Nucleic Acids Research, 1993, Vol. 21, No. 3 703–711

Assessments of DNA inhomogeneities in yeast chromosome III

Samuel Karlin, B.Edwin Blaisdell, Ronald J.Sapolsky, Lon Cardon and Chris Burge
Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA

Received September 2, 1992; Revised and Accepted November 20, 1992

ABSTRACT

With the sequencing of the first complete eukaryotic chromosome, III of yeast (YCIII) of length 315 kb, several types of questions concerning chromosomal organization and the heterogeneity of eukaryotic DNA sequences can be approached. We have undertaken extensive analysis of YCIII with the goals of: (1) discerning patterns and anomalies in the occurrences of short oligonucleotides; (2) characterizing the nature and locations of significant direct and inverted repeats; (3) delimiting regions unusually rich in particular base types (e.g., G+C, purines); and (4) analyzing the distributions of markers of interest, e.g., delta (δ) elements, *ARS* (autonomous replicating sequences), special oligonucleotides, close repeats and close dyad pairings, and gene sequences. YCIII reveals several distinctive sequence features, including: (i) a relative abundance of significant local and global repeats highlighting five genes containing substantial close or tandem DNA repeats; (ii) an anomalous distribution of δ elements involving two clusters and a long gap; (iii) a significantly even distribution of *ARS*; (iv) a relative increase in the frequency of T runs and AT iterations downstream of genes and A runs upstream of genes; and (v) two regions of complex repetitive sequences and anomalous DNA composition, 29000–31000 and 291000–295000, the latter centered at the *HMRa* locus. Interpretations of these findings for chromosomal organization and implications for regulation of gene expression are discussed.

INTRODUCTION

With the rapid accumulation of DNA sequence data, at least 15 contigs exceeding 100 kb length have been generated, including the genomes of herpesviruses HSV1, VZV, EBV, CMV, HHV-6, EHV, HVS, and CCV; the poxvirus, vaccinia; the chloroplasts of tobacco, rice, and liverwort; two substantial *E.coli* contigs; and the first complete eukaryotic chromosome, yeast chromosome III (YCIII). Many types of genomic local and global compositional heterogeneities have been characterized, including mobile elements [1], satellite repeats [2], isochore compartments [3], HTF islands [4], telomeric sequences [5], recombinational hot spots (e.g., Chi elements) [6], under-representation of TpA dinucleotides [7], CpG suppression in vertebrates [8], rarity of the tetranucleotide CTAG [9], GNN periodicity in coding regions [10], methyl transferase targets [11], and repeated extragenic palindromes (REPs) in *E.coli* [12]. Thus, genomic organization is complex and variegated on many scales.

Methods and concepts introduced below can serve as a prototype for assessment and interpretation of inhomogeneities in long DNA sequences. We focus in this paper on the sequence analysis of YCIII [13] to exemplify the methods and to point out a number of potentially interesting regions of the sequence. Accordingly, extensive statistical analysis of the YCIII sequence was performed with the aims of: discerning patterns and anomalies of di-, tri-, and tetranucleotide representations; identifying significantly long direct and inverted repeats; delimiting regions unusually rich in certain DNA types (e.g., C+G, purines); analyzing the counts and spacings of marker points such as poly-*X* (*X* = A, T, C, or G) and poly-*XY* (e.g., AT, CA iterations), delta (δ) elements, *ARS* (autonomous replicating sequences), the special tetranucleotide CTAG, 10-bp palindromes, close repeats, close dyads, and gene sequences. Marker arrays and compositional heterogeneity will be analyzed employing three principal methods: (a) *r*-scan statistics to characterize extremes in marker spacings [14, 15]; (b) plots of position-dependent marker frequencies over a sliding window; and (c) quantile tables to contrast different count distributions [15, 16], see Methods section.

It is well-established that once during each *Saccharomyces cerevisiae* haploid cell cycle an appropriate switch of the mating type occurs. This event is mediated by double-strand cleavage at the *MAT* locus of YCIII, which is repaired using homologous information at one of the silent loci *HMLα* or *HMRa* [17, 18]. The cleavage operation and subsequent DNA transpositions surely entails for YCIII DNA the potential for aberrations, instabilities, DNA duplications, and increased recombination. Abnormal consequences accompanying mating type switching may include chromosomal aneuploidy as well as excisions and amplifications in YCIII [19]. Another unusual feature of YCIII is the occasional formation of a stably propagated ring chromosome resulting from recombination between *HMLα* and *HMRa* [20].

How are the DNA rearrangements persistent in the haploid state in YCIII and the genetic diversity accumulating in the diploid state reflected in sequence features? Our analysis of YCIII reveals the following global features: (i) a relative abundance of significant local and global repeats, including five genes displaying substantial close and tandem DNA repeats; (ii) the anomalous

distribution of δ elements showing two clusters and a significant long gap; (iii) a significantly even distribution of *ARS*; (iv) biased locations of poly-A, poly-T, and poly-AT runs relative to genes; (v) two complex regions containing multiple types of repetitive sequences and anomalous DNA composition, 29000−31000 and 291000−295000, the latter centered at the *HMRa* locus.

METHODS

Evaluation of compositional biases in short oligonucleotides

Let f_x denote the frequency of the nucleotide X (A, C, G, or T) in the sequence, f_{xy} the frequency of the dinucleotide XY, f_{xyz} the frequency of the trinucleotide XYZ, etc. A standard assessment of dinucleotide bias is through an 'odds' ratio, namely $\varrho_{xy} = f_{xy}/f_x f_y$. The measure ϱ_{xy} is suitable for a single sequence, but in comparing sequences to account for the antiparallel structure of double-stranded DNA, we use the symmetrized formula $\varrho^*_{xy} = f^*_{xy}/f^*_x f^*_y$ where $f^*_x = (f_x + f_{\bar{x}})/2$ and $f^*_{xy} = (f_{xy} + f_{(xy)\bar{i}})/2$, etc. where $\bar{X}_i$ and $(XY)_{\bar{i}}$ refers to the inverted complement of X and XY, respectively [7]. The formula to evaluate trinucleotide bias is $\gamma^*_{xyz} = (f^*_{xyz} f^*_x f^*_y f^*_z)/(f^*_{xy} f^*_{xnz} f^*_{yz})$. Analogous formulas for higher order oligonucleotides exist [7]. Table 1 describes results on biases of short oligonucleotides in YCIII relative to its own composition and relative to overall yeast sequence composition.

