G+C Content Variation Along and Among *Saccharomyces cerevisiae* Chromosomes

Keith R. Bradnam, Cathal Seoighe,† Paul M. Sharp,* and Kenneth H. Wolfe†*

*Division of Genetics, University of Nottingham, United Kingdom; and †Department of Genetics, Trinity College, University of Dublin, Ireland

Past analyses of the genome of the yeast *Saccharomyces cerevisiae* have revealed substantial regional variation in G+C content. Important questions remain, though, as to the origin, nature, significance, and generality of this variation. We conducted an extensive analysis of the yeast genome to try to answer these questions. Our results indicate that open reading frames (ORFs) with similar G+C contents at silent codon positions are significantly clustered on chromosomes. This clustering can be explained by very short range correlations of silent-site G+C contents in neighboring ORFs. ORFs of high silent-site G+C content are disproportionately concentrated on shorter chromosomes, which causes a negative relationship between chromosome length and G+C content. Contrary to previous reports, there is no correlation between gene density and silent-site G+C content in yeast. Chromosome III is atypical in many regards, and possible reasons for this are discussed.

Introduction

The recent accumulation of genome sequence data has provided the opportunity to investigate aspects of genomic structure that were not previously accessible. It is now possible to ask: are chromosomes essentially just a genetic beanbag—a random distribution of genes interspersed with genetic flotsam and jetsam? Or do genomes display evidence of large-scale structure: are there patterns in the distribution of genes, and does DNA composition vary in a systematic way with respect to genomic location?

There are already some answers to these questions. For example, some large-scale structure has been detected within vertebrate genomes: regional variation in G+C content is present in the form of isochores, which are long (>100 kb) tracts of DNA that appear to be compositionally homogeneous (Bernardi 1995). However, perhaps the most illuminating insights into genome evolution will come from analyses of complete genome sequences. Analyses of complete bacterial genome sequences have revealed previously unseen patterns of base composition variation (Kerr, Peden, and Sharp 1997; McInerney 1997). Here, we investigate the budding yeast *Saccharomyces cerevisiae*, the first eukaryote for which an entire genomic sequence is available (Goffeau et al. 1997). Analysis of this unicellular eukaryote may allow us to detect fundamental patterns that shape the more complex genomes of multicellular organisms.

Even before the completion and subsequent analysis of the 12.1-Mb yeast genome sequence (Dujon 1996; Goffeau et al. 1997), many studies were conducted on the first few chromosome sequences that were available. Chromosome III was the first of the 16 chromosomes to be sequenced (Oliver et al. 1992), and base composition (G+C content) displays striking variation along that chromosome (Sharp and Lloyd 1993). This is particularly evident at the third positions of codons: each chromosome arm displays one large peak in silent-site G+C content (GC3s). Even though codon usage bias varies among yeast genes in correlation with their expression levels (Bennetzen and Hall 1982; Sharp, Tuohy, and Mosurski 1986; Sharp and Cowe 1991), this is not related to the GC3s variation, because the set of ''preferred'' codons in highly expressed yeast genes has a G+C content similar to the mean content for the whole genome (Sharp and Lloyd 1993). The next three sequenced chromosomes (XI, II, and VIII; Dujon et al. 1994; Feldmann et al. 1994; Johnston et al. 1994) were longer, and multiple periodic peaks of GC3s were seen. It was shown that the numbers of G+C-rich peaks present on these four chromosomes correlate with chromosome length, with approximately one peak per 100 kb (Sharp et al. 1995). However, periodic spacing of G+C-rich peaks was not seen in all of the remaining 12 chromosomes, and therefore the generality of this phenomenon is unclear.

In the analysis of chromosome XI, Dujon et al. (1994) also noted a correlation between increases in GC3s and increases in local gene density. Such a correlation between gene density and base composition has been seen for mammalian isochores (Bernardi 1995). This correlation was also found in chromosome III (Sharp and Matassi 1994) and the subsequent primary publications of chromosomes II, IV, VIII, IX, X, XIII, and XV all reported periodic peaks in G+C content with corresponding increases in gene density (Feldmann et al. 1994; Johnston et al. 1994; Galibert et al. 1996; Bowman et al. 1997; Churcher et al. 1997; Dujon et al. 1997; Jacq et al. 1997). However, many of these correlations are weak and do not appear convincing. Furthermore, for chromosomes VI, VII, XII, and XVI (Murakami et al. 1995; Bussey et al. 1997; Johnston et al. 1997; Tettelin et al. 1997), no such correlation was detected (although for chromosome VII, removal of Ty and LTR elements produced a correlation). For the remaining chromosomes (I, V, and XIV), analysis of G+C variation and/or gene density was not undertaken or not discussed in any detail (Bussey et al. 1995; Dietrich et al. 1997; Philippsen et al. 1997).

Key words: yeast, genome, G+C content, chromosome, isochore.

Address for correspondence and reprints: Kenneth H. Wolfe, Department of Genetics, Trinity College, University of Dublin, Dublin 2, Ireland. E-mail: khwolfe@tcd.ie.

Mol. Biol. Evol. 16(5):666–675. 1999

Thus, the apparently clear patterns of regional variation originally found in chromosomes III and XI have not been confirmed universally for other chromosomes. In this paper, we have therefore set out to clarify the nature of GC3s content variation along and among all yeast chromosomes. It is important to discount the possibility that any patterns in GC3s variation are an artifact of the methodology used. Previous methods that have detected ''peaks'' are somewhat subjective and can be sensitive to the choice of window size. Here, a more objective approach is used, which simply asks whether open reading frames (ORFs) that are similar in GC3s content are significantly clustered.

We also address aspects pertaining to the origin and maintenance of GC3s variation by testing whether patterns of GC3s variation are conserved between the large duplicated chromosomal regions identified in yeast (Wolfe and Shields 1997; Seoighe and Wolfe 1998). Furthermore, the unresolved issues of GC3s periodicity and correlations between gene density and G+C variation (Dujon et al. 1994) are investigated. Finally, we consider the variation in individual ORF GC3s values from a global perspective, examining whether ORFs of different GC3s-richness are distributed randomly throughout the yeast genome.

