

ASSESSING THE COMPOSITIONAL HETEROGENEITY OF GENOMES

ROBERT H. BARAN

*Genostat, 4508 Cheltenham Drive, Bethesda, MD 20814, USA
baran@genostat.com*

The chi-squared test of homogeneity is used to formulate a gross measure of compositional heterogeneity that is scale-invariant and independent of sequence origin. When computed from the position-dependent mononucleotide compositions of 24 microbial chromosome sequences, this (base 10) logarithmic measure varies from 1.3 to 3.6, the largest values found in sequences that are clearly separated into replichores by their skew diagrams. This heterogeneity measure is readily adapted to linear functions of the base fractions such as the G+C proportion used in identifying isochores.

Keywords: comparative genomics; compositional heterogeneity; isochore.

1. Introduction

If compositional heterogeneity pervades the genome on all scales, as suggested by Karlin and Brendel (1993), then attempts to partition nucleotide sequences into homogeneous segments will give different results depending on the method chosen, even for genomes as small as bacteriophage lambda, as discovered by Braun and Muller (1998). Yet the lack of a universally effective segmentation method does not revoke the prevailing view of the genome as a patchwork or *mosaic* of compositionally dissimilar regions. The mosaic view originates with the isochore hypothesis of Bernardi (1985, 2000) in studies of warm-blooded vertebrates but has been extended to genomes in all kingdoms including bacteria and viruses (Li, Stolovitsky, Bernaola-Galvan and Oliver (1998), Bernaola-Galvan, Carpena, Roman-Roldan and Oliver (1999)). In this view, each nucleotide y_t in the complete sequence $y_1 y_2 \dots y_N$ of a chromosome can be assigned to a compositional class (or *tile type* in the mosaic) denoted x_t and distinguished by a unique conditional distribution $b(x,y) = \Pr(y_t = y \mid x_t = x)$. Since $b(x,y)$ does *not* depend on t , the nonstationary properties of the observed sequence are transferred to the hidden variable. Lacking a segmentation method (or *tiling algorithm*) to compute the sequence $x_1 x_2 \dots x_N$ of hidden variables, how can we

determine the number of tile types required to cover the complete sequence?

A related question that may be of interest to comparative genomics is how to formulate a gross measure of heterogeneity. This could seem easy because any reasonable statistical test of the null hypothesis of homogeneity produces a p-value that indicates how strongly the hypothesis is rejected. The technical obstacle is scale dependence that makes such measures sensitive to window size on which investigators have not reached a consensus (International Human Genome Sequencing Consortium (2001), Li (2002), Li, Bernaola-Galvan, Carpena, and Oliver (2003)). For example, let the sequence be partitioned into L contiguous windows of fixed size n and a remainder of length $N - nL < n$ where N is total length. If $p_j(y)$ is the (local) fraction of y -nucleotides in the j th window, and $P(y)$ is the (global) fraction for the whole sequence, the null hypothesis implies that

$$X^2(n) = \sum_{j=1}^L n \sum_{y=a,c,g,t} \frac{[p_j(y) - P(y)]^2}{P(y)} + (N - nL) \sum_{y=a,c,g,t} \frac{[p_{L+1}(y) - P(y)]^2}{P(y)} \quad (1)$$

has a central chi-squared distribution with dL degrees of freedom where $d = 3$ is one less than the number of nucleotides.¹⁰ Then X^2/L can be regarded as an average of 3 degree of freedom chi-squared statistics in single windows and it approaches d in the homogeneous case. For example, using the approximate length and global composition [A C G T] = [0.3102 0.1916 0.1899 0.3083] of the *Haemophilus influenzae* sequence (GenBank L42023), a series of 1.831×10^6 independent bases is produced by simulation. The values of X^2/L shown in the first row of Table 1 are close to d for all window sizes.

The computation of equation 1 may be clarified by the tabular representation of Figure 1 in which rows correspond to nucleotides (left margin) and columns correspond to windows that are indexed serially by j (top margin). The sum of nucleotide counts $\{n_{y,j}\}$ in each column except the last is n (bottom margin) and the local proportions are $n_{y,j}/n = p_j(y)$ in the j th window. The global proportions $P(y)$ (right) are obtained by normalizing the row sums.

Granted that that genomic nucleotide sequences are nonstationary, X^2/L can depend on the choice of sequence origin (5' end). Therefore it is necessary to displace the windows to the right (5' to 3') in increments of Δ bases, repeating the computation of equation 1 after each displacement, and averaging the results. As the last window or remainder is displaced beyond the 3' end, the sequence wraps around to its origin. Setting Δ equal to the smallest window size, 1 kilobase (kb), displacement stops after n

increments, when n is measured in kb, since this total displacement returns us to the original partition.