r-scan statistics

In the study of genomic organization, the general problem arises of how to characterize anomalies in the spacings of a specified marker (e.g., restriction sites, purine tracts, genes, nucleosome placements, palindromes). How does one assess excessive clustering (too many neighboring short spacings), overdispersion (too many consecutive long gaps between markers), or too much evenness (too few short spacings and/or too few long gaps)? If n markers are distributed randomly on a unit interval the minimum spacing m^* follows the distribution $Prob\{m^* \geq x\} = (1-(n+1)x)^n$ for $0 < x \leq 1/(n+1)$ and the maximum spacing M^* follows the distribution $\text{Prob}\{M^* \leq y\} =$

$$\sum_{i=0}^{n+1} \binom{n+1}{i} (-1)^i [\delta\ (1-iy)\ ]^* \quad \text{for} \quad 1/(n+1) \leq y \leq 1$$

where $\delta = 1$ if $iy < 1$ and 0 otherwise. These distributions can be used to test whether m^* is too small or large and similarly for M^*, see [14, 15]. The above formulas are practical for small or moderate n. For n large, we use the r-scan statistics based on asymptotic formulas where $M^*_r(m^*_r)$ is the largest (smallest) of the cumulative lengths generated by $r+1$ successive marker points. The relevant probabilities are $\text{Prob}\{m^*_r \geq x/n^{(1+1/r)}\} \approx \exp\{-x^r/r!\}$ and $\text{Prob}\{M^*_r \leq n^{-1}[\ln n + (r-1)\ln(\ln n) + x]\} \approx \exp\{-e^{-x}/(r-1)!\}$, see [15]. The latter formulas are versatile and less sensitive to sampling fluctuations. By varying r they are also capable of discriminating DNA (or protein) sequence patterns on different scales. Table 2 (see also Figure 1 and Legend) describes results of the r-scan statistics (for $r = 1, 2, 3, 5, 7, 10$) for various marker arrays, including poly-A, poly-T, and poly-AT runs, delta elements, *ARS*, gene positions, CTAG sites, close repeats, close dyads, and all 10-bp palindromes.

Sliding window counts

The description of marker arrays via counts within sliding windows provides a graphical representation of patterns of sequence composition. This procedure can be applied broadly, but for ease of exposition we describe the procedure for close repeats (CR) and close dyads (CD). In this case, we choose a minimal stem length (typically $s \geq 8$ bp for CR and CD), and a maximal loop length ($l \leq 50$ bp or 150 bp) and specify all 5′ positions of the first stem as marker points. Such stem and loop constraints typically engender total counts of 200−1000 for sequences in the 100−500 kb range. Regions of significantly high counts in CR of lengths ≥ 8 bp with distance between copies ≤ 50 bp and ≤ 150 bp and CD with stem lengths ≥ 8 bp and loop lengths ≤ 50 bp and ≤ 150 bp were evaluated as follows. Counts were cumulated in a sliding window of 1 kb traversing YCIII and in a 111 kb contig of *E.coli* by displacements of 0.5 kb (Figure 2). Significant peaks can be determined via the formulas of the r-scan statistics. Accordingly, consider a sequence, scaled to length 1, containing n occurences of a given marker. Let $N(t)$ = the number of markers in the interval $(0, t)$. The theoretical critical value, r^*, corresponding to a given significance level, p (e.g., $p = 0.01$), for the marker counts in any window of size w can be calculated based on the formula

$$Prob\{\max_{w \leq t \leq 1} [N(t) - N(t-w)] > r\} = Prob\{m^{(r)} \leq w\} \approx 1 - e^{-x}\frac{x^r}{r!} = p$$

where $w = x/n^{1+1/r}$ and $m^{(r)}$ is the length of the minimal r-scan. Setting $x = wn^{1+1/r}$ gives $n(nw)r/r! + \log(1-p) = 0$, which may be solved numerically to find r^* for given values of n, w and p.

Segmental quantile distributions

Quantile distributions are constructed by dividing the sequence into successive segments of length w with overlap v, and using the count or frequency of the marker of interest within each segmental unit as the sampled variable. The histogram of these values constitutes the quantile density [cf. 15, 16]. We investigated the compositional spectrum of YCIII with respect to three alphabets: G, C (S = strong hydrogen bonding) versus A, T (W); A, G (R = purine) versus T, C (Y); and T, G (K = keto) versus A, C (M) alphabets, based on hydrogen bonding, steric, and/or chemical distinctions. Specifically, we examined the counts of S−W, R−Y, and K−M in a sliding window of length $w = 1.0$ kb and displacement $v = 0.5$ kb. The R−Y and K−M count histogram in YCIII were generally unimodal and centered around 0, while the S−W count histogram was unimodal but heavily skewed toward negative values, suggesting that globally YCIII is fairly homogeneous, see Figure 3. By contrast the S−W count histograms in other organisms generally show a multimodality skewed to one side depending on genomic G+C composition. For example, the S−W and R−Y count histograms of the human β-globin region (73 kb) indicate a multimodality, reflecting considerable patchiness, see Figure 3b. The corresponding histograms of YCIII are given in Figure 3a.

Score-based statistics

Segments significantly rich in a particular base type can be determined by statistics based on score assignments. For example, to detect significant purine stretches assign a score of 1 to A or G bases and a score of s to C or T bases, such that $f_{A+G} + sf_{C+T} = \mu < 0$, where f_{A+G} and f_{C+T} are the frequencies of purines and pyrimidines, respectively, in the sequence, and μ, the expected score/bp, is set to some fixed negative number. Significantly high scoring (purine-rich) segments for a given value of μ can then be determined [15]; making the value of μ more negative corresponds to greater stringency, i.e., less tolerance for nonpurine bases. For a general discussion of score-based statistics, see refs. [15, 21].

Table 1. Over- and under-representation of short oligonucleotides in yeast DNA sequences

Oligonucleotide[a]	Representation value ϱ^{*b} (γ^{*b}) Chromosome III	Chromosome IX[c]	All yeast[d]
TA	0.78	0.78	0.77
CG	0.80	0.82	0.80
CA/TG	1.11	1.09	1.09
AA/TT	1.13	1.14	1.14
CTA/TAG	0.90	0.89	0.89
CCC/GGG	0.90	0.92	0.89
ACA/TGT	0.91	0.90	0.90
GTA/TAC	1.06	1.10	1.07
ACC/GGT	1.08	1.09	1.12
CCA/TGG	1.12	1.12	1.13
CTAG	1.02	0.90	0.98
TNA	0.88	0.86	0.88
CN_2C/GN_2G	1.16	1.12	1.15
CN_5C/GN_5G	1.15	1.13	1.14
CN_8C/GN_8G	1.13	1.12	1.13
$CN_{20}C/GN_{20}G$	1.13	1.07	1.12

[a] N stands for any nucleotide. The spaced dinucleotides XN_iY were evaluated for i = 1, 2, 3, 5, 6, 7, 8, 15, 20, 30, 40, and 50.
[b] The data are shown only for oligonucleotides that have a value of ϱ^* (γ^*) ≤ 0.90 or $\varrho^*(\gamma^*) \geq 1.10$ in any one of the three collections (see Methods, part 1).
[c] 37 kb (Victoria Smith, personal communication).
[d] Combined EMBL Nucleotide Sequence Data Library yeast entries (813 sequences, total length 1.3 Mb).

RESULTS

Over-and under-representations of short oligonucleotides

Table 1 reports significant over- and under-representations of di- and trinucleotides as assessed by strand-symmetric functionals described in Methods, part 1. As in almost every organism, the dinucleotide TA is significantly under-represented [7]. The significant under-representation of the CG dinucleotide is surprising though, since yeast lacks the CpG methylase present in vertebrates. By contrast, other nonvertebrate eukaryotes (e.g., *N.crassa, C.elegans*, and *D.melanogaster*) show normal representations of CG dinucleotides [7].

