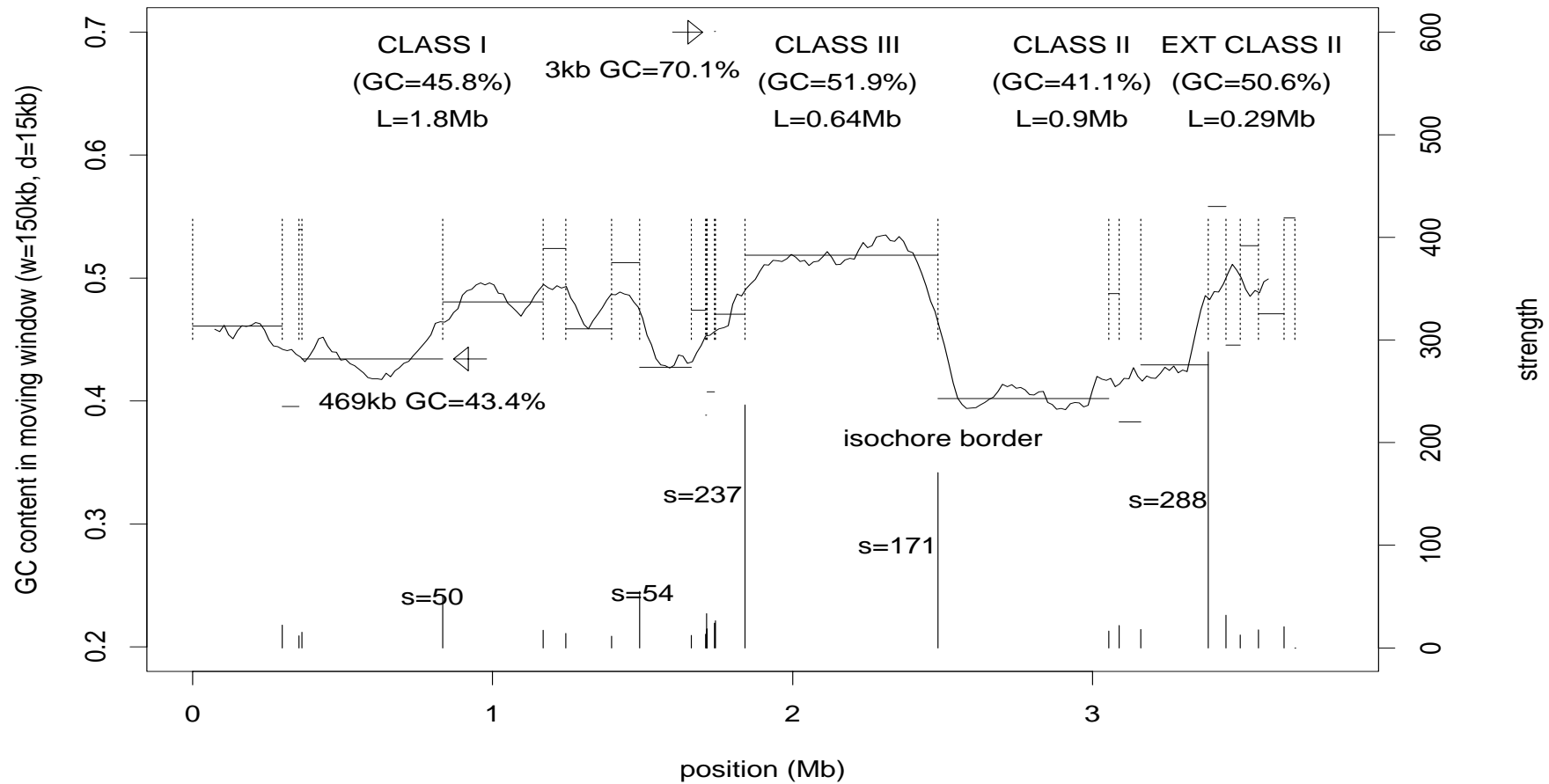
Or, sometimes, to see the “bigger picture”, stop the segmentation far before  $s$  reaches 0 ( $s > s_0 \gg 0$ ).