	1	2	3	4	5	L	$L+1$	
a	n_{a1}	n_{a2}	n_{a3}			n_{aL}		$NP(a)$
c	n_{c1}	n_{c2}	n_{c3}			n_{cL}		$NP(c)$
g	n_{g1}	n_{g2}	n_{g3}			n_{gL}		$NP(g)$
t	n_{t1}	n_{t2}	n_{t3}			n_{tL}		$NP(t)$
	n	n	n	n	n	n	$N-nL$	N

Figure 1. Tabular representation of the data in a complete sequence.

Table 1 shows average values (over all displacements) of X^2/L in windows of size $n = 1, 2, 4, \dots, 128, 256$ kb on the complete sequences of three microbial chromosomes. For the smallest size, the lowest average X^2/L is found in *Bacillus subtilis* (NC_000964) and the highest in *Plasmodium falciparum*, chromosome III (NC_000921). For $n = 128$ kb, on the other hand, *B. subtilis* is highest. Observing a maximum for the *Archaeoglobus fulgidus* (NC_000917) sequence at 128 kb, we see that heterogeneity is not always indicated more strongly as window size increases. Any rank ordering of the chromosomes with respect to this heterogeneity measure seems contingent on the choice of window size.

Table 1. Average chi-squared per window of size n on selected sequences.

sequence	window size, n (in kb)								
	1	2	4	8	16	32	64	128	256
random	3.1	3.1	3.1	3.2	3.3	3.5	2.8	3.5	2.5
<i>A. fulgidus</i>	25	37	50	65	77	82	83	83	69
<i>B. subtilis</i>	20	32	52	86	148	257	446	763	127
<i>P. falciparum</i>	48	73	109	156	219	308	289	441	450

A universally effective tiling algorithm would overcome this obstacle by producing the complete bivariate sequence $z_1 z_2 \dots z_N$ where $z_i = (x_i, y_i)$. Collecting all positions for which $x_i = x$, the normalized histogram of y provides an estimator of $b(x,y)$, and the normalized histogram of x will estimate the tile type proportions, $a(x)$. The global composition then approaches $\pi(y) = \sum_x a(x)b(x,y)$. If $\zeta(x)$ is a suitable

measure of dissimilarity between $b(x,y)$ and $\pi(y)$, its average value $\xi = \sum_x a(x)\xi(x)$ quantifies the heterogeneity of the sequence. In keeping with the approach of equation 1, an obvious choice of the dissimilarity measure is the type-specific noncentrality per nucleotide,

$$\xi(x) = \sum_{y=a,c,g,t} \frac{[b(x,y) - \pi(y)]^2}{\pi(y)} > 0. \quad (2)$$

The quantity $X^2/L - d$ can be regarded as a biased estimator of ξ that always *underestimates* the true heterogeneity. If a window of size n covers a single tile type, x' , its expected contribution to X^2 in 1 is approximately $n\xi(x')$. When the window covers more than one tile type, (say) k positions of type x' and $(n - k)$ of x'' , the expected contribution is *less than* or equal to $k\xi(x') + (n - k)\xi(x'')$. The subadditivity of the chi-squared statistic follows from Jensen's inequality and parallels a basic theorem of information theory: The entropy of a mixture equals or exceeds the weighted sum of the component entropies. The same inequality implies the positivity of the Jensen-Shannon entropic divergence measure of compositional complexity used by Bernaola-Galvan, Carpena, Roman-Roldan and Oliver (1999).

2. Method

Lacking both a tiling algorithm and an unbiased estimator, we can still attempt to compare the compositional heterogeneities of genomes in an ordinally consistent manner by removing the scale dependence from the ANOVA chi-squared statistic. Instead of selecting a single window size, we want to construct an average across all window sizes, and (once again) to average across all origins. In keeping with the notation of equation 1, let $L = 1$ be fixed, but let the size n of the window vary from δ to $N - \delta M$ in increments of δ where M is the integer part of N/δ . Then each increment produces a statistic in the series $X^2(m\delta)$, $m = 1, 2, \dots, M$, and

$$X^2(m\delta) = \frac{m\delta N}{N - m\delta} \sum_{y \in a,c,g,t} \frac{[P_m(y) - P(y)]^2}{P(y)} \quad (3)$$

in terms of the base composition P_m of the first $m\delta$ bases in the sequence and the global composition P as previously defined. The last equation follows algebraically from the formal substitutions of P_m for p_1 and $(NP - m\delta P_m)/(N - m\delta)$ for p_2 in equation 1 with $L = 1$. Note that if N/δ is an integer then the series stops after $M - 1$ increments since $P_M = P$ exactly.