Over-representation of $C(N)_nC$ or $G(N)_nG$ at displacements corresponding to multiples of period three (n=2, 5, 8 ...) are notable, and likewise, but to a lesser extent, for $A(N)_nA$ and $T(N)_nT$. All other dinucleotide dependencies for displacement between 4 and 50 have essentially normal relative frequencies (data not shown). A period-3 correlation tendency in coding regions was noted previously [10] and at the first two codon sites a G-nonG compositional bias is common [22]. Since YCIII has a large coding fraction (estimated > 70% [13]), the over-representations of GN_nG and CN_nC for n = 2, 5, 8 ... are consistent with the cited observations.

The *lowest* trinucleotide representation in YCIII and overall in yeast sequences is TAG/CTA, a phenomenon observed widely in eukaryotic sequence sets and in most prokaryotic and viral genomes [7]. At the other extreme, the trinucleotide TGG/CCA is over-represented in YCIII as in all eukaryotic sequences studied [7]. These patterns of compositional bias might result from either selection for TGG/CCA or selection against TAG/CTA, since the two trinucleotides differ by a single transition mutation.

The tetranucleotide CTAG, which is under-represented in all bacterial sequences studied and in most phage and viral genomes has normal representation in YCIII. Among other eukaryotes, CTAG is either under-represented (e.g., *N.crassa, X.laevis*, chicken, rabbit) or has normal representations but below average

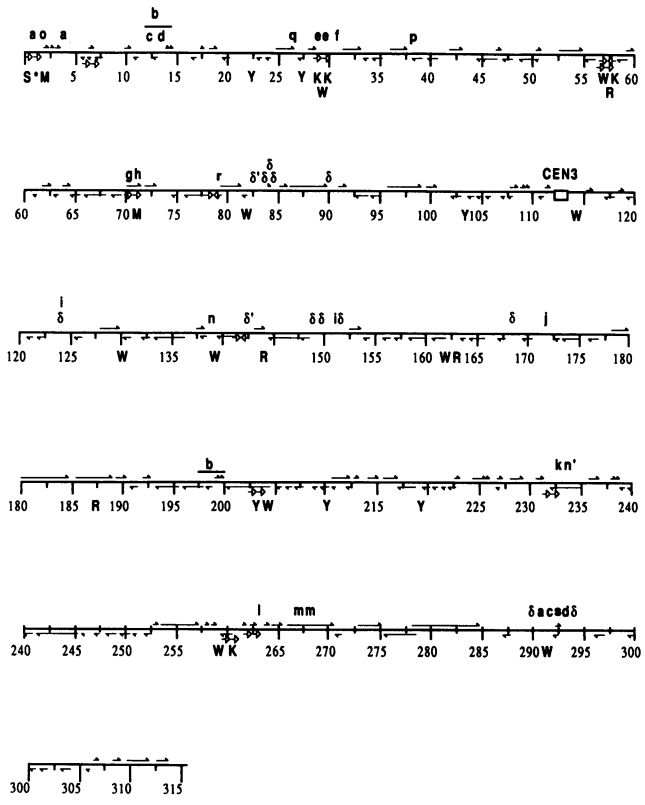

Figure 1. YCIII genome map with special sequence features.

δ delta element δ' inverted delta element

↔ significant clusters of close dyads (evaluated by *r*-scan statistics with r = 3,5,10)[1]

57172 (r = 3,5) 78439 (r = 10) 142149 (r = 3,5)

↔ significant clusters of close repeats (evaluated by *r*-scan statistics with r = 3,5,10)[1]

1--290 (r = 3,5,10) 6191 (r = 3,5) 29285 (r = 3,5,10) 57025 (r = 3) 57929 (r = 10)

70581 (r = 3,5,10) 203146 (r = 3,5) 232460 (r = 3) 260835 (r = 3,5,10) 262958 (r = 3)

High scoring segments[2] :

letter: rich in:	S C+G	W A+T	R A+G	Y C+T	K G+T	M A+C

letter	position	length	letter	position	length	letter	position	length	letter	position	length
S	1	363	W	57011	208	W	130063	88	W	204770	96
M	1	385	R	57162	57	W	138661	107	Y	210333	50
Y	22459	89	K	57917	65	R	144026	79	Y	218907	53
Y	27348	53	M	70958	96	W	162132	53	W	258722	74
K	29285	50	W	82285	71	R	162950	85	K	260583	73
W	29398	131	Y	103670	51	R	187443	95	W	291282	151
K	29427	63	W	114000	131	Y	203138	226			

Long non-delta direct and inverted repeats[3]:

repeat	position	length	adjacent genes	repeat	position	length	adjacent genes	comments
a	1172	251	TEL L76W	a	4065	250	L74W L73C	
				a	290656	246	R95C R98C	between YCA1, MATα genes
b	11499	2508	[L68C - L65W]	b	197402	2508	[R38C - R41W]	mating type genes
c	12230	703	[L67C]	c	291767	704	[R96C]	MATα gene
d	13675	266	[L66W - L65W]	d	293108	266	R97W R98C	MATa gene
e	29285	4 (x12)	L55W L54W	e	29431	4 (x17)	L55W L54W	$(TTTG)_{12}$ and $(TTTR)_{17}$ respect.
f	30842	12 (x2)	L55W L54W					12 bp tandem repeat
g	70572	6 (x10)	[L28W]					codes for $(QG)_{10}$
h	70958	33 (x3)	[L28W]					11 aa. tandem repeat
i	123770	43	R6C R7C	i	151186	43	R18C R19W	3' remnant of δ element
j	171657	18 (x3)	R28C R29C					18 bp tandem repeat
k	232454	30 (x5)	[R67C]					
l	262948	6 (x10)	[R87W]					$(CTGCTT)_{10}$ with a few errors
m	267344	276	[R89W]	m	267671	276	[R89W]	includes 84 a.a. identity
n	139102	21	[R14C]	n'	233621	21	[R67C]	codes for E7, S7 respectively
o	1536	10	TEL L76W					20 bp palindrome
p	38284	10	[L50C]					close dyad: 10 bp stem, 1 bp loop
q	26440	10	[L57W]					close dyad: 10 bp stem, 2 bp loop
r	78897	11	L25C L23C					close dyad: 11 bp stem, 2 bp loop
s	292631	10	R96C R97W					close dyad: 10 bp stem, 3 bp loop

[1] The distributions of close repeats of lengths ≥ 8 bp with distance between copies ≤ 150 bp and close dyads with stem lengths ≥ 8 bp and loop lengths ≤ 150 bp were evaluated by *r*-scan statistics for r = 3,5,10, as described in the Methods. The *r*-scan method assesses clustering, overdispersion, and excessive evenness of a marker array in a DNA sequence. No cases of overdispersion or excessive evenness were detected for either close repeats or close dyads; significant clusters (*r*-scan minimum too small) are listed (see Methods, parts 2 and 3).
[2] Regions rich in a particular base type were determined for the three natural DNA alphabets {S,W}, {R,Y} and {K,M} using score-based statistics as described in the M ethods.
[3] Long direct and inverted repeats, allowing for gaps, were found using the algorithm described in (30). Lengths of repeated segments may differ by a few bases due to insertions/deletions.