---

**How are the three isochore sequences  
segmented?**

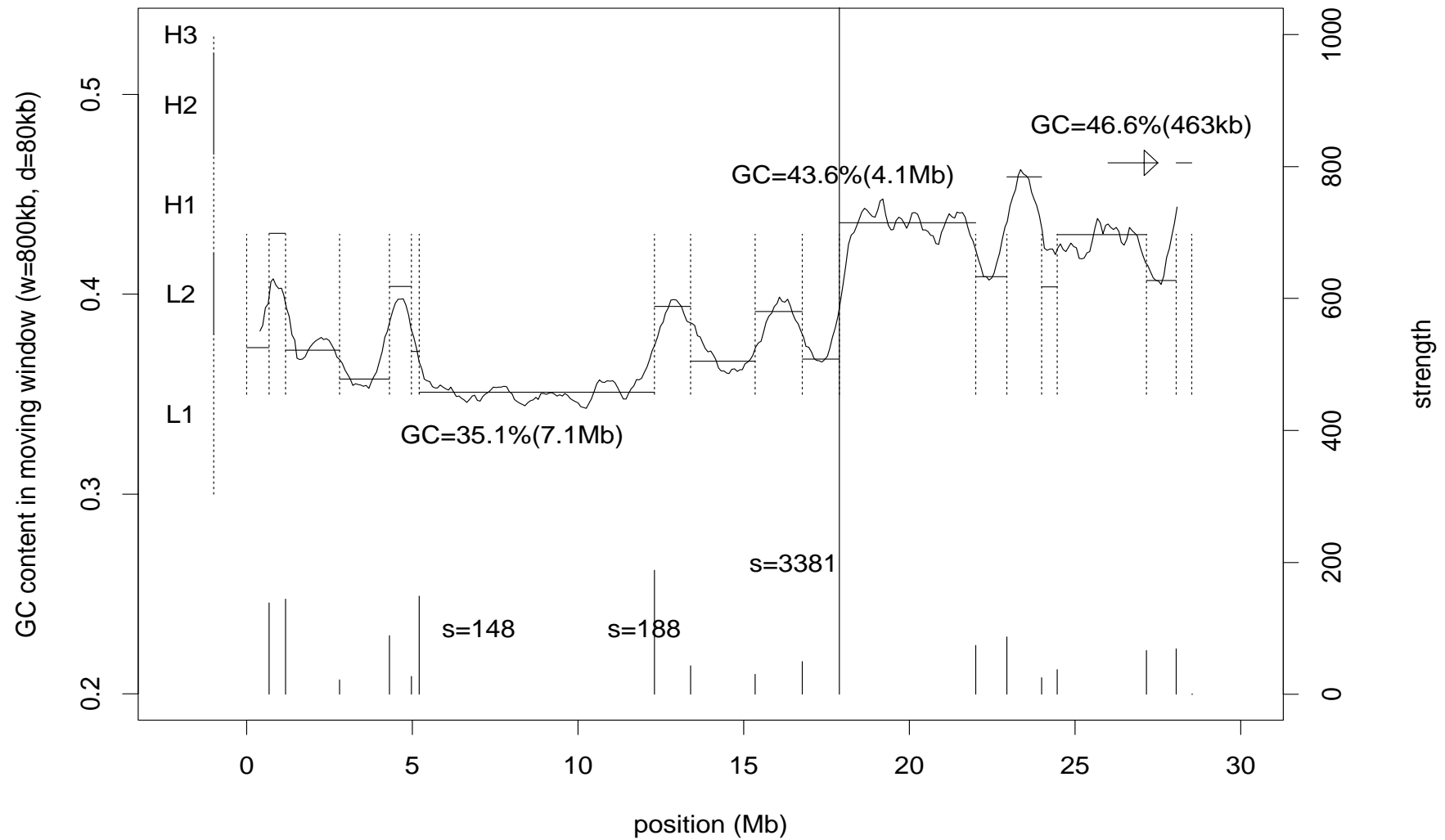
---

MHC (N=3.6Mb, s0=10, 25 domains)



The three borders have s= 237, 171, 288.

# Human 21 contig (N=28Mb, s0=20, 18 domains)



The two borders have  $s = 148, 188$ .

Summary 1: **By using segmentations with large “strength” (or high significance, or small p-values), it is possible to delineate domains that are consistent with previously known isochores. The similar strength, significance,... level can then be applied to other sequences to delineate new isochores.**