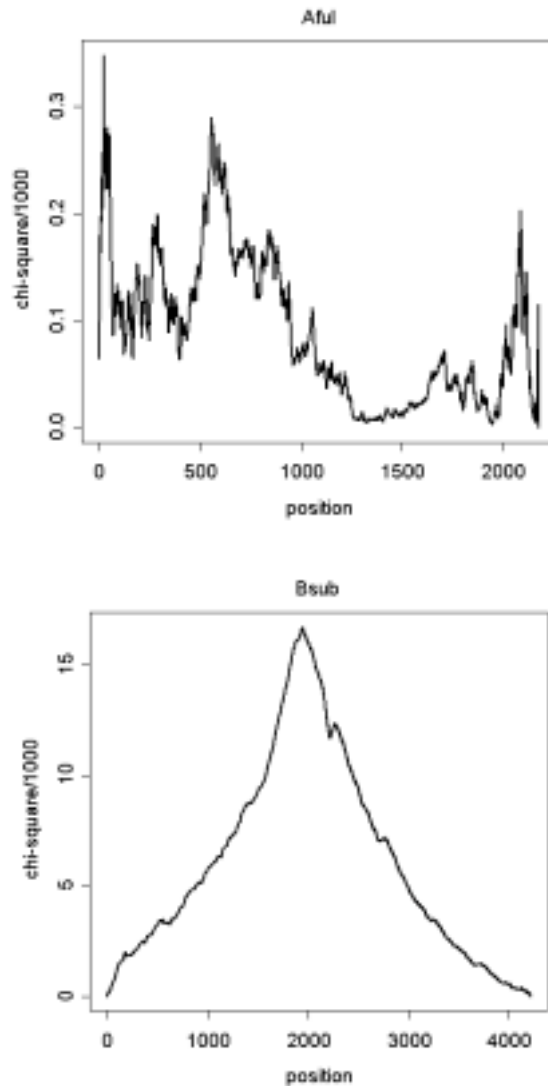


Figure 2. Plots of the chi-squared statistic (equation 3) versus position (in kb) relative to the published sequence origin for (a, above) *Archaeoglobus fulgidus* and (b, below) *Bacillus subtilis*.

Plotting $X^2(m\delta)$ versus position m , as in Figure 2, produces irregular graphs that inevitably fall toward zero as $m \rightarrow M$ (and $P_m \rightarrow P$). The average value of $X^2(m\delta)$ at the M increments now replaces the quantity X^2/L as a measure of heterogeneity. Because these values span almost three orders of magnitude in the data set of only 24 microbial chromosomes, listed in Table 2, the heterogeneity is conveniently expressed as the base ten logarithm

$$H = \log\left(\frac{1}{M} \sum_{m=1}^M X^2(m\delta)\right) \quad (4)$$

Now H is recomputed for each shift of the origin by Δ positions and the results are averaged to produce a grand mean h . The number of such shifts is the integer part of M/Δ . Ideally we would like to take $\delta = \Delta = 1$ base pair (bp) so that equation 3 is iterated N^2 times. But if these increments can be made larger without significantly affecting the result of the calculation then the computation can be greatly expedited. In the following we take $\delta = 1$ kb and choose Δ so that it roughly corresponds to a 1 degree (1°) rotation of the origin in a circular chromosome. For instance, a sequence of length $N = 910,724$ bp is partitioned into contiguous segments of length $\delta = 1$ kb beginning at the start. The window grows by increments of 1 kb until it reaches size $M\delta = 910$ kb. Then $M\delta/360^\circ = 2.53$ kb/ $^\circ$ is rounded up to 3 kb = Δ and the origin is shifted $\text{int}(910/3) = 303$ times in order to complete the calculation of $h = \text{mean}\{H\}$.

Any single measure of heterogeneity will of course be superficial in the sense that functional constraints and evolutionary forces that contribute to the effect are blurred together in a single number. Compositional heterogeneity is typically assessed with reference to certain functions of base composition. The base fractions (A , C , G , and T) sum to one and can be reduced to three components by a linear transformation. The components are usually the purine (R), strong (S), and keto (K) fractions (see Freeman, Plasterer, Smith and Mohr (1998)), defined by the linear equations

$$\begin{bmatrix} R \\ S \\ K \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} A \\ C \\ G \\ T \end{bmatrix}. \quad (5)$$

The GC-skew $(G-C)/(G+C) = (K+R-1)/S$ will typically indicate replication direction in prokaryotes as shown by Lobry (1995, 1996), Grigoriev (1998), Frank and Lobry (2000), and Lobry and Sueoka (2002). The purine excess $A+G-C-T = 2R - 1$ tends to be greater on the coding strand (see Baran, Ko and Jernigan, 2003) as explained by Francino, Chao, Riley and Ochman (1996), Francino and Ochman (1997), and Mrazek and Karlin (1998). The isochore hypothesis of Bernardi (1985, 2000) contends that $S = G+C$ exhibits clustering about specific values that reflect the localization of preferential codon usage in chromosomes of warm-blooded vertebrates. The mosaic view can be regarded as a generalization of the

isochore hypothesis to all three RSK components or equivalently to all four components of base composition.