Table 2. Assessments of various marker distributions in YCIII

r-scan[a]	observed minimum[b] at[c]		significance[d] evaluation	observed maximum[b] at[c]		significance[d] evaluation
(a) CTAG; 714 occurrences						
1	1	249236	NS	2780	273704	NS
3	63	145134	NS	4496	216479	NS
5	418	272163	LARGE	—	—	NE
0	1349	148167	LARGE	—	—	NE
(b) δ elements; 13 occurrences*						
1	76	83677	NS	121529	168261	NS
2	256	83677	SMALL	—	—	NE
3	425	82671	SMALL	—	—	NE
(c) genes; 183 occurrences						
1	1	31433	NS	3933	147938	NS
3	3	6474	SMALL	9668	286907	LARGE
5	477	48631	NS	—	—	NE
10	1490	42140	NS	—	—	NE
(d) $(A)_n$ ($n \geq 10$); 35 occurrences (Watson strand)†						
1	15	130093	NS	43753	227995	NS
2	351	56803	NS	—	—	NE
(e) $(T)_n$ ($n \geq 10$); 30 occurrences (Watson strand)†						
1	16	235503	NS	47672	30766	NS
2	3500	226207	NS	—	—	NE
(f) $(AT)_n$ ($n \geq 5$); 26 occurrences**						
1	464	223025	NS	45169	108498	NS
2	4525	50296	NS	—	—	NE
(g) nonoverlapping exact palindromes of lengths ≥ 10 bp; 393 occurrences						
1	1	306041	NS	5032	41905	NS
3	204	189005	NS	7910	12533	NS
5	634	188593	NS	—	—	NE
10	2933	31953	NS	—	—	NE

[a] A marker that occurs n times in the sequence induces $n + 1$ spacings. The length of r consecutive spacings (distances) form an r-scan or r-fragment, see Methods, part 2.

[b] A marker of length l occurring at sequence positions i and $j > i$ induces a spacing of length $j - i - (l - 1)$, where l is the length of marker. Displayed are the minimal and maximal sized r-scans over the entire sequence where markers are reduced to single points (see also Results).

[c] Sequence position of the first nucleotide in the marker.

[d] For the low frequency markers (b), (d), (e), and (f), spacings were evaluated by the exact formulas for the high frequency markers (a), (c), and (g), r-scans were evaluated by the asymptotic formulas. The significance level was set at 1%; thus, 'LARGE' means that the probability of observing a value larger than the tabulated value is at most 0.01 for a corresponding random sequence, and 'SMALL' means that the probability of observing a value smaller than the tabulated value is at most 0.01. NE = not evaluated (statistics do not apply). NS = not significant.

*δ or δ' locations: 82671 (partial, about 138 bp), 83677 (partial 151bp), 83903 (complete about 333 bp), complete copies at 84415, 90047, 123738, 142458, 149187, 149930, 151231, 168261, 290110, 293708.

† See text on poly-A and poly-T orientations.

**(AT)n, $n \geq 5$, locations all noncoding: 9194 (n = 6, distal from genes), 12325 (8, 37 bp 3' to L67C), 22369 (6, 35 bp 3' to L59C), 37761 (5, 42 bp 3' to L50C), 46946 (6, 11 bp 3' to L46W, 13 bp 3' to L44W), 50296 (13, 100 bp 5' to L43C), 52373 (9, 64 bp 3' to L40W), 54863 (5, 14 bp 3' to L39W, 41 3' to L38C), 62715 (10, 30 3' to L34W, 10 bp 3' to L33C), 65838 (5, 55 3' to L30C), 71379 (5, 49 3' to L28W), 90899 (8, 20 5' to L18W), 108498 (8, 111 3' to L5W), 153682 (6, 49 3' to R19W, 29 3' to R20C), 189005 (7+5, 108 3' to R33W), 198228 (8, 37 3' to R39C), 212602 (5, 97 5' to R50C), 223025 (7, 67 5' to R59C, 88 5' to R60W), 223502 (11, 43 3' to R60W), 223719 (10, distal), 258041 (6, 40 3' to R82W), 258761 (10, 68 3' to R83W), 272145 (5, distal), 291307 (11, distal), 291861 (9, 37 5' to R96C), 314249 (14, distal).

frequency (e.g., *C.elegans*, *D.melanogaster*, human, and overall yeast sequences). Interestingly, CTAG has the lowest representation value among all tetranucleotides in a 37-kb cosmid of yeast chromosome IX (sequence kindly communicated to us by Victoria Smith).

CTAG sites feature prominently in the consensus target sequences of the *trpR*[23] and *metJ* repressors [24]. Experiments suggest that the tetranucleotide CTAG has a tendency to or can be induced to 'kink' in these complexes and may concomitantly generally be deleterious to DNA stability. From this perspective, CTAG would be used selectively and, therefore, would occur rarely. It is intriguing that the only cluster of CTAG sites in the human cytomegalovirus genome occurs in the oriLyt region and the same is true for the Epstein–Barr virus [15, 25]. By contrast, in YCIII the spacings of CTAG sites (see Table 2) are significantly even (the minimum spacings are not as small as expected), suggesting that YCIII DNA can tolerate many widespread CTAG sites without chromatin aberrations.

Distribution of delta (δ) elements

Thirteen δ elements (of length about 333 bp) are displayed in Figure 1 and in the Legend for Table 2 (as in ref. [13]). The δ distribution in YCIII was evaluated relative to a random model of 13 independent points distributed over 315 kb using the r-scan (r = 1, 2, 3) statistics, see Methods, part 2. The minimum gap between successive deltas (76 bp) was not significant (P = .046), whereas the 2-scan and 3-scan minima (256 bp and 1125 bp, respectively) identify the region 82671–84750 as a highly significant ($P < .001$ in both cases) cluster, in accord with the designation of this region as a transpositional hot spot [26]. An intriguing observation is the presence at position 151186 of repeat **i** (Figure 3) composed of a 44 bp sequence abutting the δ at 151230. Repeat **i** is dyad-symmetric to a corresponding 3' portion of the same δ sequence. This arrangement can most easily be explained via a Ty insertion into a δ at 151230 followed by two recombination events at this location.

Spacings of gene sequences

We constructed the marker array corresponding to all gene (ORF) sequences (≥ 100 aa. length) in YCIII. Each gene sequence was reduced to a point (overlapping genes to consecutive points) reducing the effective chromosome size to 107 kb. The r-scan (r = 3, 5, 7, 10) analysis was applied to the resulting marker array. One 3-scan cluster of genes emerges comprised of four overlapping genes starting at position 6474. There was one significant overdispersion (large 3-scan) in gene locations of about 10 kb extending from 286907–296575, centered at the *HMRa*locus.

Global significant repeats

Long global direct and inverted repeats were evaluated with reference to random sequences of the same composition; precise criteria for statistical significance are reviewed in [27]. Relatively long and distant repeats often arise from transposition, recombination, RNA reverse transcription, multiple rereplication, and gene amplification, during the replication cycle or transcriptional processing, often under conditions of stress.

