GBE

Genome Biology and Evolution

SMBE

Chromosomal G + C Content Evolution in Yeasts: Systematic Interspecies Differences, and GC-Poor Troughs at Centromeres

Denise B. Lynch[1], Mary E. Logue[1], Geraldine Butler[1], and Kenneth H. Wolfe*,[2]

[1]Conway Institute of Biomedical and Biomolecular Sciences, University College Dublin, Belfield, Dublin 4, Ireland

[2]Smurfit Institute of Genetics, Trinity College, Dublin 2, Ireland

*Corresponding author: E-mail: khwolfe@tcd.ie.

Accepted: 3 July 2010

Abstract

The G + C content at synonymous codon positions (GC3s) in genes varies along chromosomes in most eukaryotes. In *Saccharomyces cerevisiae*, regions of high GC3s are correlated with recombination hot spots, probably due to biased gene conversion. Here we examined how GC3s differs among groups of related yeast species in the *Saccharomyces* and *Candida* clades. The chromosomal locations of GC3s peaks and troughs are conserved among four *Saccharomyces* species, but we find that there have been highly consistent small shifts in their GC3s values. For instance, 84% of all *S. cerevisiae* genes have a lower GC3s value than their *S. bayanus* orthologs. There are extensive interspecies differences in the *Candida* clade both in the median value of GC3s (ranging from 22% to 49%) and in the variance of GC3s among genes. In three species—*Candida lusitaniae*, *Pichia stipitis*, and *Yarrowia lipolytica*—there is one region on each chromosome in which GC3s is markedly reduced. We propose that these GC-poor troughs indicate the positions of centromeres because in *Y. lipolytica* they coincide with the five experimentally identified centromeres. In *P. stipitis*, the troughs contain clusters of the retrotransposon Tps5. Likewise, in *Debaryomyces hansenii*, there is one cluster of the retrotransposon Tdh5 per chromosome, and all these clusters are located in GC-poor troughs. Locally reduced G + C content around centromeres is consistent with a model in which G + C content correlates with recombination rate, and recombination is suppressed around centromeres, although the troughs are unexpectedly wide (100–300 kb).

Key words: centromeres, G+C content, fungi, *Saccharomyces*, *Candida*.

Introduction

G + C content varies substantially within the genomes of most eukaryotes, a phenomenon first noted by Bernardi et al. (1985) who introduced the term "isochore" to describe a region of homogeneous G + C content. In the era of genomics, it has become possible to study the isochore structures of genomes in detail and to examine the process by which the G + C content of a gene or genomic region can change during evolution (Eyre-Walker and Hurst 2001; Duret et al. 2006; Duret and Galtier 2009). Pronounced differences can exist both within a genome (e.g., in rice, where the G + C content at synonymous codon positions—GC3s—of different genes ranges from about 43% to 92%; Wang et al. 2004) and between genomes (e.g., human genes tend have more extreme GC3s values, at both ends of the scale, than their mouse orthologs; Mouchiroud et al. 1988).

Variation in the G + C content along yeast chromosomes was first reported in *Saccharomyces cerevisiae* by Sharp and Lloyd (1993). They used a sliding-window approach and found that groups of consecutive genes on *S. cerevisiae* chromosome III ranged from about 35% to 50% in their GC3s values. Later analyses confirmed that similar patterns exist on most other *S. cerevisiae* chromosomes (Dujon 1996; Bradnam et al. 1999). A link between G + C content variation and recombination was initially suggested by the discoveries that smaller chromosomes are slightly richer in GC3s (Bradnam et al. 1999) and have a higher recombination rate per unit length (because a minimum of one crossover must occur on each chromosome per meiosis; Kaback

et al. 1992; Mancera et al. 2008) and that recombination hot spots tend to occur in local peaks of G + C content (Gerton et al. 2000; Birdsell 2002; Mancera et al. 2008). For several years, it was unclear which of these two factors drove the other: whether a high local G + C content could increase the local recombination rate (Gerton et al. 2000; Petes and Merker 2002) or whether the presence of a recombination hot spot could (over time) elevate the G + C content in the nearby genomic region (Birdsell 2002; Marais 2003). Recent evidence from many different eukaryotic systems has pointed toward the latter alternative, and biased gene conversion is likely to be the molecular mechanism by which it occurs (Duret and Galtier 2009). However, this conclusion was called into question recently by Marsolier-Kergoat and Yeramian (2009) and Tsai et al. (2010) who found that recent nucleotide substitution patterns in different genomic regions did not fit the predictions of the biased gene conversion model. Tsai et al. (2010) proposed that this inconsistency is the result of a change in the life cycle of *Saccharomyces* species, caused by the emergence of mating-type switching.