The present method is adaptable to the RSK components and to arbitrary linear combinations of them. Let the vector on the left in equation 5 be Q for the complete sequence, where the index y runs from 1 to 4 as written, and let $Q_m(i)$ be one of the RSK components in the window spanning the first $m\delta$ bases. When the components are treated individually, the chi-squared statistic

$$X_i^2(m\delta) = \left(\frac{m\delta N}{N - m\delta} \right) \frac{[Q_m(i) - Q(i)]^2}{Q(i)[1 - Q(i)]} \quad (6)$$

has $d = 1$ degree of freedom for $i = R, S,$ or K . Substituting 6 for the summand in 4 gives heterogeneity measures H_i specific to the three components.

3. Heterogeneity in whole chromosomes

The first numeric column of Table 2 lists the heterogeneity measures H computed from equation 3 for the 24 sequences. They range from 1.228 (*Synechocystis*) to 3.823 (*Staphylococcus aureus*). Shifting the origin to the right by increments of Δ such that $\Delta/M \approx 1/360 = 1^\circ$ produces the measurement sample $\{H\} = \{H(k), k = 0, 1, 2, \dots, \text{int}(M/360)\}$; and the sample average is the overall heterogeneity $h = \text{mean}\{H\}$, listed in the second numeric column. These values range from 1.350 (*Saccharomyces cerevisiae*, chromosome XV) to 3.581 (*S. aureus*). Although $10^h < \text{mean}\{10^{H(k)}\}$ in general, the difference is fairly insubstantial, ranging from 0.3% of h in *Mycoplasma genitalium* to 2.3% in *S. cerevisiae*, chromosome XI, and falling below 1% in 16 out of 24 cases. (The larger percentages are in cases where h is small.) These differences are also small compared to the corresponding sample standard deviations $\sigma\{H\}$.

In Figure 2(a), the plot of chi-squared versus position in the unshifted ($k = 0$) *A. fulgidus* sequence is fairly typical of those that score below $h = 2.0$, showing irregular fluctuations and a general declining trend. In contrast, *B. subtilis* exhibits a sharp peak near the midpoint, at position $m = 1944$ in Figure 2(b), reminiscent of the graph of cumulative GC-skew versus position, which peaks around the 1940 kb position and exhibits the same "notch" around 2250 kb.^{14,16} Such similarity is not surprising if the chromosome is seen as a concatenation of two chirochores (or replichores) of nearly equal length that differ with respect to whether the published sequence or its reverse complement is the leading strand of replication.¹²⁻¹⁶ The composition of the leading strand may not differ very much between the chirochores but, so long as $A \neq T$ and/or $G \neq$

Table 2. Results from application of the method to the nucleotide sequences of 24 microbial chromosomes listed alphabetically by species with GenBank accession number.

Species, chr.; GenBank									
	H(0)	h	$\sigma\{H\}$	H(k*)	Δk^*	k**	M		
Archaea									
Arcaheoglobus fulgidus									
	1.951	1.913	.131	2.321	2088	n.a.	2178		
Methanobacter thermoautotrophicus; AE000666									
	2.178	2.197	.127	2.500	400	409t	1710		
Menthanococcus jannaschii; L77117									
	2.614	2.269	.170	2.737	1645	1665t	1664		
Bacteria									
Bacillus subtilis; NC_000964									
	3.782	3.560	.135	3.882	1944	1940t	4214		
Borrelia burgdorferi; AE000783									
	3.741	3.450	.145	3.742	459	458o	910		
Campylobacter jejuni; AL111168			3.318	3.132	.149	3.471	815	835t	1641
Chlamydia pneumoniae; BA000008			2.855	2.986	.134	3.298	207	210t	1229
Chlamydia trachomatis; AE001273			3.074	3.212	.139	3.501	720	721o	1042
Escherichia coli; U00096			2.702	2.822	.111	3.044	1560	1550t	4639
Haemophilus influenzae; L42023			2.308	2.422	.149	2.816	1475	1475t	1830
Helicobacter pylori; AE001439			2.622	2.550	.095	2.819	1555	1598o	1643
Mycobacterium tuberculosis; AE000516			3.038	2.770	.143	3.047	2208	2240t	4403
Mycoplasma genitalium; NC_000908			2.755	2.701	.082	2.850	286	297t	580
Mycoplasma pneumoniae; NC_000912			2.458	2.330	.103	2.562	114	418t	816
Salmonella enterica; NC_003198			2.903	3.054	.126	3.279	1469	1454t	4809
Staphylococcus aureus; NC_002145			3.823	3.581	.148	3.912	1408	1383t	2814
Streptomyces coelicolor; NC_003888			3.029	2.867	.108	3.124	336	177o	8667
Synechocystis sp.; NC_000911			1.228	1.425	.127	1.807	2730	n.a.	3573
Vibrio cholerae, I; AE003852			3.310	3.049	.112	3.310	1	1o	2961
Vibrio cholerae, II; NC_002506			2.870	2.730	.135	3.010	519	507t	1072
Other									
Plasmodium falciparum,II; NC_000910			3.259	2.876	.173	3.312	918	n.a.	947
Plasmodium falciparum,III; NC_000921			3.093	2.726	.184	3.195	39	n.a.	1059
Saccharomyces cerevisiae, XI; NC_001143			1.443	1.388	.168	1.790	548	n.a.	666
Saccharomyces cerevisiae, XV; NC_001147			1.437	1.350	.111	1.683	117	n.a.	1091