Repeat **b** in Figure 1 reflects the duplication of the *HMLα* and *MAT* loci; repeats **c** and **d** correspond to the common sequences of the *HMLα* and *HMRa* loci [17, 18]. Several repetitive structures including **g** and **h** (see Figure 1) occur in the coding region of YCL28w after its transmembrane domain, most notably a 10-fold iteration of the diresidue Gln-Gly encoded by the perfect DNA repeat $(CAAGGT)_{10}$. This precise identity suggests that these repeats are of recent origin. Four other more complex

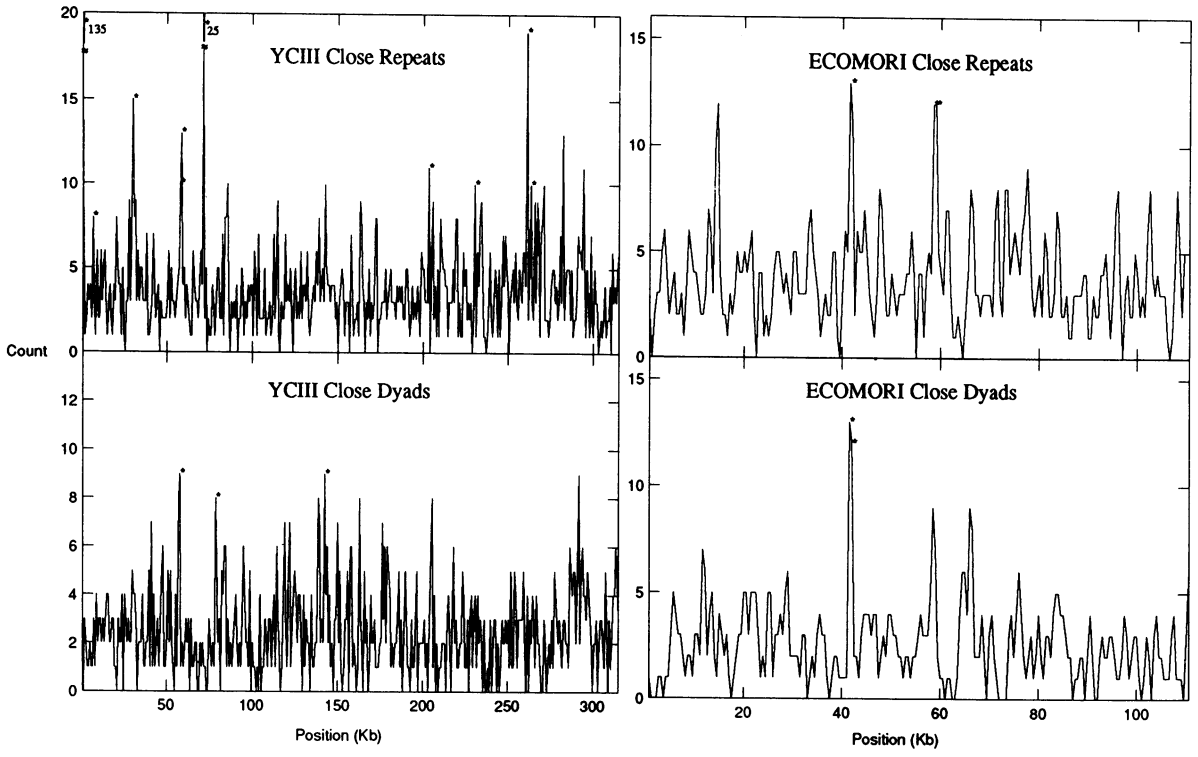

Figure 2. Sliding window plots showing counts of close repeats (top) and close short dyads (bottom) in YCIII and the 111 kb ECOMORI *E.coli*/contig [37]. Close repeats and dyads are of length ≥ 8 bp with ≤ 150 bp between stems or copies. Counts in the sliding windows are cumulated in 1 kb segments with 500 bp displacement. Asterisks indicate statistically significant clusters based on the method of *r*-scans (r = 1, 3, 5, 10).

repetitive sequences are present in the protein, downstream of the Gln-Gly repeat, including a 33-bp three-fold tandem repeat also rich in Gln residues. High glutamine regions are often associated with open coil tertiary structures and putatively are important to gene transactivation function [28].

The large 276-bp DNA repeats labeled **m** in Figure 1 in the gene YCR89w encode an 84-aa. identity with only two errors. At the protein level these extend to a five-fold 36-aa. repeat with diminished DNA conservation for three of the copies. These identities establish similarity to the *AGA1* gene product (the *a*-agglutinin core subunit) containing two imperfect copies of the 36-aa. segment.

The approximate ten-fold 6-bp tandem repeat (**l**) coding for $(LV)_3(LL)_6LV$ in the gene YCR87w is noteworthy. The lengthy C-terminal portion of YCR67c harbors several Ser homopeptides, S_{10}, S_6, S_4, and six copies of S_3. The S_{10} homopeptide derives from the single codon TCT, S_6 is translated from $(TCTTCA)_3$ and S_4 from TCY codons. Interestingly, the C-terminal S_3 alone is encoded from the alternative codon form $(AGT)_3$. Near the C-terminus, YCR67c carries a five-fold 10-aa tandem repeat (**k**), which derives from a significant 30-bp repeat at the DNA level. The oligonucleotide **n** = $(GAA)_7$ is translated to $(Glu)_7$ in gene YCL14c, whereas an inverted complement n′ is part of the $(T-CT)_{10}$ iteration in gene YCR67c, mentioned above.

Regions rich in close repeats

Our criteria for a close repeat and for a close dyad are defined in Methods, part 3. From the *r*-scan statistics and a sliding window assessment, we determined three regions abundant in close dyads and ten regions abundant in close repeats. The telomere element, 1−360, contains a multitude of tandemly repeated C_nA polynucleotides including $C_3ACACA-C_2ACAC_3ACA$ six times and C_2ACAC_3ACACA occurs ten times. All A nucleotides are singletons. The other regions rich in close repeats highlight five genes (see also previous section). Thus, the segment 269250−270750 (enclosed in the gene YCR89w) scores significantly high in close repeats preponderant in Ser homopeptides and coded by homo-oligonucleotides. The segment 57346−58743 in YCL37c, an unknown gene, carries four aa. repeats of two copies each (QEDE, RKKK, KDGF, HNSN, one letter code) based mostly on DNA identities. Among the 116 YCIII ORFs studied, the YCR33w gene product is distinguished by a significant excess of multiplets (homopeptides X_2, X_3, X_4, etc., see [29]). The count of multiplets provides a measure of homopeptide density in the protein sequence. The mechanisms for and significance of high multiplet counts are unknown. Examples of proteins in other species with a high aggregate of multiplets include the *myc* gene family and several *Drosophila* developmental control proteins (e.g., *sevenless, cut*). Among 12 known yeast proteins in the SWISS-PROT database containing high multiplet counts, five are transcription factors, or regulators of such factors (*ABF1, HAP4, REB1, SNF2, SCH9*), while three others are cell cycle control proteins (*CDC25, CLS24, SSD1*).

The segment 260000−261750 carrying high counts of local repeats covers the genes YCR84c and YCR85w; the former translates to the protein *TUP1* responsible for transcriptional repression in several systems [31]. This protein is especially rich in Pro, Thr, and Gln clusters encoded in each case mostly by a single codon.