Materials and Methods

Sequences and Data

DNA sequences and details of ORFs were obtained from the Saccharomyces Genome Database (SGD; Cherry et al. 1998). Specifically, the chromosome sequences were downloaded (January 1998) from ftp://genome-ftp.stanford.edu/pub/yeast/genome_seq/, and the ORF location tables were downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/tables/ORF_Locations/. All ORFs that were completely contained within larger ORFs were excluded from subsequent analysis. ORFs encoded by yeast transposons (Ty elements) were removed from the analysis. This left 6,145 ORFs, of which 2,721 had been annotated as being genes (i.e., they had genetic names as well as ORF designations). Information on duplicated blocks in the yeast genome is available from http://acer.gen.tcd.ie/~khwolfe/yeast/.

Test for the Significant Clustering of ORFs of Similar GC3s Values

To try to delimit significant clusters of ORFs with similar GC3s values on each chromosome, the following method was devised. We began with a sliding window of two ORFs and tried every window size up to half the length of the chromosome under analysis. In this way, every set of adjacent ORFs was tested against the remaining ORFs on the chromosome using *t*-tests to distinguish whether there was a statistically significant difference in mean GC3s. Nominally statistically significant results were recorded.

We then tried to exclude the possibility that many of these significant *t*-test results were simply due to the large number of windows examined. The order of the ORFs on each chromosome was shuffled randomly, and *t*-tests were performed on these shuffled data. One thousand simulations were made in this manner. We returned to the *t* values for the unshuffled data and asked for each significant *t* value: in how many of the 1,000 simulations was this *t* value exceeded? From these data, a list of windows on different chromosomes whose average GC3s seemed significantly different ($P < 0.05$) from the average GC3s of the rest of the chromosome was produced. Many of these significant windows overlap with each other; for example, a highly significant window of two ORFs might be overlapped by a larger, slightly less significant, window of five ORFs. This technique overcomes biases introduced by choice of window size by considering all possible windows.

A modified shuffling method was also used. In this shuffling scheme, chromosomes were built from randomly chosen ORFs which were sequentially accepted or rejected with a probability based on the difference between their GC3s contents and the GC3s content of the preceding ORF. For example, for a given ORF, if this difference in GC3s content was 0.02, and if the fraction of ORFs in the real data that had 0.02 of a difference in GC3s content as compared with their nearest neighbor was 0.1, then that randomly chosen ORF would be accepted with a probability of 0.1.

Results

Variation in Silent-Site G+C Content

Variation in GC3s was calculated and plotted for all 16 chromosomes (fig. 1) using a sliding window of 15 adjacent genes (as in Sharp and Lloyd 1993). The plotting of all chromosomes to the same scale allows for an objective comparison between chromosomes. Chromosome III exhibits the most striking patterns of GC3s variation and is the only chromosome for which a window of GC3s exceeds 50%. Only two other chromosomes (I and XI) have regions of average GC3s that exceed 45%. Chromosome IV contrasts starkly with chromosome III; GC3s variation is confined within a narrow band and does not exceed 40%. Overall, there seem to be no clear and consistent patterns among all 16 chromosomes. For some chromosomes (such as XI), GC3s variation appears to be almost periodic, and ''peaks'' of GC3s are roughly uniform in size. Other chromosomes (such as X) display a far less predictable pattern of GC3s variation.

GC3s Variation in Duplicated Chromosomal Regions

Initial studies of GC3s variation in yeast concentrated on the peaks of high GC3s that are seen when the data have been smoothed by a window of 15 ORFs (fig. 1). These peaks, which span up to approximately 40 ORFs, might be produced by long-range effects that require stability of a chromosome segment over a long period of time. If this is so, then large regions that have been undisrupted by chromosomal rearrangements might be expected to contain some of the highest peaks. The large duplicated blocks identified by Wolfe and Shields (1997) provide us with segments that have not undergone large-scale rearrangement since genome duplication about 10^8

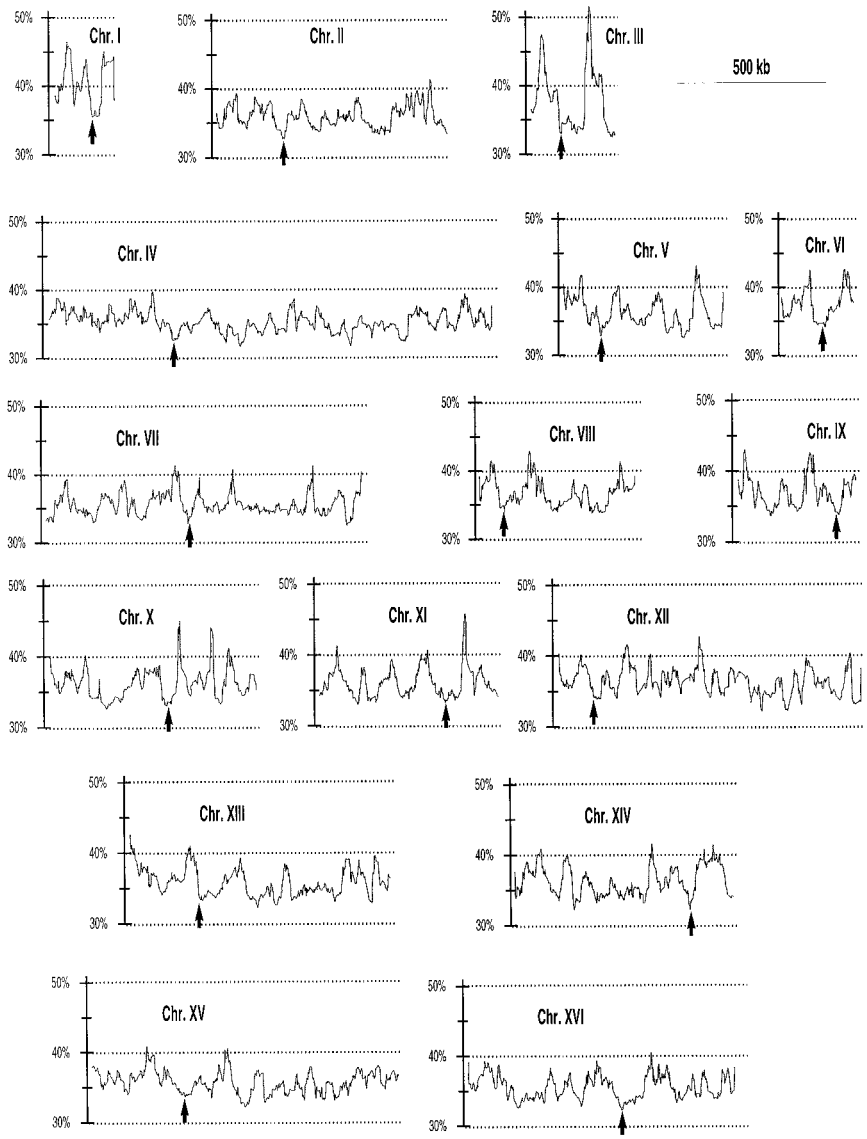

FIG. 1.—Variation in silent-site G+C content (GC3s) along 16 yeast chromosomes. GC3s was calculated as in Sharp and Lloyd (1993) using a sliding window of 15 ORFs, but plotted as a line rather than as a series of points for ease of presentation. All chromosomes are drawn to the same scale. Arrows denote approximate positions of centromeres. Dotted lines denote 30%, 40%, and 50% GC3s.