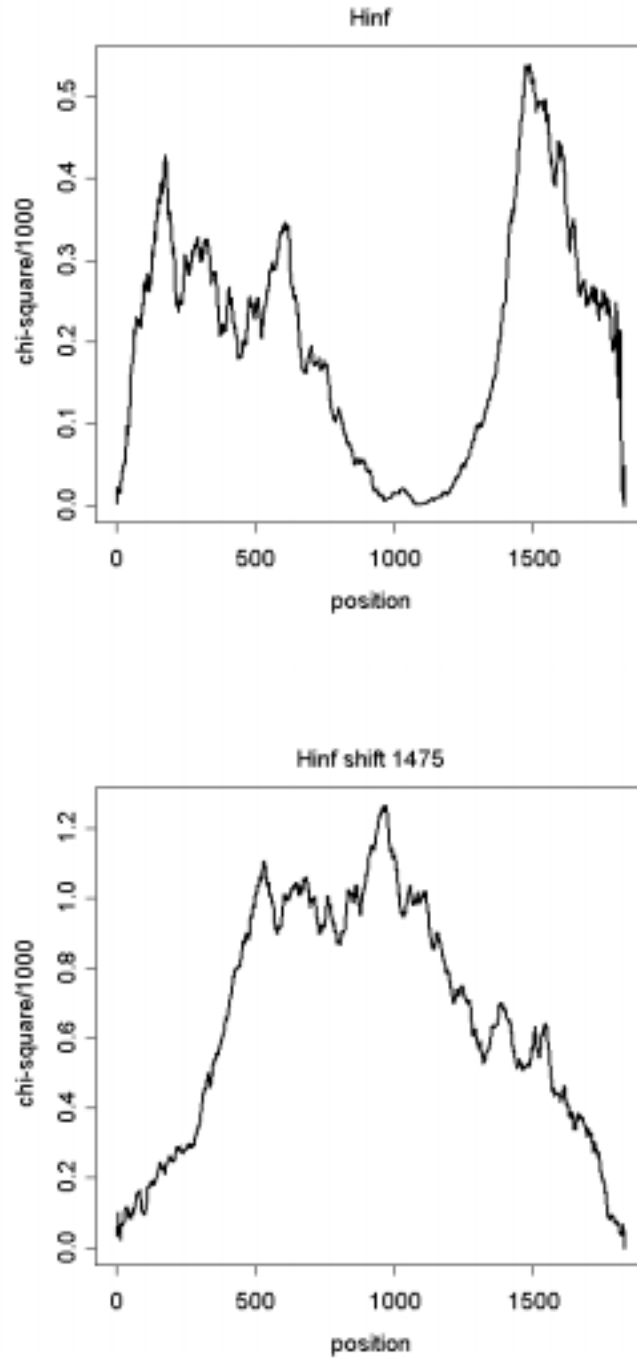


Figure 3. Plots of the chi-squared statistic (equation 3) versus position (in kb) for the *Haemophilus influenzae* sequence (above) before and (below) after a shift of 1475 kb in the 3' direction.

C within them, the composition of the published strand will differ between them. Thus compositional dissimilarity between two parts of the sequence will tend to be maximal when the parts are chirochores.

When the published sequence origin coincides with the replication origin (*ori*), $X^2(m\delta)$ will tend to grow until $m\delta$ passes the replication terminus (*ter*), and decline thereafter until $m = M$. When the sequence starts at *ter*, X^2 will tend to grow until $m\delta$ passes *ori*. Thus the peak in X^2 versus m can indicate *ori* or *ter* when the sequence starts at *vice versa*. When the sequence starts between *ori* and *ter*, the plot of X^2 versus m can show peaks separated by a central valley, as illustrated in Figure 3(a) for the unshifted *H. influenzae* sequence. A shift of 1475 kb to the right will place the sequence origin where X^2 is maximum and typically produce a plot that shows a single, central peak, as in Figure 2(b). Moreover, $H(k)$ tends to be maximal when such a central peak is evident.