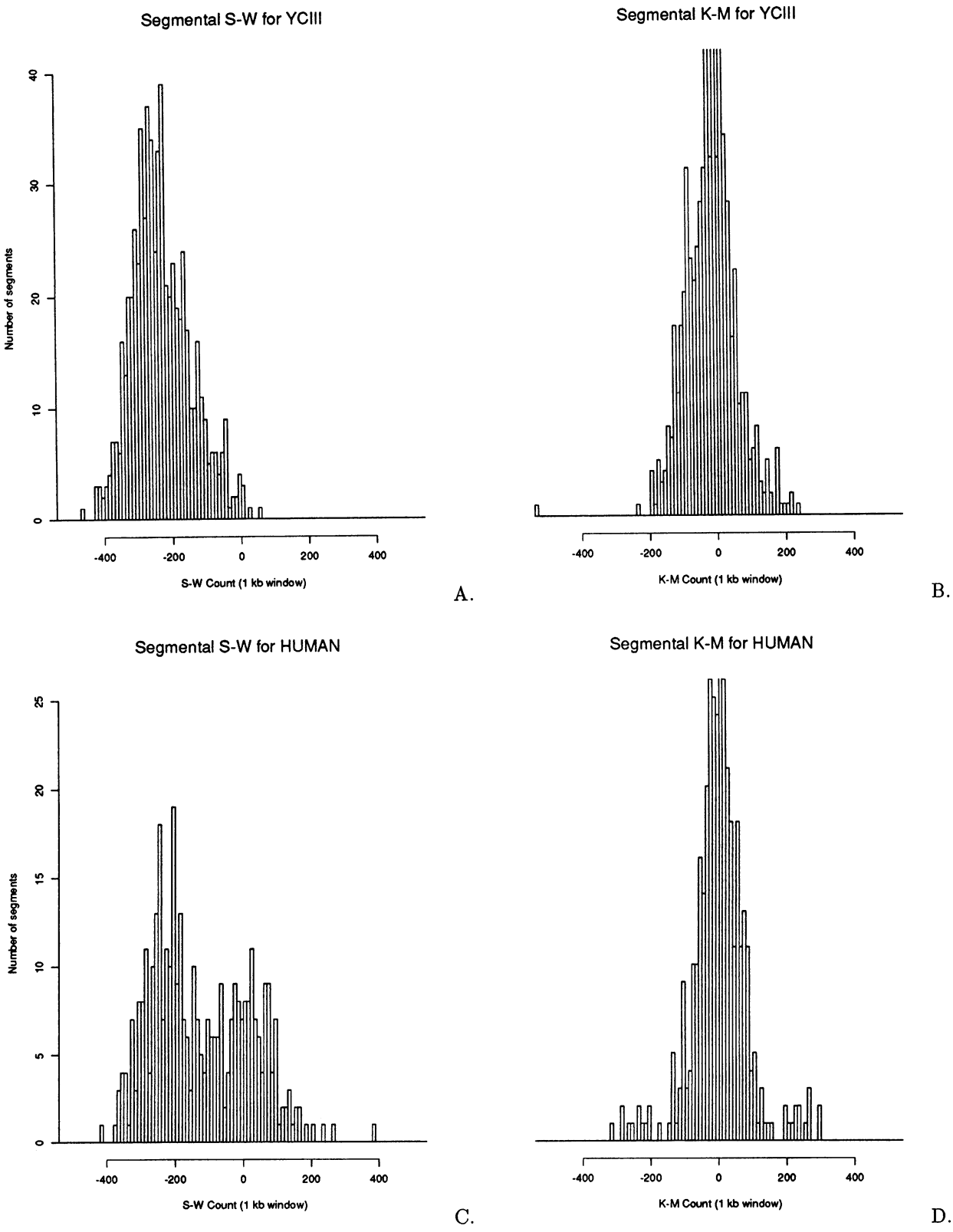

Figure 3. Histograms of S−W (A+G)−(C+T) and K−M (G+T)−(A+C) counts in sliding 1 kb windows with 500 bp displacement. A. and B. show S−W and K−M distributions for YCIII; C. and D. show corresponding plots for a collection of human sequences of combined length 184 kb. Human sequences include the beta globin region (HUMHBB), adenosine deaminase gene (HUMADAG), factor IX gene (HUMFIXG), and the tissue plasminogen activator gene (HUMTPA).

Distributions of palindrome and close dyad symmetry pairings

There occur 393 distinct exact palindromes 10 bp or longer in YCIII. The *r*-scan statistics revealed no distributional anomalies of these palindrome occurrences. This contrasts with similar analyses of the human cytomegalovirus genome, which revealed two 5- and 10-scan 10 bp palindrome clusters, one at the oriLyt region and the other in a major enhancer region (see [25]). Similarly, the Epstein–Barr virus genome shows a single 5-scan cluster of 10-bp palindromes, also occurring at its oriLyt domain [15].

The 3-scan evaluations for the marker array induced by close dyad pairings (stem $\geq$ 8 bp and loop $\leq$ 50 bp) identified a single statistically significant segment commencing at 114007 (near the centromere) and extending 17 bp (data not shown). For dyads with loops $\leq$ 150 bp, highly significant 3- and 5-scan clusters occur at 57172 (between genes YCL38c and YCL37c) and at 142149 (between genes YCR15c and YCR16w and proximal to a δ).

Distributions of poly-A, poly-T and poly-AT sites

On the Watson strand there are 33 poly-A runs and 30 poly-T runs, each determined as containing a core A_{10} (T_{10}) or longer sequence (multiple A_n or T_n iterations $n \geq 10$, separated by less than three errors are united). Five of these are contained in gene regions, four encoding polylysine peptides and one encoding polyphenylalanine. There are 26 $(AT)_n$ iterations with $n \geq 5$, all located in noncoding regions, of which 22 are proximal (within 120 bp) to a gene. The distributions in YCIII of poly-A, of poly-T, and of poly-AT (see Table 2), show minimum and maximum gaps within the ranges expected for random distributions.

Regions of complex repetitive structures

The noncoding region 29000–31000 is replete with distinctive DNA repetitive structures including a concentrated pyrimidine stretch, two concentrated keto (T or G) stretches, and a concentrated weak (A or T) nucleotide stretch (see Figure 1). This region also contains the extended tandem repeats $(TTTG)_{12}$ (with four errors) at 29285 and $(TTT(A/G))_{17}$ (with five errors) at 29431, including three copies of $(TTTA)_2TTTG$. The two 24 bp tandem repeats $(ATT_3AATCGA_4CT(G/A)\ CAGCATGT)_2$ at 30842 also stand out. The region 290000–295550 is also rich in local and global repetitive sequences. The sequence at 295512, $AAAA(CAAA)_2TGCT\ (CAAA)_2$ with about 25% mismatches might be construed as a feasible dyad component to the sequences at 29284 or at 29451. A large exact palindrome occurs at 292131 (stem length 10 bp, loop length 3).

There are 24 *ARS* consensus sequences in YCIII. An *r*-scan analysis of their locations reveals the 5-scan and 10-scan having significantly even spacings (a high minimum). Such an even distribution for potential origins of replication seems advantageous. The positions 57051, 152338, 232099 and 291408 contain the core *ARS* consensus (T/A)TTTAYRTTT(T/A) in the midst of highly repetitive and anomalous DNA structures, see Figure 1.