years ago. Locations of GC3s peaks exceeding 40% (see dotted line in fig. 1) were examined against the coordinates of the 22 largest undisrupted segments in the yeast genome, spanning about 17% of the genome. Of 63 GC3s peaks, only 10 were found to be even partially within the large duplicated blocks. This was not more than would be expected by chance. Analysis of GC3s in pairs of genes that have remained in duplicate since genome duplication shows only a very weak, although significant, correlation ($r = 0.34$, $N = 406$, $P < 0.01$), indicating that change of GC3s content is rapid compared with the time estimated for genome duplication.

Are ORFs that Are Similar in GC3s Significantly Clustered?

The methodology used to display GC3s variation in figure 1 has a natural bias toward finding peaks in the data. It is perhaps better to consider GC3s variation in terms of the possible physical clustering of ORFs with similar GC3s values. A method was devised to delimit such clusters (see *Materials and Methods*) using *t*-tests. Most chromosomes were found to have only a few of these clusters; some chromosomes (I and II) had none at all. Both high-GC3s and low-GC3s clusters were found. The high-GC3s clusters were typically small (less than 10 ORFs), whereas the low-GC3s clusters were usually much longer (up to 200 ORFs). Chromosomes X and XI produced many high-GC3s clusters, but it was chromosome III that again stood out. Most of chromosome III is occupied by two sets of overlapping clusters, one of which includes a window of six GC3s-rich ORFs on the right arm that is the most significant cluster in the entire genome, and the other of which is a large GC3s-poor area in the middle of the chromosome (data not shown; see fig. 1). It seems, therefore, that there are

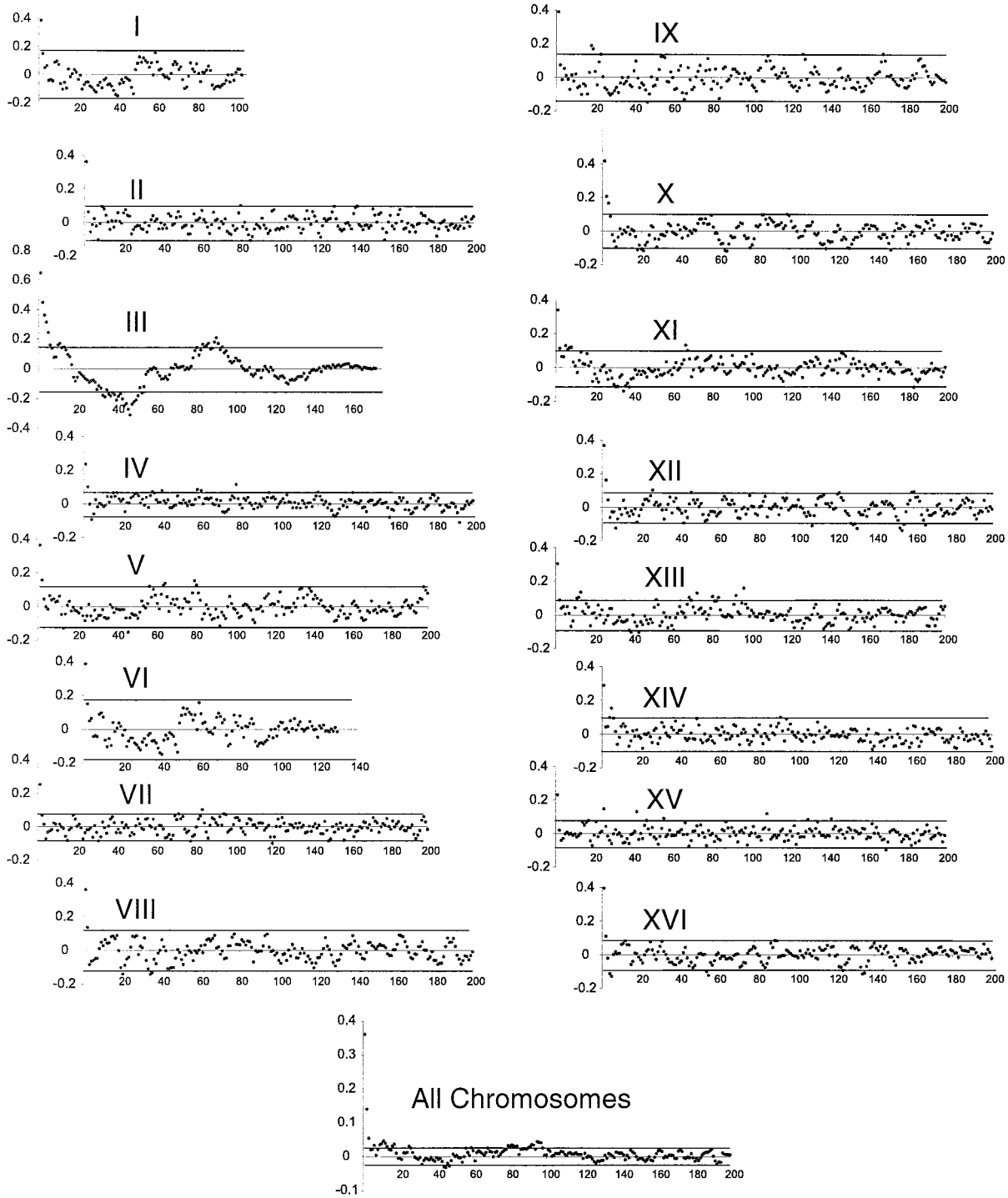

FIG. 2.—Correlograms for yeast chromosomes. The Y-axis in each plot shows the autocorrelation coefficient, r_k, which is a measure of the correlation between the GC3s values of all ORFs that are a distance k ORFs apart on the chromosome in question. Distances (k) between ORFs are shown on the X-axis for values up to $k = 200$. The 5% significance levels for autocorrelation coefficients are shown as two solid lines. Points lying outside these lines represent significant correlations. The bottom panel shows the pooled result when all chromosomes are considered together.

regions of the genome (particularly chromosome III) where GC3s variation is significantly nonrandom.