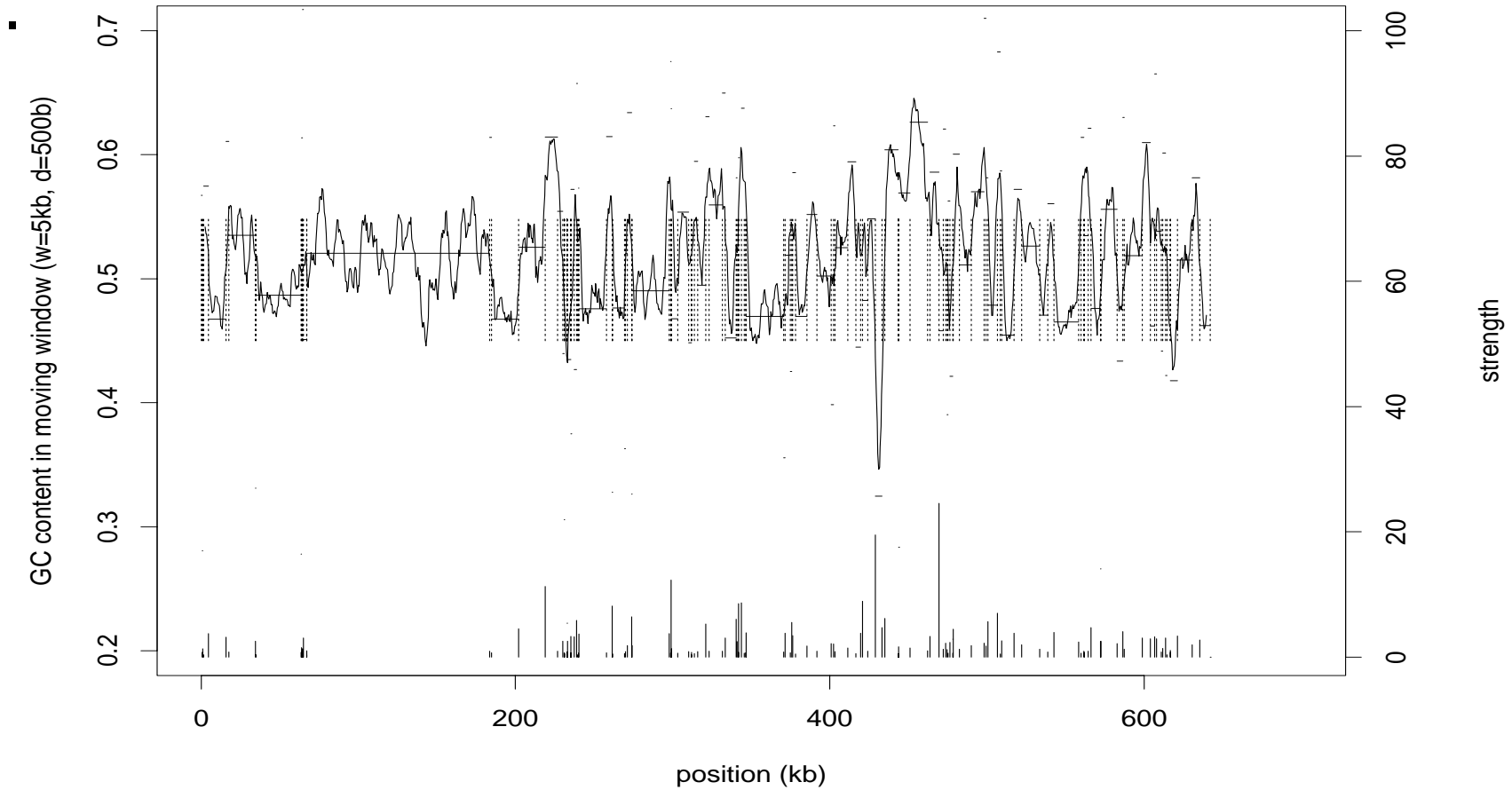
---

## Can the Three Isochore Sequences be Further Segmented?

---

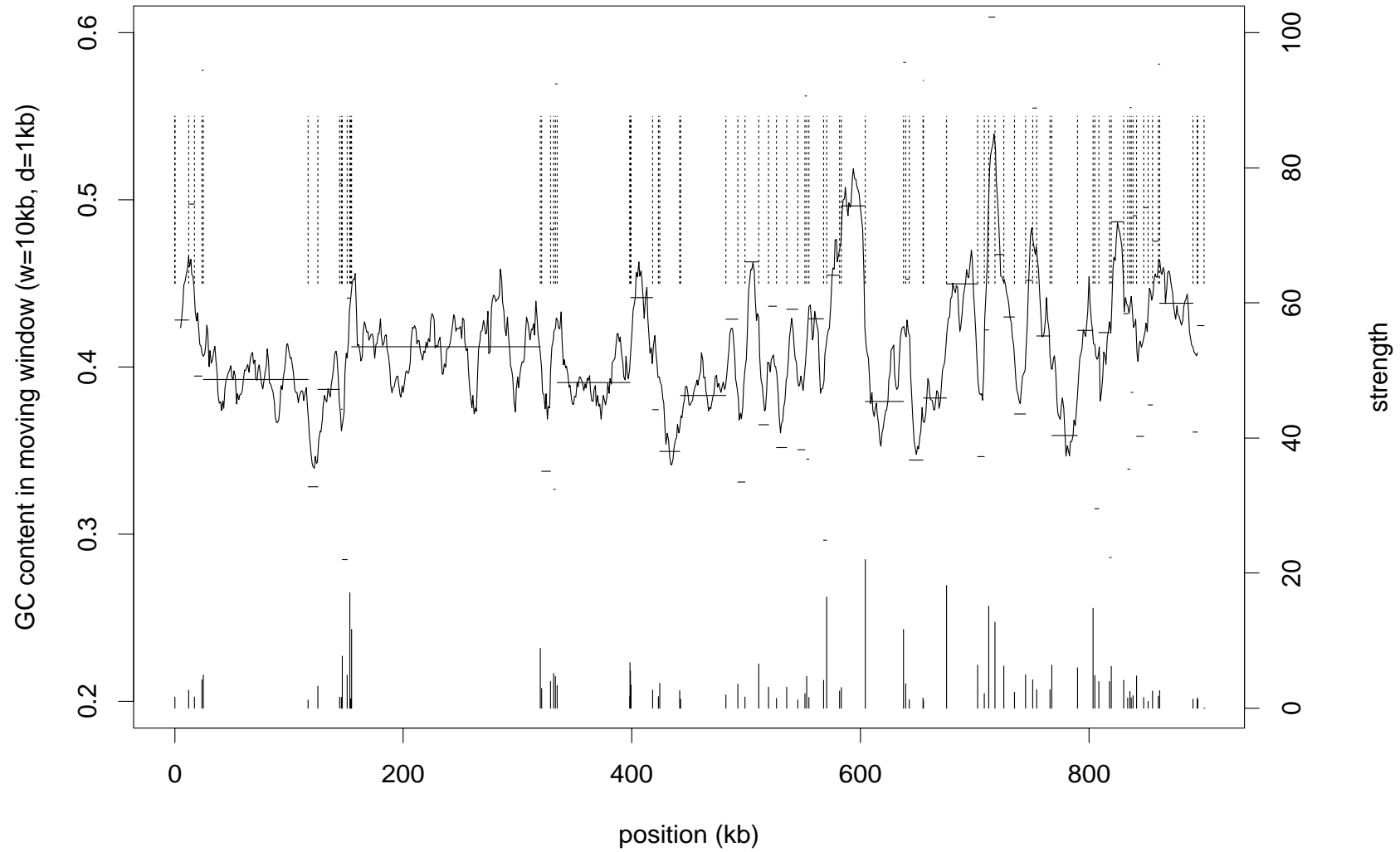
Reducing the  $s_0$  from a few hundreds (100,00%) to 1 (100%) or 0.333 (33.3%), or 0.