Let Δk^* be the shift that maximizes the heterogeneity, so that $H(k^*)$ is the maximum of the chi-squared plot, and let $\delta k^{**} = k^{**}$ kb denote the position of the closest global extremum in cumulative GC skew to the nearest 1 kb. When k^{**} indicates the global maximum of the skew diagram, it is followed by a "t" (for *ter*); and when k^{**} marks the global minimum of the skew diagram, it is followed by an "o" (for *ori*). Table 2 compares these peak locations in all of the bacterial sequences (except *Synechocystis*) and in two archaea. A large discrepancy, 418–114 = 312 kb, is found in *Mycoplasma pneumoniae*; but otherwise the numbers are mainly concordant, agreeing to within 20 kb in 12 cases.

The skew diagram fails to yield a clear indication of chirochore boundaries in six "refractory" cases^{14,16} including *Synechocystis*, *A. fulgidus*, and four eukaryotic sequences, which have "n.a." for k^{**} in Table 2. Can $H(k)$ be used to clarify this issue? With reference to Figure 4, we see that $H(k)$ can vary in a nearly periodic manner, with successive peaks separated by about 180° , in bacterial sequences where sharp boundaries are already indicated by the skew diagram. $H(k)$ attains its maxima in the telomeres of the *Plasmodium* chromosomes and peaks near the 3' end of the *A. fulgidus* sequence. The graph of $H(k)$ for *M. pneumoniae* is particularly disappointing since the global peak around 50° is contradicted by the deep depression 180° later.

These results are fairly insensitive to the choice of increments ($\delta = 1$ kb and $\Delta \approx 1^\circ$). When the computations are repeated with $\delta = 10$ kb and $\Delta \approx 30^\circ$, h changes by an average of -0.003 and a median of $+0.001$. The individual changes range from -0.052 (*P. falciparum* II) to $+0.18$ (*S. cerevisiae* XI). Larger increments must flatten the extrema of the chi-squared plots and thus forfeit the possibility of using this approach as an adjunct to established methods for locating replication boundaries.²⁰

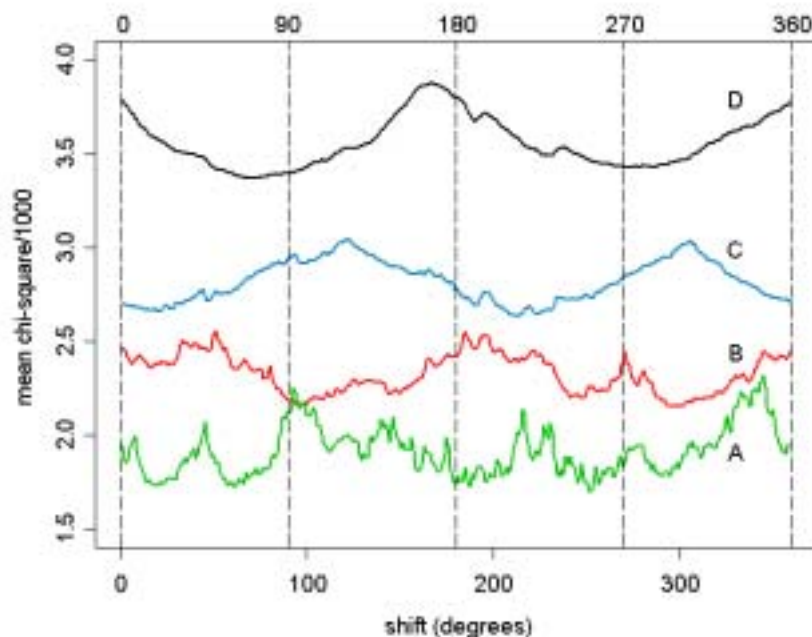


Figure 4. Heterogeneity versus degrees of shift in the sequence origin for (A) *Archaeoglobus fulgidus*, (B) *Bacillus subtilis*, (C) *Salmonella enterica*, (D) *Plasmodium falciparum*, chromosome III, and (E) *Mycoplasma pneumoniae*.

Li *et al.*⁹ applied an ANOVA test to the *Borrelia burgdorferi* genome and confirmed its homogeneity when viewed through windows of a certain size. Table 2 appears to contradict this finding in so far as *B. burgdorferi* ranks third highest in heterogeneity with $h = 3.450$. The discrepancy is basically semantic as the cited work considers only the local fluctuation of $S = G+C$ relative to the global proportion (28.59%). Linear functions of the base composition can be treated by the present method as noted in connection with equation 6. For the unshifted sequence, the $G+C$ heterogeneity is only $H_S(0) = 1.138$ compared to $H(0) = 3.782$ for all components of the composition. The corresponding reduction factor is $10^{-2.644} = 2.3 \times 10^{-3}$.