Is GC3s Variation Periodic?

Results from the first few chromosomes to be sequenced hinted at periodicity in GC3s. Methods from time series analysis were used to search for any such periodicity in the yeast genome (e.g., see Chatfield 1989, chapter 2). Correlograms for each chromosome (fig. 2) show the correlation (as measured by the autocorrelation coefficient, r_k) between the GC3s values of any two ORFs as a function of the distance (k, measured in

ORFs) between them. With completely random data, a correlogram would be expected to fluctuate randomly about $r_k = 0$. Periodicity in the data should produce regular significant autocorrelations at large distances. For most chromosomes, there was no significant autocorrelation at large distances (fig. 2). In each chromosome, however, there was a significant short-range autocorrelation normally extending only to nearest-neighboring ORFs ($k = 1$), or sometimes to second-nearest neighbors ($k = 2$). However, chromosome III again stands out from the other chromosomes (fig. 2). It is the only chromosome to exhibit significant long-range correlations (with a trough at approximately $k = 45$ and a peak at $k = 90$, both $P < 0.05$, and a weaker trough at $k = 130$). Short-range correlations (over distances of $k = 1$–5) are also much stronger on this chromosome than on others. This strong short-range correlation may be sufficient to give the impression of longer-range effects. There also appears to be less random ''noise'' in r_k at different values of k for chromosome III than for other chromosomes; this is a consequence of the stronger short-range correlations (fig. 2).

For the whole genome (bottom of fig. 2), there are statistically significant autocorrelations for nearest neighbors ($r_1 = 0.36$) and for ORFs separated by one intervening ORF ($r_2 = 0.14$), and the correlation for $k = 3$ is of borderline significance ($r_3 = 0.05$). The 5% statistical significance level for the whole-genome plot corresponds to $r = 0.03$, and the value of $r_1 = 0.36$ was not exceeded in 10,000 random permutations of the data. The slight trough (near $k = 45$) and peak (near $k = 90$) in the plot for the whole genome are attributable to chromosome III and disappear if that chromosome is excluded.

The Nature of Short-Range Correlations

The short-range correlations in GC3s between ORFs occur regardless of the strand on which the neighboring ORFs are located. Furthermore, the correlations do not appear to be due to correlated levels of expression. The frequency of optimal codons (Fop; Ikemura 1981; Sharp and Cowe 1991) was used as an indicator of expression level, and no correlation was detected when the Fop values of neighboring ORFs were compared. Interestingly, the strength of GC3s correlation between neighboring ORFs depends strongly on their distance of separation, measured in base pairs. In the whole genome, for the 5% of neighboring ORFs that are most distantly separated, $r_1 = 0.19$; for the 5% of ORFs that are closest, $r_1 = 0.43$.

We investigated whether the short-range correlations observed in figure 2 could explain the significant clustering of ORFs of similar GC3s values. To do this, the multiple-*t*-test methodology outlined earlier was repeated, but a bias was introduced into the way the ORFs on a chromosome were shuffled. Chromosome III was chosen because it contains the most pronounced clustering of high-GC3s ORFs. Instead of shuffling ORFs randomly, they were shuffled taking into account the tendency for neighboring ORFs to have similar GC3s values. To do this, we first observed the range of differences in the GC3s values of adjacent ORFs in the real, unshuffled data. From this, we could determine how likely it is that two ORFs with a given difference in GC3s will be adjacent to each other, so shuffled data sets could be produced that have short-range correlation profiles similar to that of the real data. Using this modified shuffling technique, all the significant clusters of ORFs on chromosome III could easily be reproduced. Therefore, the nearest-neighbor correlation seems sufficient to explain the significant clustering of ORFs of similar GC3s values.

In contrast to the short-range GC3s correlations between neighboring ORFs, there is no significant positive correlation between the G+C content of adjacent noncoding regions (even when taking into account the fact that these tend to be farther apart than neighboring ORFs). However, if noncoding regions are split in two, the G+C contents of the two halves are weakly correlated ($r = 0.27$, $N = 6{,}300$, $P < 0.01$). This suggests that there is a very short range correlation in the base composition of noncoding regions.

Correlation Between Chromosome Length and Chromosome G+C Content

We noted from figure 1 and from shuffling experiments that shorter chromosomes tend to have clusters with very high GC3s contents. There is a strong negative relationship (fig. 3*A*) between the length of a chromosome and its crude G+C content ($r = -0.78$, $P < 0.01$). If repetitive elements and remnants of transposable elements are removed from the calculation of chromosome length and G+C content, the correlation between the two increases ($r = -0.84$). We will refer to this measurement of chromosome length as ''adjusted chromosome length.''

Because noncoding regions of the yeast genome typically have lower G+C contents than genes, the relationship observed in figure 3*A* could be due to longer chromosomes containing a higher proportion of noncoding DNA, but, in fact, the opposite is true. The shortest chromosomes have a disproportionately greater concentration of noncoding sequence (defined here as DNA that is not in an ORF, RNA sequence, Ty element, or solo LTR). Approximately 35% of chromosome I (the shortest chromosome) is noncoding, as compared to 25% of chromosome IV (the longest chromosome). It has been postulated that the shorter yeast chromosomes may contain ''filler'' material to increase their stability (Bussey et al. 1995; Oliver 1995).