### MHC class3 (s0=0.333, 138 domains)



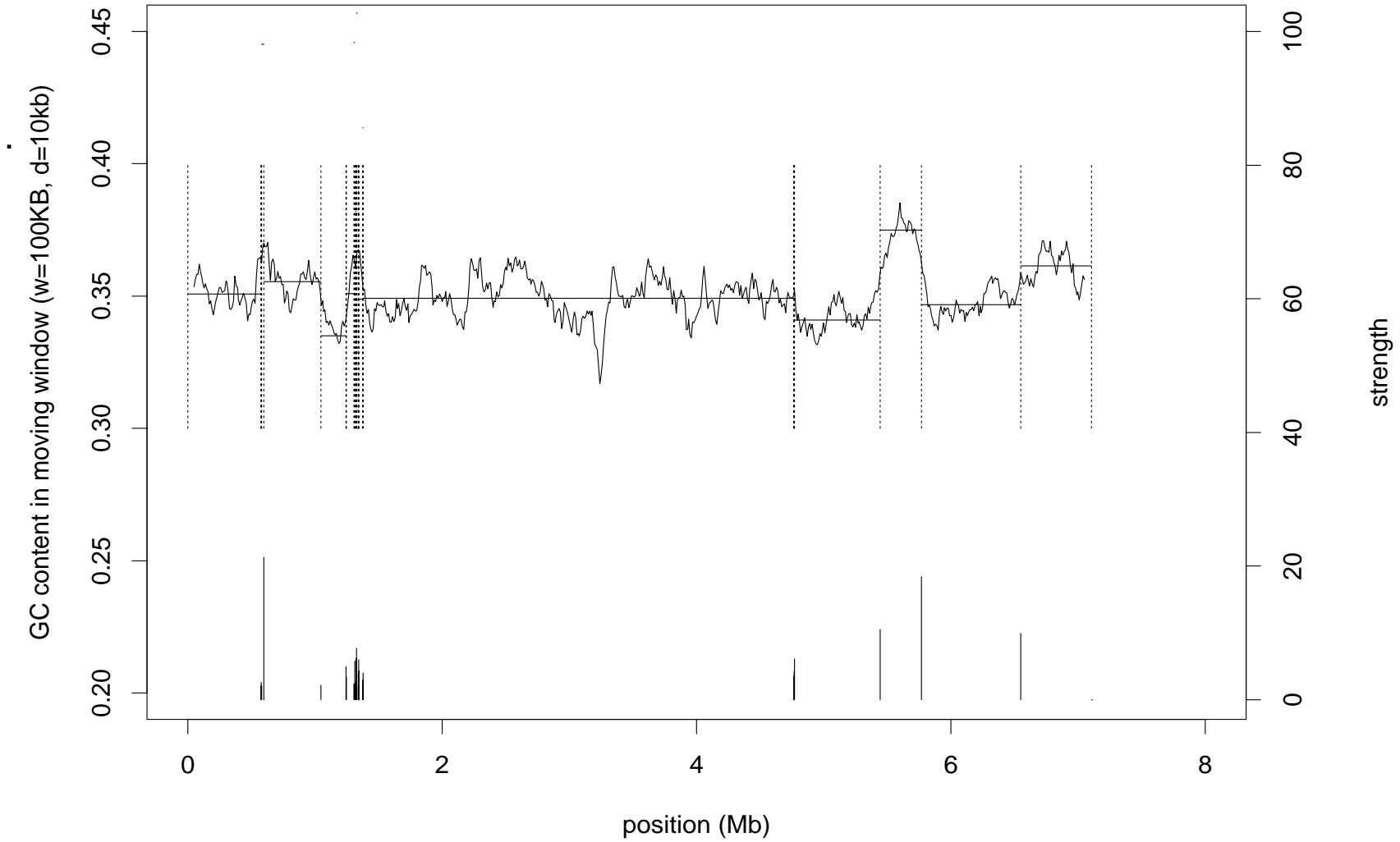
MHC class 3 (left): there are two segmentations with s as large as 20