The mean heterogeneities $h_i \equiv \text{mean}\{H_i\}$ with respect to the RSK components, computed from equation 6 with $\delta = 5$ kb and $\Delta \approx 10^\circ$, are shown in Table 3 for all sequences. The S-component heterogeneity of *B. burgdorferi* ($h_S = 1.205$) is fourth from smallest among the bacteria and archaea in the data set. The difference $h - h_S = 2.243$ for *B. burgdorferi* is the maximum among all $3 \times 24 = 72$ differences of this form ($h - h_i$). The second largest is $h - h_K = 2.234$ for *S. aureus*. Thus the S component

contributes less than 1% of the gross heterogeneity of *B. burgdorferi* and the K component contributes <1% in *S. aureus*. The means of $h - h_i$ in the data set are 0.700 for R, 1.031 for S, and 0.812 for K. When the $h - h_i < \log(2) = 0.301$, more than half of gross heterogeneity can be attributed to just the i component. This occurs four times for R (in *S. aureus*, *B. subtilis*, *M. thermoautotrophicum*, and *M. jannaschii*), 4 times for K (in *B. burgdorferi*, *H. influenzae*, *M. tuberculosis*, and *S. coelicolor*), and once for S (in *Synechocystis*).

Table 3. Mean heterogeneities with respect to the RSK components. The right-most column copies h from Table 2.

species	mean heterogeneity, h						
	R	S	K	all	R	K	all
<i>A. fulgidus</i>	1.513	1.314	0.685	1.912	-	-	-
<i>B. subtilis</i>	3.315	2.614	1.798	3.557	1.743	1.394	2.790
<i>B. burgdorferi</i>	1.852	1.205	3.279	3.448	1.623	0.945	1.868
<i>C. jejuni</i>	2.425	1.320	2.541	3.130	1.461	1.489	2.095
<i>C. pneumoniae</i>	1.996	1.599	2.681	2.985	0.835	1.236	1.959
<i>C. trachomatis</i>	2.461	1.427	2.852	3.212	0.522	1.008	1.717
<i>E. coli</i>	2.147	2.058	2.364	2.823	1.155	0.973	2.291
<i>H. influenzae</i>	1.265	1.006	2.145	2.426	1.070	1.712	2.066
<i>H. pylori</i>	2.058	1.356	1.891	2.549	1.354	0.711	1.838
<i>M. thermoautotro..</i>	1.951	1.254	0.871	2.196	-	-	-
<i>M. jannaschii</i>	1.976	1.314	0.598	2.264	-	-	-
<i>M. tuberculosis</i>	1.827	1.088	2.619	2.768	0.638	0.576	1.464
<i>M. genitalium</i>	2.132	2.386	1.445	2.701	1.195	1.348	2.548
<i>M. pneumoniae</i>	1.989	1.623	1.535	2.330	1.275	1.163	1.992
<i>P. falciparum II</i>	2.317	2.285	1.293	2.855	-	-	-
<i>P. falciparum III</i>	1.527	2.411	1.767	2.726	-	-	-
<i>S. cerevisiae XI</i>	0.930	0.644	0.781	1.396	-	-	-
<i>S. cerevisiae XV</i>	0.737	0.689	0.562	1.291	-	-	-
<i>S. enterica</i>	2.388	1.940	2.725	3.056	1.175	1.101	2.210
<i>S. aureus</i>	3.368	1.786	1.346	3.580	1.781	1.446	2.250
<i>S. coelicolor</i>	2.067	1.908	2.640	2.862	1.161	1.757	2.347
<i>Synechocystis sp.</i>	0.329	1.155	0.314	1.421	-	-	-
<i>V. cholerae I</i>	2.251	2.090	2.691	3.050	0.749	1.321	2.300
<i>V. cholerae II</i>	1.639	2.049	2.356	2.731	0.570	1.288	2.261

4. Heterogeneity in single genes

These results suggest a new solution to the problem, posed at the outset, of partitioning the genome into compositionally homogeneous segments. The maximum of X^2 from equation 3 specifies a partition of the complete sequence into two parts flanking the position Δk^* from the published origin. These two parts, which are loosely called "halves," are inhomogeneous to the extent that X^2 exceeds the critical value of the chi-squared test at significance level α . Now each "half" can be analyzed by the same method and segmented into "quarters" which in turn give rise to eighths, sixteenths, and so on. The procedure stops when no partition of the segment yields $X^2 > \chi_d^2(\alpha)$. For example, the 3 degree of freedom test at 1.8% significance has critical value $\chi_3^2(1 - .018) \approx 10.0$. Figure 5 shows the end result of segmenting the bacteriophage lambda (GenBank NC_001416) sequence with $\alpha = 1.8\%$ significance and $\Delta = 1$ kb resolution as before. When the boundaries, indicated by the vertical bars, are only 1 kb apart, the implication is that inhomogeneity persists down to a finer scale, and hence a smaller Δ could be desirable. The longest homogeneous segments are between positions 1 and 20 kb where the coat protein is encoded in several long open reading frames.