GC-biased gene conversion (gBGC) is a phenomenon in which heteroduplex DNA containing a mismatch between an A/T base and a G/C base is preferentially corrected in favor of the G/C base (for instance, repairing an A:G mismatch so that it becomes a C:G pair rather than an A:T pair). Consequently, a heterozygote will produce more gametes carrying the G/C allele than the A/T allele, increasing the chance that the G/C allele will eventually become fixed in the population. gBGC has been observed to occur in many eukaryotes including *S. cerevisiae* (Birdsell 2002; Marais 2003; Mancera et al. 2008). Regions of heteroduplex DNA are primarily formed during meiotic recombination, so if a species is prone to gBGC then regions of its genome that are recombination hot spots will tend to increase in G + C content to a greater extent than other regions. These increases will be particularly evident at less-constrained positions, such as the synonymous third positions of codons, and will lead to a local correlation between recombination rate and GC3s in species that have variable recombination rates along chromosomes (Galtier et al. 2001; Birdsell 2002; Duret and Galtier 2009). G + C content of intergenic regions may also be informative, but because these regions cannot readily be aligned among the yeast species we consider here (Cliften et al. 2001), it is difficult to be certain that any base composition differences in intergenic regions are the result of mutation processes rather than insertions/deletions. For this reason, our analyses are mostly based on GC3s profiles.

G + C content variation in the *S. cerevisiae* genome was analyzed in great detail because it was the first eukaryotic genome sequence to become available (Goffeau et al. 1996), but this aspect has received much less attention in the other yeast genomes that have been published more recently (more than two dozen species in the Saccharomycotina). Because some of these genomes are known to have average G + C contents that are substantially different from that in *S. cerevisiae* (Dietrich et al. 2004; Souciet et al. 2009) and in view of the dramatic G + C content discontinuity that was recently demonstrated in the *Lachancea kluyveri* genome (Payen et al. 2009), we wondered how the isochore structure of yeast genomes evolves. To examine this topic, we chose to focus mainly on two groups of closely related yeasts for which genome sequences are available: the *Saccharomyces* species (formerly called *Saccharomyces sensu stricto*; Kellis et al. 2003) and the *Candida* clade species (Butler et al. 2009).

Methods

Sequence Data and Species Nomenclature

Sources of downloaded data are given in supplementary table S1 (Supplementary Material online). For data obtained from National Center for Biotechnology Information (*Pichia stipitis*, *P. pastoris*, *Kluyveromyces thermotolerans*, *Zygosaccharomyces rouxii*, *Yarrowia lipolytica*, and *Eremothecium gossypii*) or Génolevures (*L. kluyveri*, *Z. rouxii*, *Candida glabrata*, and *K. lactis*), Perl scripts were used to obtain coding sequences using GenBank—or EMBL—format description files and chromosomal sequences. For *C. parapsilosis*, contig sequences were obtained from the Sanger Institute, and a Broad Institute annotation supplemented with in-house annotations was used. Following the revision of the names of yeast genera by Kurtzman (2003), no species are classified as "*Saccharomyces sensu lato*" and the genus *Saccharomyces* is monophyletic. Therefore, we refer to the clade consisting of *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus* simply as *Saccharomyces* instead of "*Saccharomyces sensu stricto*."

Calculation of GC3s Values

Perl scripts were written to calculate GC3s percentages for each gene. GC3s is the G + C content at the third position of all codons except TAA, TAG, TGA, TGG, and ATG codons. For species that translate the CTG codon as serine instead of leucine, we also excluded CTG codons. Genes with fewer than 100 codons, or whose annotated coding region lengths were not a multiple of three base pairs, were not used in GC3s calculations.

Identification of Orthologs between Species

Orthology was defined by best reciprocal Blast hits of protein sequences. The analysis reported in figure 1 and table 1 used only genes that have orthologs in *S. cerevisiae*, and the analyses shown in figures 2, 4, and 5 and supplementary figure S2H–K (Supplementary Material online) used only