# MHC class2 (s0=1, 83 domains)



MHC class 2 (right): there are many segmentations with s larger than 15

# human ch21 (s0=2, 26 domains)



ch21: a larger  $s_0$  value (2) is chosen for this plot because of its longer length

Summary 2: **It is clear that isochore sequences can be further segmented. In other words, in the strict sense of homogeneity on the base level, isochore sequences will not pass the test. Perhaps no long DNA sequences can ever pass the same test.**

# ANOVA test of The Three Isochore Sequences

---

Groups: *subsequences*

Members: *sub-subsequences*

Testing whether the means of each group are the same (null). If the p-value is smaller than a pre-set significance level, reject the null hypothesis.

## MHC class 3

# group	members per g	size per m	p-value
2	10	32104 bp	0.199
2	100	3210	0.10
10	10	6420	0.2834
20	10	3210	0.0424

**Just when you think various ANOVA lead to the same conclusion that MHC class 3 is homogeneous...**

# group	members per g	size per m	p-value
100	90	71	<b>0</b>

**When the member corresponds to smaller segments, heterogeneity is revealed, and the test will reject the null (homogeneity) hypothesis.**

## MHC class 2

# group	members per g	size per m	p-value
2	10	45047 bp	0.079
2	100	4504	<b>0.0069</b>
10	10	9009	0.099
20	10	4504	<b>0.0028</b>

**More members per group, or more groups, will lead to smaller segments, thus revealing possible heterogeneity at these smaller length scales.**

## human ch21 contig

# group	members per g	size per m	p-value
2	10	355201 bp	0.52
2	100	35520	0.32
10	10	71040	0.053
20	10	35520	<b>0.00065</b>

Summary 3: **For the three isochore sequences, ANOVA test may or may not reject the null (homogeneous, all group means are the same), depending on the choice of groups and members.**

One may redefine groups and members by domains obtained from recursive segmentation at two different significance levels (or two  $s$  values). [*JL Oliver, W Li (1998), "Quantitative analysis of compositional heterogeneity in long DNA sequences: the two-level segmentation test", preprint*] This method tries to maximize the difference between groups, thus is more likely to reject the null than the equal-size-domain method.