When separate analyses are carried out for different classes of sequence (fig. 3*B*), the correlations with chromosome length are strongest for crude ORF G+C content ($r = -0.83$, $P < 0.01$) and for weighted mean ORF GC3s ($r = -0.81$, $P < 0.01$), and they are weaker for noncoding regions ($r = -0.57$, $P < 0.05$). ORFs on the shortest chromosome have an average silent-site G+C content (GC3s) that is 5% higher than that on the longest chromosome. The high value for chromosome I might indicate a curved relationship rather than a linear one between the two variables. The values of these correlations increase slightly if the measure of adjusted chro-

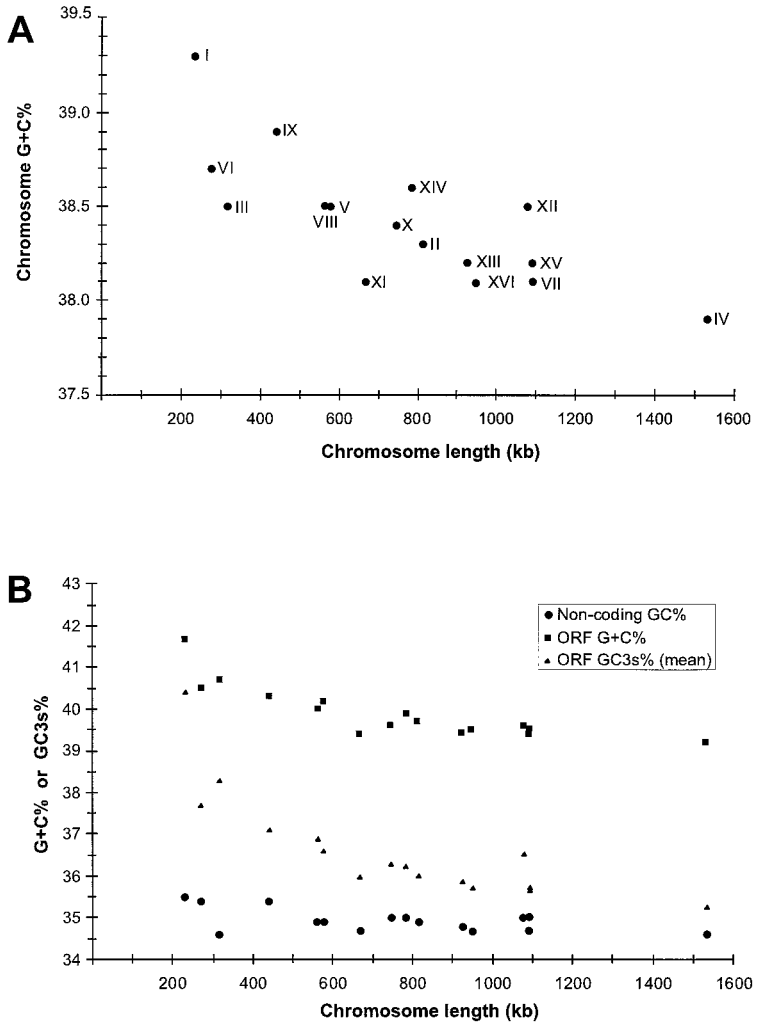

FIG. 3.—Relationships between chromosome length and base composition. *A*, Correlation between chromosome length and chromosome G+C content (GC%). Roman numerals denote chromosome numbers. *B*, Correlations between chromosome length and ORF GC%, ORF GC3s%, and noncoding GC%.

mosome length is used. These relationships hold when chromosome arms are considered separately, although the correlation values are lower (data not shown).

Distribution of All GC3s Values

From figure 3, the question arises of why shorter chromosomes tend to have ORFs with higher G+C and GC3s contents. This also has implications for how we should interpret figure 1. Higher peaks in GC3s were observed on the shorter chromosomes, but this might just be a reflection of these chromosomes' higher underlying GC3s values. If the modal GC3s value, rather than the mean, is used, the relationship between chromosome length and GC3s does not hold ($r = -0.14$, NS). This suggests that the GC3s values for all yeast ORFs are not normally distributed. In fact, there is considerable skew, with twice as many ORFs on the right-hand side of the mode of the distribution as on the left (fig. 4*A*). If the right-hand side of the distribution resembled the left-hand side, it might be expected that the right hand side would tail off at about 48% GC3s. Interestingly, the distribution of G+C values for noncoding regions shows no significant positive tail (fig. 4*B*), and there is actually a slight negative tail.

The uneven distribution of ORFs with high GC3s values among chromosomes is confirmed by examination of the locations of the ORFs in the tail of the distribution (fig. 4*A*). Shorter chromosomes have a higher proportion of high-GC3s ORFs than might be expected by chance (table 1). In fact, the number of ORFs with GC3s $>$ 48% on each chromosome seems to be independent of chromosome length ($r = 0.47$, NS).

Relationship Between Gene Density and GC3s

Many authors have cited a correlation between regional increases in GC3s and regional increases in gene density (see *Introduction*). To test whether this is true, each chromosome was divided into adjacent 50-kb windows. For example, chromosome III (315 kb) was divided into six windows of 50 kb centred in the chromosome. The remaining fragments (7.5 kb at each telomere) were removed. For each window, gene density and the average ORF GC3s was calculated. Incomplete parts of ORFs at the ends of windows were included in