DISCUSSION

Large scale sequence data are forthcoming including physical, genetic and sequence maps from the genomes of many organisms. Acquisition of data generally runs considerably ahead of interpretation. Thus, it would seem timely to develop methods for assessing, classifying, and contrasting heterogeneities within and among long genomic sequences. A key goal is to identify significant departures in the distribution of sequence markers from a random distribution. Relevant statistics encompass: criteria for discerning over- and under-representations of short oligonucleotides; procedures for ascertaining significant local and global direct and dyad repeats; and analyses of counts and spacings of marker points along the sequence such as special oligonucleotides, 10-bp palindromes, insertion elements, iterated dinucleotides, replication origins, and genes. Other marker arrays amenable to distributional analyses (by *r*-scan statistics, quantile distributions, score-based statistics, see Methods and for reviews [15]), but not evaluated in this paper would include single or aggregate versions of recognized regulatory sequences (e.g., AP1, SP1, TATA-box, CCAAT, polyadenylation signals), nucleosome locations, specific or aggregate type II restriction sites, and methylase targets. In this paper we apply these statistics and perspectives to the recently sequenced YCIII. A number of interesting regions of the sequence are identified. We highlight several of our key findings and venture some interpretations, speculations, and experiments suggested by our observations (see also Results).

Repeats

A striking property of YCIII is the abundance of short and long DNA repeats occurring mostly in tandem and within ORFs (excluding the large *MAT* and δ element repeats). The profusion and complexity of DNA identities in the protein products of YCL28w and YCR89w are most notable; four other genes display substantial close repeats (see Results). The near perfect DNA close repeats in several YCIII ORFs may be of recent origin resulting from DNA polymerase slippage and/or unequal crossing-over events. They may have little consequence to the protein's function, or they may provide flexibility as links between domains of conserved structure. Manipulations (mutagenesis, contractions, expansions, rearrangements) on these repeat regions could provide clues as to the function or neutral role of these repeats.

Counts of close repeats (CR) and close dyads (CD)

Close multiple dyad pairings may offer target sequences that can fold into elaborate secondary structures concomitant with efficient binding of dimeric or multimeric proteins. Appropriate close repeat elements might allow for cooperative binding interactions with multiple transcription or replication factors. The position-dependent plots show significantly more CR clusters than CD clusters in YCIII, in the *E.coli* EcoMORI contig (Figure 2), and this appears to be a ubiquitious phenomenon across long DNA sequences. Such asymmetry, numbers of CR occurrences more than CD occurrences, might be accounted for by the relative facility and frequency of polymerase slippage and unequal crossing-over events. Also, clusters of close dyads often appear as part of promoter, enhancer, and terminator sequences governing crucial transcription and replication functions and might thereby be used selectively. The CR and CD plots associated with the *E.coli* contig (Figure 2) highlight a region about position 41190 containing simultaneously significant CR and CD clusters. This dyad and conjoined repeat conglomerate may be an important regulatory sequence [32].

The noncoding regions, 29000–31000 and 290000–295000

of YCIII, are especially rich in local repetitive sequences and in other compositional anomalies, see Figure 1 and Results. The segment from positions 290000–295000, centering at the *HMRa* locus and including an *ARS* site, may be an important chromosomal control region. Because of their complex repetitive structures, these regions seem attractive for experimental manipulation to test for replicational and transcriptional regulatory function.

Segmental quantile distributions for S–W, R–Y, and K–M in YCIII vs. human

The quantile distributions for counts of S–W, R–Y, and K–M in a sliding 1 kb window (see Methods, parts 3 and 4) were calculated for YCIII and a combined set of four long human genomic sequences (beta-globin, factor IX, tissue plasminogen activator, and adenosine deaminase) of total length about 184.7 kb); histograms for S–W and K–M are displayed in Figure 3. The R–Y plots (data not shown) were essentially unimodal for both YCIII and the human sequences with mean/mode around zero and similar variances. The S–W plots, however, are very different. YCIII is essentially unimodal with mode/mean at about −230 corresponding to the A+T bias of yeast DNA. By contrast, the S–W plot for the human sequences is multimodal with a primary mode at about −220 and a secondary mode at about +40. The bimodality of the human plot presumably reflects isochore compartments [3], a phenomenon which apparently does not occur in yeast (at least not in YCIII). The K–M histograms are also contrasting, with YCIII essentially unimodal apart from a single outlier point at the left telomere, whereas the human sequences carry several significant outlier segments. The strongest M concentrations occur in the *line* element *Kpn1*, in the β-globin and factor IX sequences. A strong K region occurs in the β-globin expanse between the ε-globin and G^{γ}-globin genes.

Poly-X and poly-XY runs

Of the combined poly-A and poly-T occurrences, adapted to the orientation of neighboring genes, there are 12 poly-T and five poly-A sequences 3′ to a gene less than 120 bp from the stop codon, whereas, there are 10 poly-A but only three poly-T sequences within 120 bp 5′ to a gene. Why this asymmetry? The following argument may be applicable. In prokaryotes, poly-T runs are common at the 3′ end of transcription units. They correspond to thermodynamically unstable dA:rU hybrids [33]. In this respect, yeast transcription termination may be similar to that in prokaryotes. Another deterrent to poly-A runs downstream in a gene or proximal to it may relate to such a run occasionally acting as an erroneous substrate for a poly-A binding protein (PABP). Selection against poly-A runs in mRNA avoids incorrect PABP interactions with the message [34]. Similarly, poly-T would be expected to occur less often than poly-A at the 5′ flank of genes to reduce possibilities of premature transcriptional release, since dA:rU is less stable than dT:rA.

All 26 poly-AT sequences (see Legend to Table 2) occur in noncoding regions. It is striking that 13 are proximal downstream of a gene, four proximal convergent (downstream of two genes), four upstream proximal, and one proximal divergent. Four poly-AT occurrences are distal (more than 400 bp from any gene). It has recently been established that an AT iteration in yeast functions as a recognition site for mRNA end-formation [35]. This is consistent with the relative abundance of poly-AT occurrences proximal to stop codons compared to other locations of such iterates. Upstream of genes, AT iterates might occur as TATA-box signals.

There were four occurrences of $\{ (AC)_n/(GT)_n,\ n \geq 5 \}$ in YCIII. The $(AC)_n/(GT)_n$ sequences are used widely as microsatellite markers for constructing physical and genetic maps. They are estimated in higher eukaryotes to be, on average, 30 kb apart, but their locations in YCIII are confined to the 5′ half of the chromosome.

We ascertained all poly-X runs, X_n, $n \geq 8$ and $(XY)_n$, $n \geq 4$ in two recently published contigs of *E.coli*, a 91 kb stretch at 85 min. on the circular chromosome [36] and a second 111 kb stretch at 0 min. [37]. Virtually all poly-T and poly-A in both contigs were located in noncoding regions with poly-T mostly proximal downstream of gene segments and poly-A proximal upstream, paralleling the findings in YCIII. The first contig contained a single $(AT)_4$ proximal downstream of a gene and the second contig contained no poly-AT sequence. Both *E.coli* contigs carry many $(CG)_n$, $n \geq 4$ iterates (9 in the former, 10 in the latter), all in genes, consistent with the fact that *E.coli* is known to be relatively rich in alanine and arginine. YCIII has no corresponding CG iterate, although several 10-bp or longer oligonucleotides of only CG components occur in YCIII.