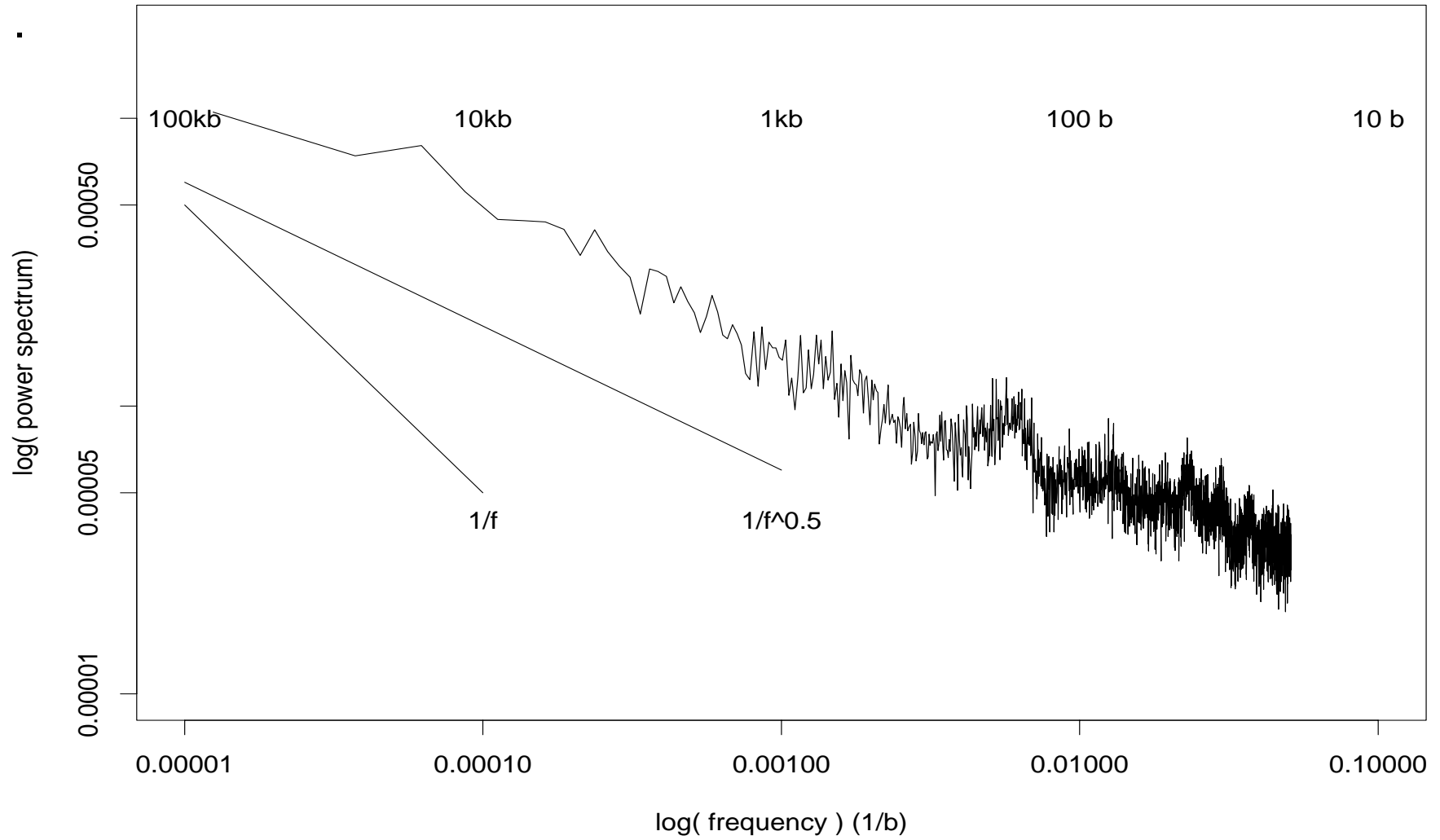
# Complex Base Composition Patterns

---

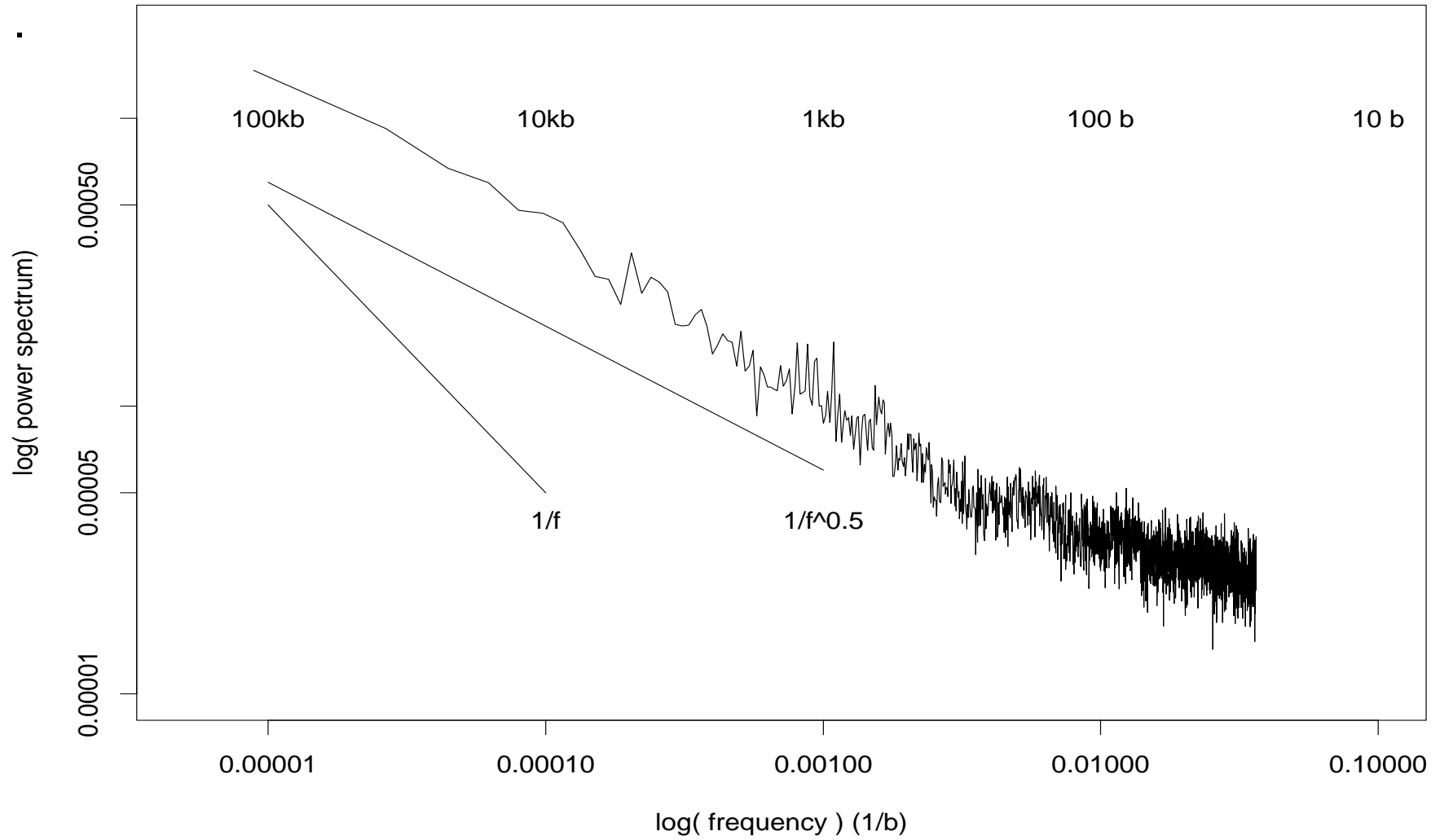
When a relatively homogeneous sequence (isochore) becomes inhomogeneous under a more stringent criterion, we call this “**domains within domains**” .