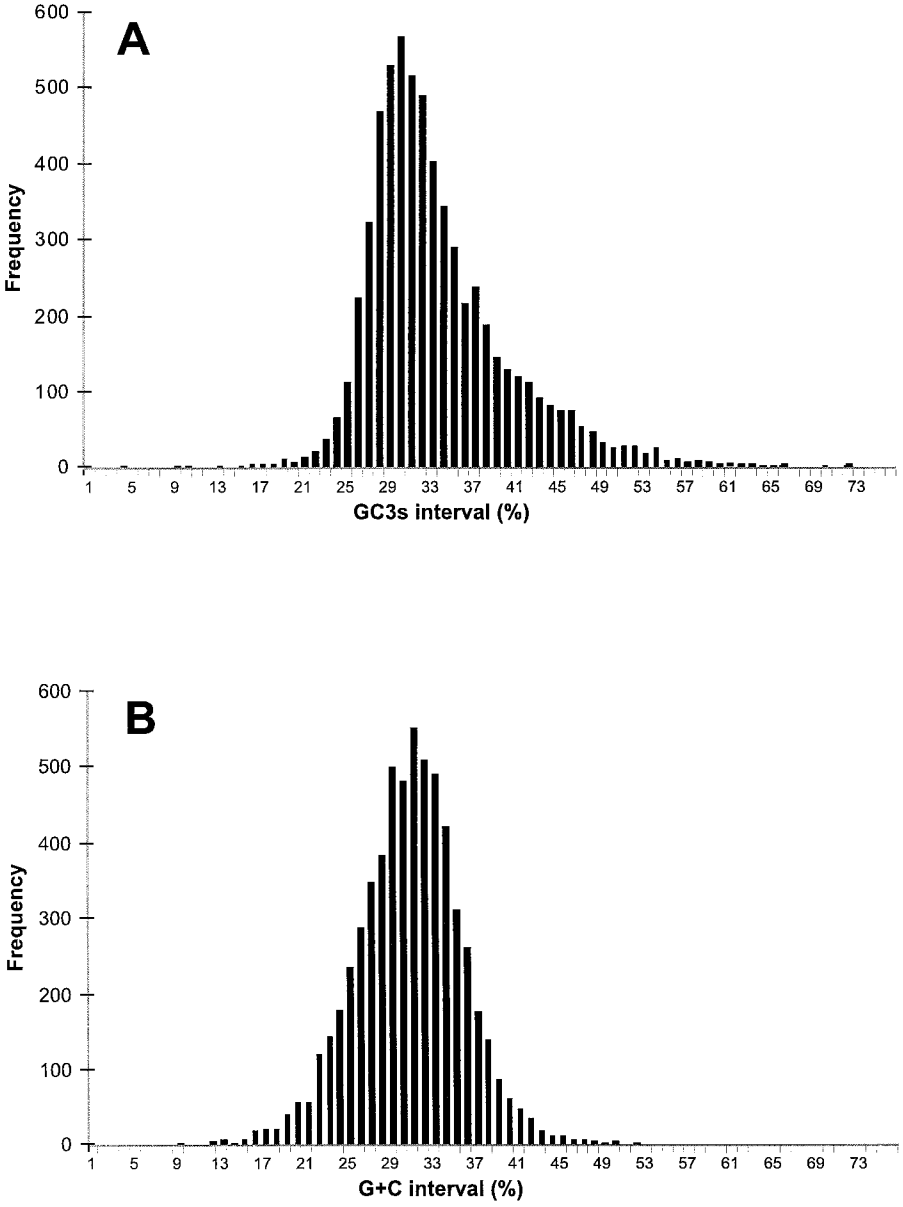

FIG. 4.—Distribution of G+C and GC3s values in the yeast genome (interval size, 1% G+C or GC3s). *A*, Distribution of all ORF GC3s values from 6,145 ORFs. *B*, Distribution of G+C values from 6,004 noncoding regions (regions of less than 75 bp were excluded from the analysis).

the calculations. There is no correlation between the variables (fig. 5). Making allowance for the fact that ORFs on shorter chromosomes have a higher average GC3s by subtracting the average chromosome GC3s value from each ORF still results in no correlation (data not shown).

Discussion

Our analyses have produced two apparently independent results. First, while variation in GC3s is not completely random, the observed clusters of ORFs of similar GC3s values can be accounted for by considering the very short range correlations between neighboring ORFs. Second, high-GC3s ORFs are located preferentially on shorter chromosomes, and the distribution of all GC3s values is not normal. Furthermore, both of these results are apparent only when considering the silent sites of ORFs and not when considering noncoding regions. Noncoding regions in yeast are typically short and may consist largely of regulatory elements that are under selective constraint (Sharp and Lloyd 1993). This hypothesis is supported by the finding that intergenic regions are more conserved than silent sites when closely related *Saccharomyces* species are compared (Adjiri et al. 1994; unpublished data). The results for noncoding regions are therefore inconclusive. They reflect either a fundamental lack of pattern or a pattern that is largely obscured by selective constraints.

Nearest-Neighbor Effects

Regional variation in base composition has largely been inferred to be due to regional variation in mutation patterns (Filipski 1987; Sharp and Lloyd 1993). One

Table 1
Chromosome Distribution of ORFs Above Various GC3s Cut-Off Levels

GC3s Cut-Off (%)	ORFs[a]	Percentage of ORFs on each chromosome that are above cut-off[b]														
		I (0.2 Mb)	VI (0.3 Mb)	III (0.3 Mb)	IX (0.4 Mb)	VIII (0.6 Mb)	V (0.6 Mb)	XI (0.7 Mb)	X (0.7 Mb)	XIV (0.8 Mb)	II (0.8 Mb)	XIII (0.9 Mb)	XVI (0.9 Mb)	XII (1.1 Mb)	VII (1.1 Mb)	XV (1.1 Mb)	IV (1.5 Mb)
55	131 (2%)	7.7	3.8	9.7	3.7	3.2	3.2	1.8	3.2	1.7	1.2	0.6	1.0	3.0	1.1	1.4	0.9
50	312 (5%)	16.3	7.6	14.3	7.3	6.1	6.5	3.3	5.3	5.8	3.1	4.0	3.9	7.3	3.4	3.4	3.3
45	767 (12%)	22.1	21.2	22.9	17.8	15.1	13.0	9.6	14.5	13.5	10.8	12.1	11.1	15.0	10.0	7.7	10.1
40	1623 (26%)	45.2	37.1	36.0	34.7	30.6	25.3	24.9	28.2	29.2	26.4	26.1	23.6	28.4	22.9	21.6	21.5
35	3516 (57%)	76.0	67.4	61.7	64.8	62.2	56.7	58.3	57.1	55.7	57.3	56.8	50.7	59.8	56.9	54.4	53.5

[a] Number of ORFs in the genome which exceed the GC3s cutoff. Values in parentheses are number of ORFs above GC3s cut-off as a percentage of all ORFs in the genome.
[b] Chromosomes are ordered from left to right by increasing size.

possible cause of mutation pattern variation is that different regions of the genome are replicated at different times (Wolfe, Sharp, and Li 1989; Eyre-Walker 1992). A second possibility is that regional mutation patterns reflect differences in the local frequency of recombination. Recombination involves DNA repair, a process known to be biased toward G+C-richness in mammals (Brown and Jiricny 1988). It might therefore be expected that recombination hot spots would have elevated G+C content, and this is true at least for chromosome III, where hot spots for double-strand breaks (DSBs) coincide with G+C-rich areas of the chromosome (Baudat and Nicolas 1997). Because DSBs tend to be located in intergenic sequences, the ensuing DNA repair may affect the ORFs on each side of the DSB and thus contribute to the correlation of GC3s in neighboring genes. A third possibility is raised by the discovery that the genome is partitioned into distinct replicational and transcriptional domains in the nucleus during S-phase (Wei et al. 1998). If these domains are set up anew during each cell cycle, then neighboring genes may tend to experience similar chemical environments during their evolution, whereas genes that are not close together may not have this shared history.