Spacings of δ elements

The strain of YCIII sequenced contains 13 δ elements with a single complete Ty2 insertion element (see Figure 1). The number of solo deltas throughout the haploid yeast genome estimated on the basis of hybridization experiments is about 100 with about 30–40 complete Ty elements [38–40]. From these numbers it appears that YCIII is high in solo deltas but low in Ty elements. Is YCIII more vulnerable to Ty insertions which by recombination reduce quickly to solo deltas? We might speculate that the conversion of Ty insertions to solo deltas results from a misdirected *HO*-nuclease cleavage in the ε portion of Ty with enlargement of the double-stranded break by DNA degradation culminating in recombination between flanking δ elements. Alternatively, ectopic recombination may be more frequent in YCIII than in other chromosomes [40].

The δ elements are not randomly distributed but involve two clusters and one long gap. The interval of 122 kb free of δ elements (168261–290110) is borderline statistically significant ($P \approx .016$), suggesting that this region may be resistant to Ty insertions or that occurrences of δ or Ty are more deleterious in this region than in others.

Perspectives

The relative abundance of DNA repeats in YCIII compared to overall yeast sequences, more than required to code for the observed amino acid repeats in several genes, suggests that YCIII is in a dynamic state. The persistent mating type gene rearrangements of YCIII suggest a higher order chromosome structure which preferentially, as required, brings *HML* or *HMR* to the *MAT* locus. In this context, translocating the left arm of YCIII to a different chromosome (maintaining *HML* in context), essentially abrogates the common mating type switching (I.Herskowitz and J.Margolskee—personal communication). Moreover, interchanging the *HML* and *HMR* regions does not change the mating choice that is apparently determined by the content of the *MAT* locus. However, mutations at the *HML* and *HMR* loci or in their vicinity may curtail appropriate silencing during the process of mating type alterations [41]. Thus, the

mating type switching, accompanying processes and concomitant sequences putatively incorporate mechanisms and controls that promote or modulate DNA flux. To some extent, the dynamic nature of YCIII could be evaluated relative to the other chromosomes by sampling in the wild and then assessing (e.g., by PCR) the nature of genomic variation in the sample focusing especially on the gene segments containing significant DNA repeats.

ACKNOWLEDGEMENTS

We gratefully acknowledge Drs. V.Brendel, A.M.Campbell, and E.M.Mocarski for helpful comments on the manuscript. We are especially indebted to I.Herskowitz for a variety of discussions and clarifications concerning processes of the haploid and diploid life states. Supported in part by NIH Grants HG00335-04, GM10452-29 and NSF Grant DMS91-06974 to S.K.

REFERENCES

1. Berg, D. E. and Howe, M. M. (1989) Mobile DNA. American Society for Microbiology, Washington, D.C.
2. Willard, H. F. and Waye, J. S. (1987) *Trends Genet.*, **3**, 192–198.
3. Bernardi, G., Mouchirond, D., Bautier, C., and Bernardi, G. (1988) *J.Mol.Evol.*, **28**, 7–18.
4. Bird, A. P. (1986) *Nature*, **321**, 209–213.
5. Blackburn, E. H. (1991) *Nature*, **350**, 569–573.
6. Gilson, E., Saurin, W., Perrin, D., Bachellier, S., Hofnung, M. (1991) *Nucleic Acids Res.* **19**, 1375–1383.
7. Burge, C., Campbell, A., Karlin, S. (1992) *Proc.Natl.Acad.Sci.U.S.A.*, **89**, 1358–1362.
8. Josse, J., Kaiser, A. D., Kornberg, A. (1961) *J.Biol.Chem.*, **236**, 864–875.
9. Karlin, S., Burge, C., Campbell, A. (1992) *Nucleic Acids Res.*, **20**, 1363–1370.
10. Ficket, J. W. (1982) *Nucleic Acids Res.*, **10**, 5303–5318.
11. Nelson M. and McClelland, M. (1992) *Nucleic Acids Res.*, **19**, 2045–2071.
12. Krawiec, S. and Riley, M. (1990) *Microbiol.Rev.*, **54**, 502–539.
13. Oliver, S. G. et al. (1992) *Nature*, **357**, 38–46.
14. Karlin, S. and Macken, C. (1991) *Nucleic Acids Res.*, **19**, 4241–4246.
15. Karlin, S. and Brendel, V. (1992) *Science*, **257**, 39–49.
16. Karlin, S., Blaisdell, B. E. and Bucher, P. (1992) *Prot. Eng.*, **5**, 729–738.
17. Herskowitz, I. (1988) *Microbio. Rev.*, **Dec.** 536–553.
18. Herskowitz, I. (1989) *Nature*, **342**, 749–757.
19. Klar, A. J. S., Strathern, J. N., Hicks, J. B. and Prudente, D. (1983) *Mol. and Cell Biol.*, **3**, 803–810.
20. Brandriss, M. C., Soll, L. and Botstein, D. (1979) *Genetics*, **79**, 551–560.
21. Karlin, S. and Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 2264–2268.
22. Shepherd, J. C. W. (1981) *J. Mol. Evol.*, **17**, 94–102.
23. Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. and Sigler, P. B. (1988) *Nature*, **335**, 321–326.
24. Rafferty, J. B., Somers, W. S., Saint-Girons, I. and Phillips, S. E. (1989) *Nature*, **34**, 705–710.
25. Masse, M. J. O., Karlin, S., Schachtel, G. A. and Mocarski, E. S. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 5246–5250.
26. Warmington, J. R., Anwar, R., Newlon, C. S., Waring, R. B., Davies, R. W., Indge, K. J. and Oliver, S. G. (1986) *Nucleic Acids Res.*, **14**, 3475–3485.
27. Karlin, S., Ost, F. and Blaisdell, B. E. (1989) In Waterman, M. (ed.) Mathematical Methods for DNA Sequences. CRC Press, Inc., Boca Raton, 133–157.
28. Mitchell, P. J. and Tjian, R. (1989) *Science*, **245**, 371–378.
29. Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E., and Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 2002–2006.
30. Leung, M.-Y., Blaisdell, B. E., Burge, C. and Karlin S. (1991) *J. Mol. Bio.*, **221**, 1367–1378.
31. Zharg, M., Rosenblum-Vos, L. S., Lowry, C. V., Boakye, K. A. and Zitomer, R. S. (1991) *Gene*, **97**, 153–161.
32. Cardon, L, Burge, C., Schachtel, G., Blaisdell, B. E. and Karlin, S. Comparative DNA sequence features in two long *E.coli* contigs, submitted.
33. von Hipple, R. H. and Yager, T. D. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 2307–2311.
34. Sachs, A. B. (1990) *Curr. Opinions in Cell Bio.* **2**, 1092–1098.
35. Sapolsky, R. J. (1991) Unpublished Ph.D. thesis, 'ARS Protection and mRNA 3′ end formation by the yeast GAL7 transcriptional terminator,' Stanford University.
36. Daniels, D. L., Punkett, G., Burland, V. B. and Blattner, F. R. (1992) *Science*, **237**, 771–778.
37. Yura, T., Mori, H. et al. (1992) *Nucl. Acids Res.*, **20**, 3305–3308.
38. Curcio, M. J., Hedge, A. M., Boeke, J. D., Garfinkel, D. J. (1990) *Mol. Gen. Genetics*, **220**, 213–221.
39. Cameron, J. R., Loh, E. Y., Davis, R. W. (1979) *Cell*, **16**, 739–751.
40. Kupiec, M. and Petes, T. D. (1988) *Genetics*, **119**, 549–559.
41. Weiler, K. and Broach, J. R. (1992) *Genetics*, in press.