A more convenient and natural way to analyze base composition fluctuation at different scales is the **spectral analysis**. Borrowing from time series analysis, a spectrum is a decomposition of the sequence by **periodic functions of different wavelengths**.

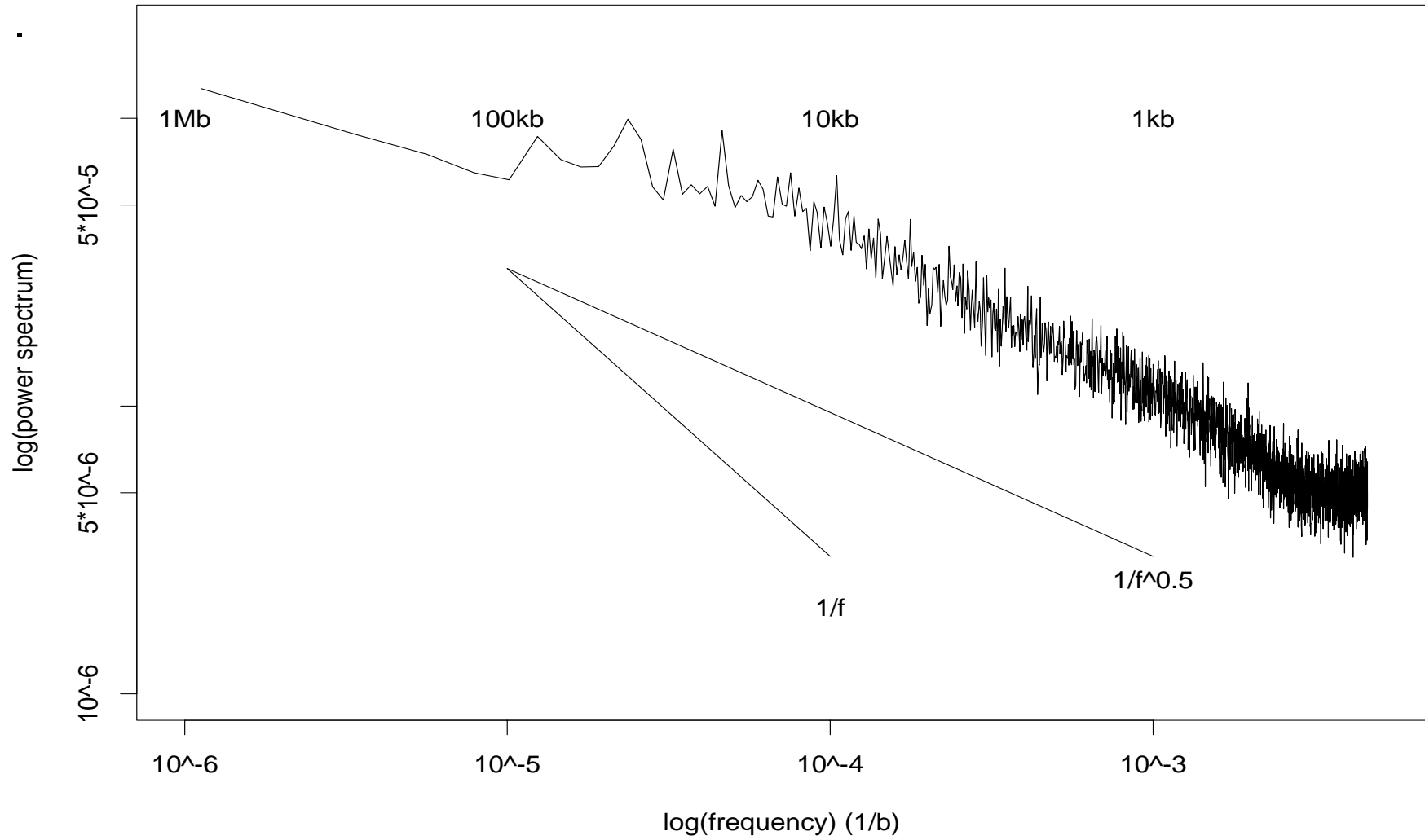
# MHC class3 (2<sup>16</sup> pts, ave 2<sup>4</sup>)