Interchromosomal Differences in G+C Content

The disproportionate concentration of high-GC3s ORFs on shorter chromosomes gives rise to the negative correlation between chromosome length and chromosome G+C content. A negative correlation has also been reported for the relationship between chromosome length and genetic map length per kilobase (Mortimer, Contopoulou, and King 1992). It is a requirement for meiosis that there be at least one chiasma per chromosome, and this results in a higher chiasma density and a longer map length per kilobase on shorter chromosomes. High chiasmata density is associated with high G+C content in humans (Ikemura and Wada 1991), so the differences in recombination rates per kilobase between chromosomes could cause the observed phenomenon. The relationship between chromosome length and GC3s content might therefore have been anticipated. What is interesting is that this relationship is produced by ORFs of high GC3s content and not reflected in a difference in the modal GC3s content. This suggests that the GC3s content of only a subset of the ORFs is affected by DSBs.

The Paradox of Chromosome III

Chromosome III has consistently shown the strongest clustering effects under the objective criteria that have been applied in this study. No other chromosome displays such pronounced regional variation in GC3s (fig. 1) or such autocorrelations at either short or long distances (fig. 2). Why is chromosome III different? We consider three possibilities below.

First, the chromosome III sequence was published in 1992 and is widely thought to be less accurate than other yeast chromosome sequences. Frameshift sequence errors could increase GC3s values for some genes, but such errors seem unlikely to produce the

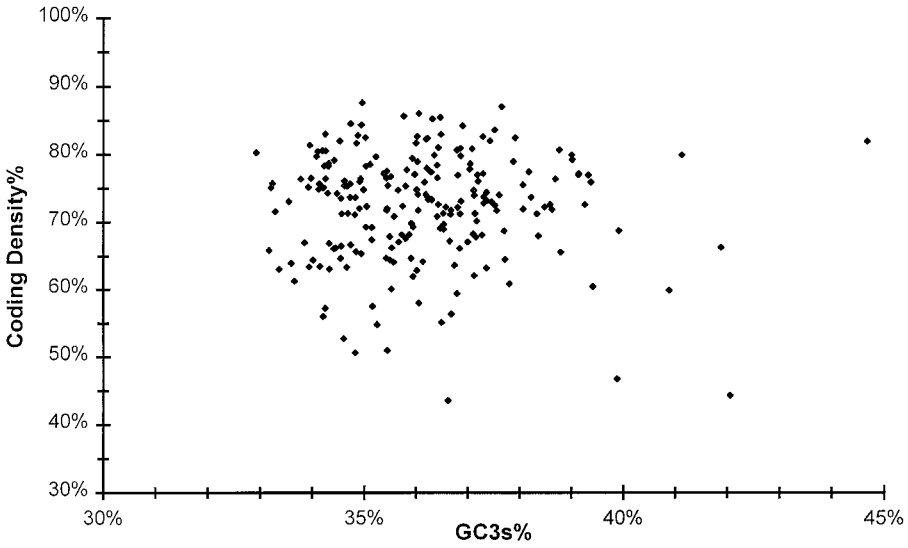

FIG. 5.—Relationship between gene density and GC3s. Each point represents a 50-kb window of sequence from the yeast genome. See text for calculation of gene density.

strong GC3s correlations between neighboring genes seen in figure 2.

A second possible (although unlikely) reason for the differences in chromosome III is that parts of its sequence could be derived from yeast species other than *S. cerevisiae.* The laboratory yeast strain (S288C) whose genome was sequenced is derived largely from a single natural isolate of *S. cerevisiae* (EM93), but small fractions of its genome (probably less than 5% in total) come from two other species: ''*S. microellipsoides*'' strain NRRL-210 (which is possibly *Zygosaccharomyces microellipsoideus*) and the lager yeast *Saccharomyces carlsbergensis* (Mortimer and Johnston 1986). Some sequences from *S. carlsbergensis* show only 82%–84% DNA sequence identity to *S. cerevisiae* (e.g., MET2; Hansen and Kielland-Brandt 1994). *Saccharomyces cerevisiae* strain EM93 has been preserved in yeast stock centers, so it should be possible to use chip technology (e.g., Winzeler et al. 1998) to identify which parts of the S288C genome do not come from this isolate. At present, there is no information about the location of this ''foreign'' DNA in S288C and no reason to suspect that there is more of it on chromosome III than elsewhere.

A third possibility, which we consider to be the most likely, is that chromosome III is unique among yeast chromosomes because it contains the mating-type loci. These comprise the MAT locus and the two silent mating-type cassettes (HML and HMR) located near the two ends of the chromosome. The mating type switches each generation because MAT**a** cells tend to select the cassette on the left arm (HML, which contains a silent copy of the α gene) as a donor for gene conversion at MAT, whereas MATα cells tend to select HMR (which contains a silent copy of the **a** gene). If one of the silent cassettes is relocated to a different chromosome, mating-type switching still occurs, but its efficiency is greatly reduced, because the bias in donor selection is lost (Weiler, Szeto, and Broach 1995). It is likely, therefore, that there is selective pressure to preserve mating-type switching as an intrachromosomal reaction, and so to keep most of chromosome III (between HML and HMR) intact. If chromosome III has been largely free from structural disruption, then its pattern of GC3s variation may represent a fundamental pattern of mutation. Other chromosomes, which are not so constrained, may never be able to reveal such clear trends. In this sense, chromosome III might be considered a ''model'' chromosome for the study of mutational phenomena in eukaryotic genomes.

Acknowledgments

This study was supported by the European Communities' 4th Framework Biotechnology Programme (contract BIO4-CT95-0130) and the BBSRC (G04905). We are grateful to Andrew Lloyd and Bénédicte Lafay for helpful discussion.

LITERATURE CITED

ADJIRI, A., R. CHANET, C. MEZARD, and F. FABRE. 1994. Sequence comparison of the *ARG4* chromosomal regions from the two related yeasts, *Saccharomyces cerevisiae* and *Saccharomyces douglasii.* Yeast **10**:309–317.

BAUDAT, F., and A. NICOLAS. 1997. Clustering of meiotic double-strand breaks on yeast chromosome III. Proc. Natl. Acad. Sci. USA **94**:5213–5218.

BENNETZEN, J. L., and B. D. HALL. 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

BERNARDI, G. 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. **29**:445–476.

BOWMAN, S., C. CHURCHER, K. BADCOCK et al. (22 co-authors). 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIII. Nature **387**:90–93.

BROWN, T. C., and J. JIRICNY. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell **54**:705–711.