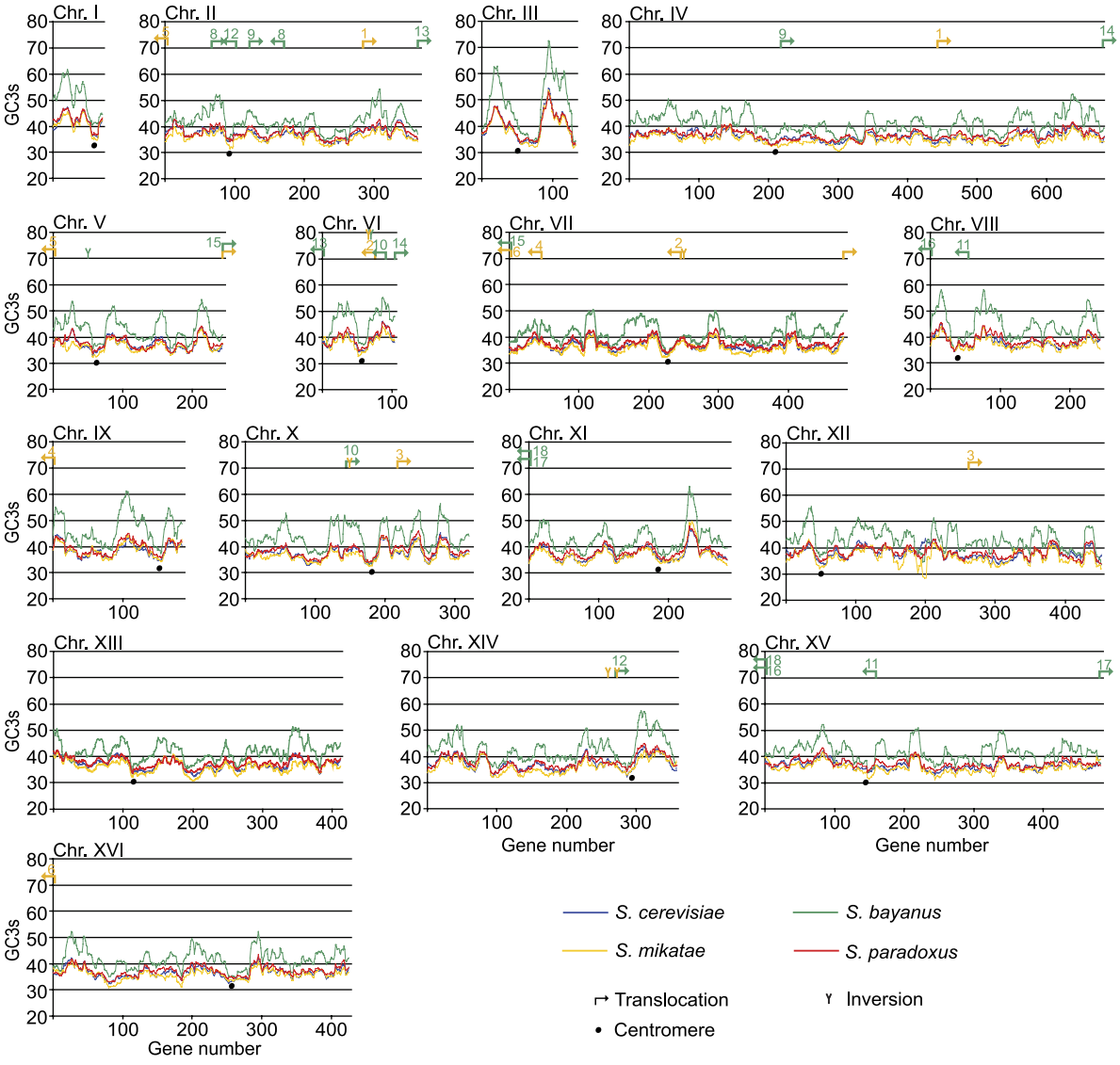

Fig. 1.—GC3s profiles of chromosomes in *Saccharomyces* species. For *Saccharomyces cerevisiae*, the values plotted are a moving average of 15 adjacent genes along each chromosome. Complete chromosomal sequences from *S. paradoxus*, *S. mikatae*, and *S. bayanus* are not available (instead, we have large contigs), so for these species we used the *S. cerevisiae* gene order to calculate the moving average. The locations of known inversions and translocations (Kellis et al. 2003) are indicated; numbers on arrows indicate paired breakpoints. Dots mark the positions of *S. cerevisiae* centromeres. Centromeres in the other three species are inferred to be at the same locations based on experimental data for five *S. bayanus* centromeres (*SbCENa*, *SbCEN1*, *SbCEN6*, *SbCEN7*, and *SbCEN8* corresponding to *S. cerevisiae CEN2*, *CEN1*, *CEN11*, *CEN10*, and *CEN14*, respectively; Huberman et al. 1986; Yamane et al. 1999) and on the presence of the point centromere consensus sequence at all 16 sites in each of *S. paradoxus*, *S. mikatae*, and *S. bayanus* (supplementary file S2e, Supplementary Material online, of Kellis et al. 2003).

genes that have orthologs in *C. albicans*. The analysis in figure 3 used only genes with orthologs in all nine *Candida* clade species. All other analyses used all annotated genes for which a GC3s percentage could be calculated.