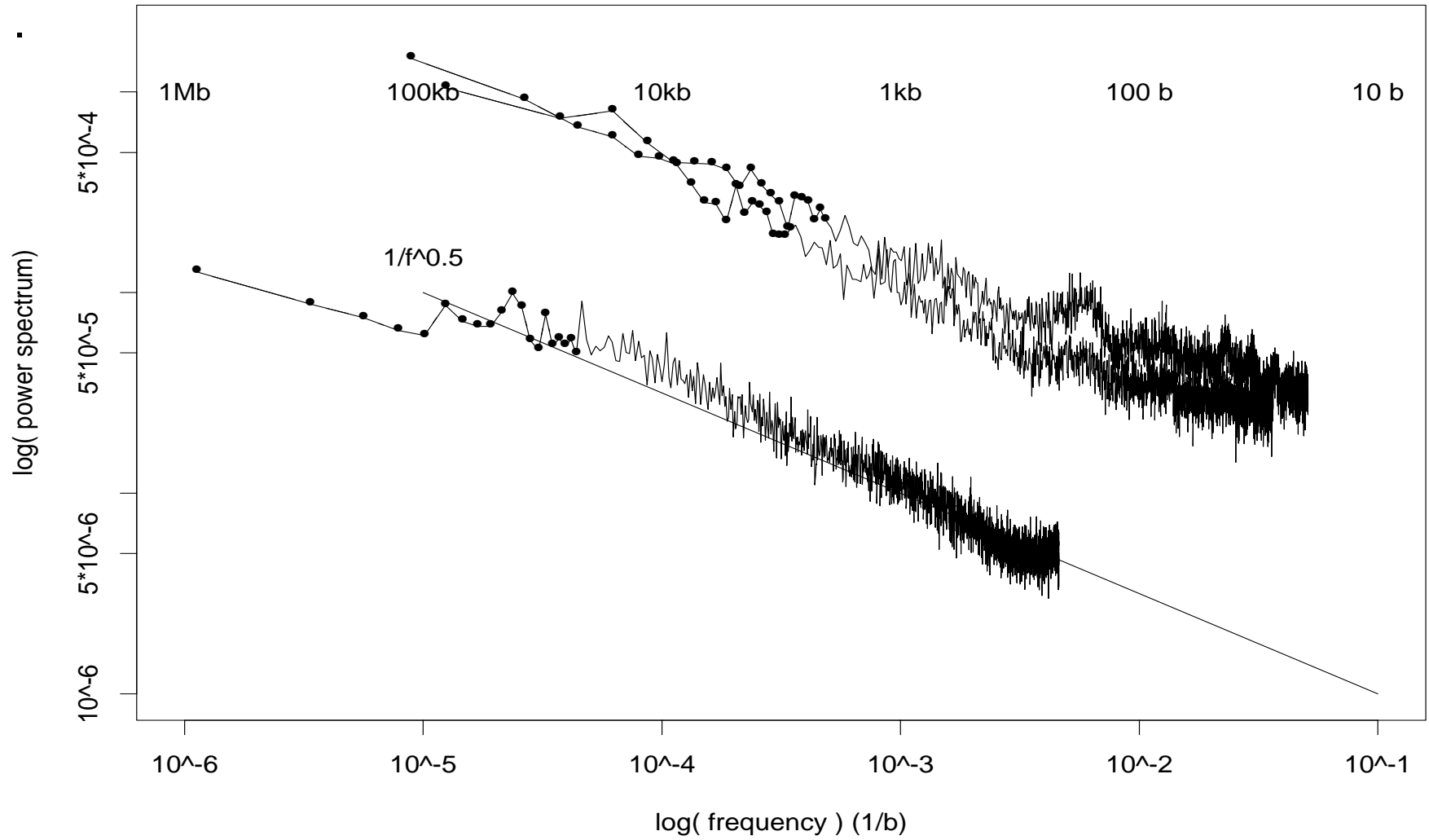
# MHC class2 (2<sup>16</sup> pts, ave 2<sup>4</sup>)



human ch21 (2<sup>16</sup> pts, ave 2<sup>4</sup>)



all three isochores (2<sup>16</sup> pts, ave 2<sup>4</sup>)



Summary 4: **Random sequences exhibit flat “white” power spectra. All three isochore sequences, however, exhibit  $1/f^\alpha$  ( $\alpha \approx 0.5$ ) “colored” or “pink” power spectra. It is the best evidence for variances/fluctuations at different length scales.**

**The alternation of isochores with different GC contents was proposed as an explanation of the observed  $1/f^\alpha$  spectra in DNA sequences.**

**Since even a single isochore to exhibit  $1/f$  spectra, it is clear that this explanation is not correct.**

**For ch21 isochore, the spectrum at  $> 100$  kb scale does flatten out a bit. Is it an indication of the homogeneity at this length scale?**

## *CONCLUSION*

The IHGSC paper did not present a state-of-art summary of our understanding on the isochore issues from the sequence analysis perspective. The domains-within-domain phenomenon is more a rule than an exception for DNA sequences. The ubiquitous  $1/f^\alpha$  power spectra in isochore sequences is a clear indication of this complex pattern. But despite of this, isochore sequences can still pass homogeneity test (e.g. by ANOVA) when the length scale, on which the homogeneity concept applies, is appropriately specified.