BUSSEY, H., D. B. KABACK, W. ZHONG et al. (13 co-authors). 1995. The nucleotide sequence of chromosome I from *Sac-*

charomyces cerevisiae. Proc. Natl. Acad. Sci. USA **92**: 3809–3813.

BUSSEY, H., R. K. STORMS, A. AHMED et al. (89 co-authors). 1997. The nucleotide sequence of Saccharomyces cerevisiae chromosome XVI. Nature **387**:103–105.

CHATFIELD, C. 1989. The analysis of time series, an introduction. Chapman and Hall, London.

CHERRY, J. M., C. ADLER, C. BALL et al. (12 co-authors). 1998. SGD: Saccharomyces Genome Database. Nucleic Acids Res. **26**:73–79.

CHURCHER, C., S. BOWMAN, K. BADCOCK et al. (27 co-authors). 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IX. Nature **387**:84–87.

DIETRICH, F. S., J. MULLIGAN, K. HENNESSY et al. (39 co-authors). 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome V. Nature **387**:78–81.

DUJON, B. 1996. The yeast genome project: what did we learn? Trends Genet. **12**:263–270.

DUJON, B., K. ALBERMANN, M. ALDEA et al. (97 co-authors). 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XV. Nature **387**(Suppl.):98–102.

DUJON, B., D. ALEXANDRAKI, B. ANDRE et al. (108 co-authors). 1994. Complete DNA sequence of yeast chromosome XI. Nature **369**:371–378.

EYRE-WALKER, A. 1992. The role of DNA replication and isochores in generating mutation and silent substitution rate variance in mammals. Genet. Res. **60**:61–67.

FELDMANN, H., M. AIGLE, G. ALJINOVIC et al. (97 co-authors). 1994. Complete DNA sequence of yeast chromosome II. EMBO J. **13**:5795–5809.

FILIPSKI, J. 1987. Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. **217**:184–186.

GALIBERT, F., D. ALEXANDRAKI, A. BAUR et al. (57 co-authors). 1996. Complete nucleotide sequence of Saccharomyces cerevisiae chromosome X. EMBO J. **15**:2031–2049.

GOFFEAU, A., R. AERT, M. L. AGOSTINI-CARBONE et al. (633 co-authors). 1997. The Yeast Genome Directory. Nature **387**(Suppl.):5–105.

HANSEN, J., and M. C. KIELLAND-BRANDT. 1994. *Saccharomyces carlsbergensis* contains two functional *MET2* alleles similar to homologues from *S. cerevisiae* and *S. monacensis.* Gene **140**:33–40.

IKEMURA, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.

IKEMURA, T., and K. WADA. 1991. Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. Nucleic Acids Res. **19**:4333–4339.

JACQ, C., J. ALT-MORBE, B. ANDRE et al. (136 co-authors). 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV. Nature **387**:75–78.

JOHNSTON, M., S. ANDREWS, R. BRINKMAN et al. (35 co-authors). 1994. Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. Science **265**:2077–2082.

JOHNSTON, M., L. HILLIER, L. RILES et al. (57 co-authors). 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. Nature **387**:87–90.

KERR, A. R. W., J. F. PEDEN, and P. M. SHARP. 1997. Systematic base composition variation around the genome of *Mycoplasma genitalium,* but not *Mycoplasma pneumoniae.* Mol. Microbiol. **25**:1177–1179.

MCINERNEY, J. O. 1997. Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb. Comp. Genom. **2**:1–10.

MORTIMER, R. K., C. R. CONTOPOULOU, and J. S. KING. 1992. Genetic and physical maps of *Saccharomyces cerevisiae,* edition 11. Yeast **8**:817–902.

MORTIMER, R. K., and J. R. JOHNSTON. 1986. Genealogy of principal strains of the yeast genetic stock center. Genetics **113**:35–43.

MURAKAMI, Y., M. NAITOU, H. HAGIWARA et al. (13 co-authors). 1995. Analysis of the nucleotide sequence of chromosome VI from *Saccharomyces cerevisiae.* Nat. Genet. **10**:261–268.

OLIVER, S. 1995. Size is important, but. Nat. Genet. **10**:253–254.

OLIVER, S. G., Q. J. VAN DER AART, M. L. AGOSTONI-CARBONE et al. (147 co-authors). 1992. The complete DNA sequence of yeast chromosome III. Nature **357**:38–46.

PHILIPPSEN, P., K. KLEINE, R. POHLMANN et al. (88 co-authors). 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV and its evolutionary implications. Nature **387**(Suppl.):93–98.

SEOIGHE, C., and K. H. WOLFE. 1998. Extent of genomic rearrangement after genome duplication in yeast. Proc. Natl. Acad. Sci. USA **95**:4447–4452.

SHARP, P. M., M. AVEROF, A. T. LLOYD, G. MATASSI, and J. F. PEDEN. 1995. DNA sequence evolution: the sounds of silence. Philos. Trans. R. Soc. Lond. B Biol. Sci. **349**:241–247.

SHARP, P. M., and E. COWE. 1991. Synonymous codon usage in *Saccharomyces cerevisiae.* Yeast **7**:657–678.

SHARP, P. M., and A. T. LLOYD. 1993. Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. Nucleic Acids Res. **21**:179–183.

SHARP, P. M., and G. MATASSI. 1994. Codon usage and genome evolution. Curr. Opin. Genet. Dev. **4**:851–860.

SHARP, P. M., T. M. TUOHY, and K. R. MOSURSKI. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14**:5125–5143.

TETTELIN, H., M. L. AGOSTINI-CARBONE, K. ALBERMANN et al. (115 co-authors). 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome VII. Nature **387**:81–84.

WEI, X., J. SAMARABANDU, R. S. DEVDHAR, A. J. SIEGEL, R. ACHARYA, and R. BEREZNEY. 1998. Segregation of transcription and replication sites into higher order domains. Science **281**:1502–1506.

WEILER, K. S., L. SZETO, and J. R. BROACH. 1995. Mutations affecting donor preference during mating type interconversion in *Saccharomyces cerevisiae.* Genetics **139**:1495–1510.

WINZELER, E. A., D. R. RICHARDS, A. R. CONWAY et al. (11 co-authors). 1998. Direct allelic variation scanning of the yeast genome. Science **281**:1194–1197.

WOLFE, K. H., P. M. SHARP, and W. H. LI. 1989. Mutation rates differ among regions of the mammalian genome. Nature **337**:283–285.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

HOWARD OCHMAN, reviewing editor

Accepted January 27, 1999