Sliding-Window Analysis Method

For *Saccharomyces* species (fig. 1), genes were ordered as found in the *S. cerevisiae* genome. In figure 2, both *C. albicans* and *C. dubliniensis* genes were ordered according to the *C. albicans* genome. For all other species, genes are plotted in their native order. A sliding-window method similar to that of Sharp and Lloyd (1993) was used. We plotted the average GC3s of 15 adjacent genes using a step size of one gene. In figure 1, where a species did not have orthologs of all 15 genes in an *S. cerevisiae* window, the GC3s average was calculated from the remaining available genes. This procedure allows the same number of windows to be obtained from each interspecies comparison.

Table 1
Regression Analysis for GC3s Values in *Saccharomyces* Species

Species 1	Species 2	No. genes compared	Correlation coefficient (*r*)	Slope (*m*)	SE of slope	Intercept (*c*)	SE of intercept
Saccharomyces cerevisiae	*S. bayanus*	4,731	0.86	1.592	0.013	−0.163	0.005
S. cerevisiae	*S. mikatae*	4,917	0.88	0.908	0.007	0.027	0.003
S. cerevisiae	*S. paradoxus*	5,107	0.92	0.910	0.005	0.041	0.002
S. bayanus	*S. mikatae*	4,479	0.86	0.476	0.004	0.160	0.002
S. bayanus	*S. paradoxus*	4,619	0.87	0.468	0.004	0.178	0.002
S. mikatae	*S. paradoxus*	4,807	0.89	0.853	0.006	0.068	0.002

NOTE.—The equation of the best-fit regression line for each comparison is $y = mx + c$, where y is the GC3s value of a gene in species 2 and x is its GC3s value in species 1. All slopes are significantly different from unity ($P < 10^{-39}$ by t-test, calculated using the TDIST function in Microsoft Excel). SE, standard error.

Results

Systematic Shifts in GC3s in *Saccharomyces* Species and between *C. albicans* and *C. dubliniensis*

Saccharomyces paradoxus, *S. mikatae*, and *S. bayanus* are the closest known relatives of *S. cerevisiae* (Kurtzman and Robnett 2003). In terms of their levels of synonymous nucleotide sequence divergence (*K*s) or their levels of amino acid sequence divergence, the *Saccharomyces* species are approximately as different from each other as humans are from rodents (Kellis et al. 2003; Dujon 2006). The 4 *Saccharomyces* species all have 16 chromosomes and these are colinear except for 9 reciprocal translocations and 20 inversions among them (Fischer et al. 2000; Kellis et al. 2003).

We identified orthologous genes between each of the other *Saccharomyces* species and *S. cerevisiae*, calculated GC3s for a moving average of 15 adjacent genes in each species, and plotted these values according to the chromosomal order of the genes in *S. cerevisiae*. For *S. cerevisiae*, this is identical to the approach previously used by Sharp and