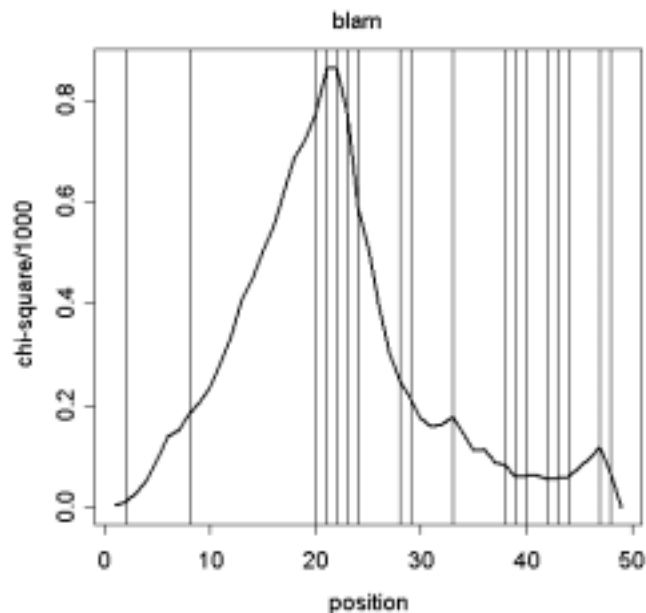


Figure 5. Plot of the chi-squared statistic (equation 3) versus position (in kb) for the bacteriophage lambda sequence with vertical bars partitioning it into homogeneous segments at the 1.8% significance level with 1kb resolution.

References

1. S. Karlin and V. Brendel, "Patchiness and correlations in DNA sequences," *Science* **259**: 667-679 (1993).
2. P. Bernaola-Galván, P. Carpena, R. Román-Roldán and J.L. Oliver, "Compositional complexity of DNA sequence models," *Computer Physics Communications* **121-122**: 136-138 (1999).
3. W. Li, G. Stolovitzky, P. Bernaola-Galván, and J.L. Oliver, "Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes," *Genome Research* **8**: 916-928 (1998).
4. J.V. Braun and H-G. Müller, "Statistical methods for DNA sequence segmentation," *Statistical Science* **13**: 142-162 (1998).
5. G. Bernardi et al., "The mosaic genome of warm-blooded vertebrates," *Science* **228**: 953-958 (1985).
6. G. Bernardi, "Isochores and the evolutionary genomics of vertebrates," *Gene* **241**: 953-17 (2000).
7. International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature* **409**: 860-921 (2001).
8. W. Li, "Are isochore sequences homogeneous?" *Gene* **300**: 129-139 (2002).
9. W. Li, P. Bernaola-Galván, P. Carpena and J.L. Oliver, "Isochores merit the prefix 'iso'," *Computational Biology and Chemistry* **27**: 5-10 (2003).
10. A. Agresti, *Categorical Data Analysis*, New York, Wiley (1990).
11. J.M. Freeman, T.N. Plasterer, T.F. Smith and S.C. Mohr, "Patterns of genome organization in bacteria," *Science* **279**: 1827-1829 (1998).
12. J.R. Lobry, "Properties of a general model of DNA evolution under no-strand-bias conditions," *J. Mol. Evol.* **40**, 326-330 (1995).
13. J.R. Lobry, "Asymmetric substitution patterns in the two DNA strands of bacteria," *Mol. Biol. Evol.* **13**, 660-665 (1996).
14. A. Grigoriev, "Analyzing genomes with cumulative skew diagrams," *Nucleic Acids Res.* **26**, 2286-2290 (1998).
15. J.R. Lobry and N. Sueoka, "Asymmetric directional mutation pressures in bacteria," *Genome Biology* **3**, 58.1-58.14 (2002).
16. R.H. Baran, H. Ko and R.W. Jernigan, "Methods for comparing sources of strand compositional asymmetry in microbial chromosomes," *DNA Research* **10**, 85-95 (1993).
17. M.P. Francino, L. Chao, M.A. Riley, and H. Ochman, "Asymmetries generated by transcription-coupled repair in enterobacterial genes," *Science* **272**, 107-109 (1996).
18. M.P. Francino and H. Ochman, "Strand asymmetries in DNA evolution," *Trends Genet.* **13**, 240-245 (1997).
19. J. Mrázek and S. Karlin, "Strand compositional asymmetry in bacterial and large viral genomes," *Proc. Natl. Acad. Sci. USA* **95**, 3720-3725 (1998).
20. A.C. Frank and J.R. Lobry, "Oriloc: prediction of replication boundaries in un-annotated bacterial chromosomes," *Bioinformatics* **16**: 560-561 (2000).