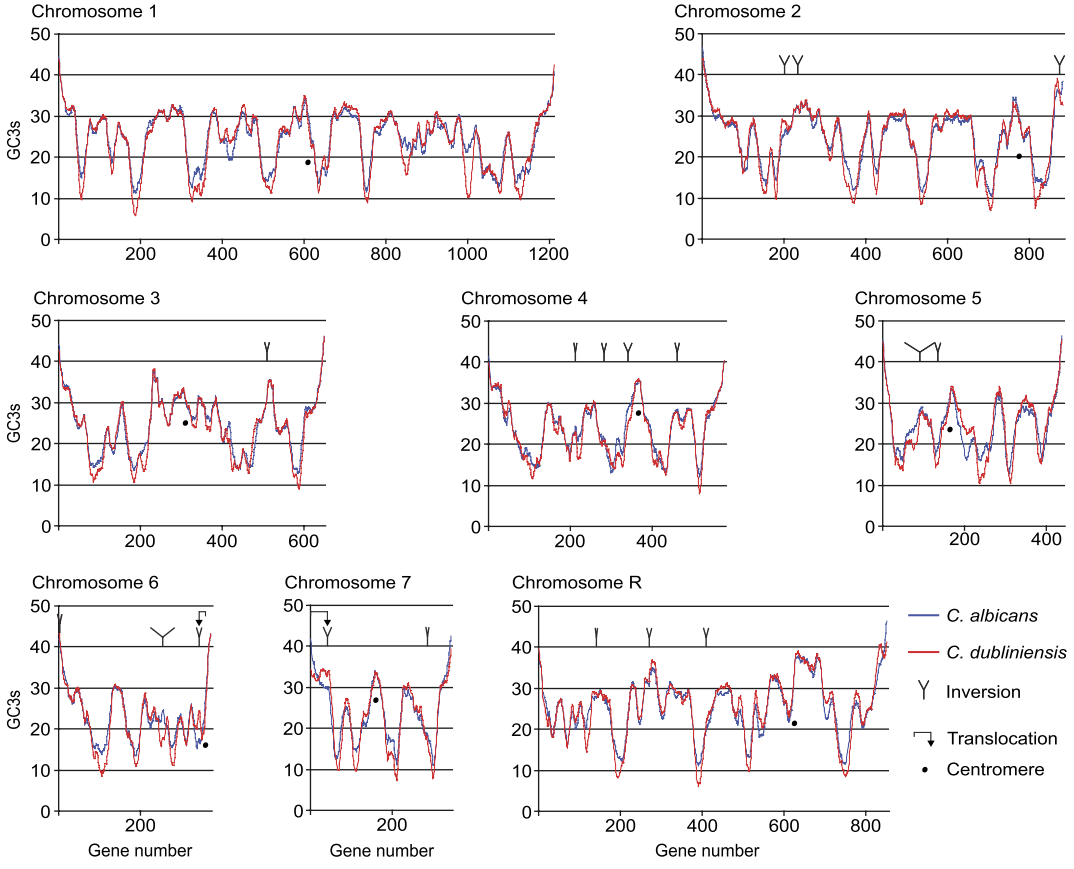

FIG. 2.—Chromosomal GC3s profiles in *Candida albicans* and *C. dubliniensis*. Chromosomal rearrangements of five or more genes are indicated by "Y" for inversions and hooked arrows for translocations. Genes are shown in the order that they occur in *C. albicans*. Dots show the positions of centromeres in *C. albicans* (Sanyal et al. 2004).

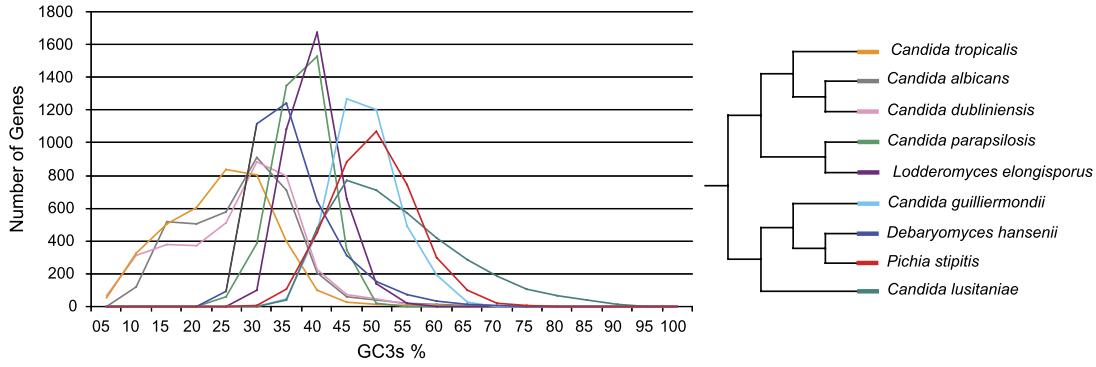

Fig. 3.—Distribution of GC3s values among nine species in the *Candida* clade. A set of 3,687 orthologous genes was identified among all these species, and the values for each species were grouped into bins of 5% intervals. The phylogenetic tree on the right is modified from Butler et al. (2009) and is derived from 644 single-gene families.

Lloyd (1993) and Bradnam et al. (1999), except that we did not weight genes by length. In general, the locations of peaks and troughs of GC3s in the four species coincide (fig. 1), so that, for example, the most G + C–rich region in each of the four genomes is on the right arm of chromosome III. There are, however, some strikingly consistent differences among the species. *Saccharomyces bayanus* has the highest GC3s values and *S. mikatae* has the lowest values throughout the whole genome. The interspecies differences are greatest in the areas around GC3s peaks, whereas in the troughs, all species have more similar GC3s values. When the GC3s values for individual genes are compared between *S. bayanus* and *S. cerevisiae*, the values are seen to be highly correlated ($r = 0.86$), but the slope of the best-fit line is significantly different from 1 (table 1; supplementary fig. S1, Supplementary Material online). A slope different from 1 indicates that the difference in base composition between the species is not uniform across all genes but instead varies systematically, with greater divergence in GC-rich genes than in GC-poor genes. The most GC-rich genes in *S. bayanus* have GC3s of about 90%, whereas their *S. cerevisiae* orthologs have GC3s of only about 67% (supplementary fig. S1, Supplementary Material online). Similarly, the slopes of the best-fit regression lines for all other pairs of *Saccharomyces* species are significantly different from 1 (table 1), indicating systematic and statistically significant patterns of divergence.

The patterns of variation of GC3s in *Candida* species have not been examined before. We used the same sliding-window approach and found that although the *C. albicans* genome is overall more GC poor than the *S. cerevisiae* genome (median GC3s values 26% and 36%, respectively), the *C. albicans* genome contains a similar pattern of alternating regions with different GC3s contents (fig. 2). The troughs in this species reach 11% GC3s, whereas the "GC-rich" peaks are generally only about 35% GC3s, except near the telomeres where GC3s reaches 40–45%. The elevation of GC3s near the telomeres is not restricted to subtelomeric gene families (which are less extensive than in *S. cerevisiae*; van het Hoog et al. 2007) but extends well into chromosomal regions that contain single-copy genes with conserved synteny in other *Candida* species. For example, near the left end of *C. albicans* chromosome 1, the first gene with GC3s <35% is *orf19.6090* which is 28 kb (16 annotated genes) away from the beginning of the chromosome sequence.

The extent of sequence divergence between *C. albicans* and *C. dubliniensis* is approximately the same as among the *Saccharomyces* species (Jackson et al. 2009). To compare their GC3s profiles, we sorted the *C. dubliniensis* genes into the same chromosomal order as their *C. albicans* orthologs and then applied a sliding window of 15 genes (fig. 2). The peaks and troughs in the two *Candida* species are at similar chromosomal locations, but the troughs in *C. dubliniensis* are even more GC poor (reaching a minimum of 6% GC3s).

Variation of GC3s among Species in the *Candida* Clade

We examined the evolution of yeast G + C contents and isochore structures on a broader evolutionary scale using the genome sequences of nine species in the "*Candida* clade"—the clade of species that translate the codon CTG as serine instead of leucine (Butler et al. 2009). The GC3s values of these *Candida* species vary quite widely. The most GC-poor species is *C. tropicalis* with a median GC3s of 22%, and the most GC rich is *C. lusitaniae* with a median of 49% (fig. 3). Viewed from a phylogenetic perspective (Butler et al. 2009), species that are more closely related to each other tend to have more similar G + C contents. There is a GC-poor clade (*C. tropicalis*, *C. albicans*, and *C. dubliniensis*), an intermediate clade (*C. parapsilosis* and *Lodderomyces elongisporus*), and a GC-rich clade (*C. lusitaniae*, *C. guilliermondii*, and *P. stipitis*). *Debaryomyces hansenii* is the only exception to this trend: its GC3s

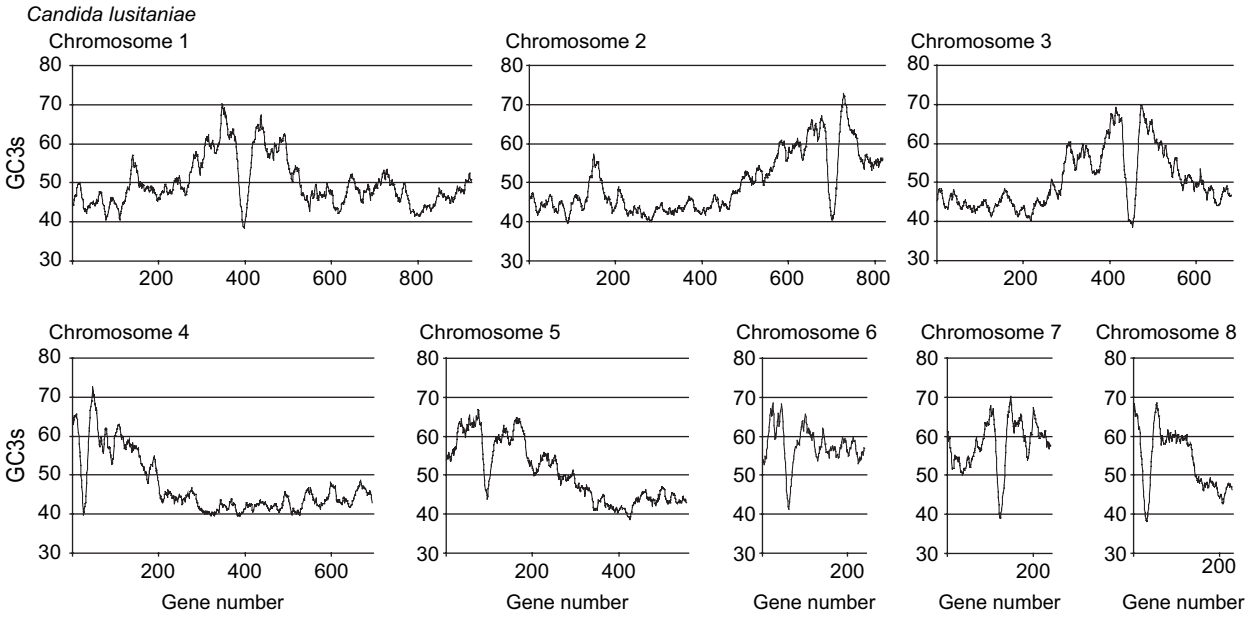

Fig. 4.—GC3s profiles of *Candida lusitaniae* chromosomes.

values are intermediate but it clusters phylogenetically within the high-GC clade, which suggests that *D. hansenii* may be declining in G + C content.

Species at the center of the distribution in figure 3, such as *L. elongisporus*, show a relatively low variance of GC3s values among genes. In contrast, the species at the extremities show both a larger variance and a skewed distribution, with a long tail of GC3s-rich genes in *C. lusitaniae* and long tails of GC3s-poor genes in *C. tropicalis* and *C. dubliniensis*. Sliding-window plots of GC3s content for individual *Candida* clade species are included in supplementary figure S2 (Supplementary Material online). The differences between species in the amount of GC3s variance are illustrated by comparison of the plot for *L. elongisporus*, which is almost flat, to that for *C. tropicalis*, which contains deep troughs (supplementary fig. S2I and J, Supplementary Material online). We do not know the G + C content of the common ancestor of the *Candida* clade, but it is clear that the G + C content of at least some *Candida* lineages must have changed extensively in the time since this ancestor existed and that the G + C contents of groups of neighboring genes on the chromosome tend to change in concert.

GC-Poor Troughs and Centromere Locations

The chromosomal GC3s profiles of some of the more GC-rich *Candida* clade species are very distinctive. The pattern on the larger chromosomes of *C. lusitaniae* is particularly interesting (fig. 4). GC3s is relatively low at the telomeres (~45%), increases gradually toward the center of the chromosomes reaching peaks of ~70%, and then plunges to a narrow trough of ~40%. All eight chromosomes in this species show similar patterns, with one obvious GC-poor trough on each chromosome, but for some chromosomes, the trough is close to one end (chromosomes 2, 4, and 8) and the corresponding telomeric region is not as GC poor as on other chromosomes. *Pichia stipitis* shows a similar pattern of one deep GC-poor trough per chromosome (fig. 5), but this species does not show the same pattern of elevated G + C content on each side of the trough that was seen in *C. lusitaniae*.

Because there is one striking trough per chromosome in both *C. lusitaniae* and *P. stipitis*, we hypothesized that these troughs might mark the locations of centromeres. The centromere is the only known genomic feature that occurs exactly once per chromosome, so there are no other obvious candidate causes of the troughs. Our approach uses a sliding window (to reduce sampling error), so each trough in *C. lusitaniae* and *P. stipitis* cannot be mapped directly to a single specific intergenic region but instead to a window of 15 consecutive genes. However, for every trough, we were able to identify one unusually large intergenic region within the window with the lowest GC3s, and we propose these as candidate centromere locations (table 2). Centromere locations have not been determined experimentally in any species of the *Candida* clade (fig. 3) except *C. albicans* and *C. dubliniensis* (Sanyal et al. 2004; Padmanabhan et al. 2008). Even though the centromeres in *C. albicans* and *C. dubliniensis* are not located in GC-poor troughs (fig. 2), we nevertheless suggest that the centromeres of *C. lusitaniae* and *P. stipitis* have been subject to a mutational process that has formed the troughs by making the DNA in the region around the centromere become GC poor (see Discussion).

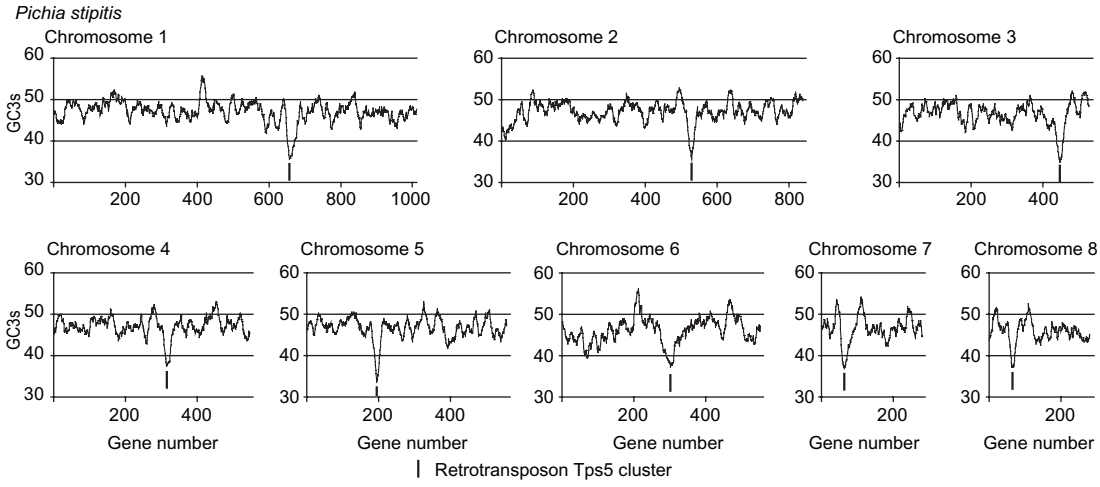

Fig. 5.—GC3s profiles of *Pichia stipitis* chromosomes. Vertical bars mark the positions of Tps5 retrotransposon clusters.

Analysis of the *Y. lipolytica* genome provides support for the hypothesis that GC3s-poor troughs can, for some species, indicate centromere locations. *Yarrowia lipolytica* is an outgroup to both the *Candida* and *Saccharomyces* clades (Dujon et al. 2004; Butler et al. 2009) and is one of the few yeast species in which centromeres have been identified experimentally (Vernis et al. 1997, 2001). We find that the GC3s profiles of *Y. lipolytica* chromosomes each contain one distinct GC-poor trough and that these coincide with the locations of the five experimentally determined centromeres on chromosomes A–E (fig. 6). Chromosome F is the only *Y. lipolytica* chromosome for which no centromere was cloned (Vernis et al. 2001). A centromere location for chromosome F was predicted bioinformatically when the genome was sequenced (Dujon et al. 2004), but this prediction does not coincide with the location of the GC-poor trough. We suggest that the location predicted by Dujon et al. (2004) is incorrect and that the centromere of chromosome F lies at the bottom of the adjacent trough (fig. 6).

Retroelements and Centromere Locations

In *P. stipitis*, Jeffries et al. (2007) noted that all the copies of the Ty5-like retrotransposon Tps5 occurred in clusters and

Table 2
Locations of GC-Poorest 15-Gene Windows in *Pichia stipitis and Candida lusitaniae* and Sizes of the Largest Intergenic Regions within Them

	Window[a]				Longest intergenic[b]		
Chromosome	**GC3s (%)**	**Start bp[c]**	**End bp[c]**	**Nadir[d]**	**From**	**To**	**Length**
P. stipitis chromosome 1	35.4	2,230,753	2,332,231	PICST_53251	PICST_28862	PICST_53466	14,594
P. stipitis chromosome 2	35.4	1,661,496	1,736,483	PICST_30000	PICST_70083	PICST_41273	38,214
P. stipitis chromosome 3	34.9	1,399,674	1,477,779	PICST_35641	PICST_30981	PICST_30986	24,208
P. stipitis chromosome 4	37.3	1,011,423	1,076,401	PICST_58121	PICST_58121	PICST_31542	26,877
P. stipitis chromosome 5	33.4	624,197	682,494	PICST_46516	PICST_32086	PICST_46124	17,264
P. stipitis chromosome 6	36.8	856,936	916,894	PICST_32891	PICST_78946	PICST_32901	30,150
P. stipitis chromosome 7	36.9	235,679	324,031	PICST_33311	PICST_14352	PICST_73528	16,278
P. stipitis chromosome 8	37.0	265,747	351,425	PICST_50504	PICST_91563	PICST_33721	36,077
C. lusitaniae chromosome 1	38.3	1,045,088	1,078,544	CLUG_00526	CLUG_00522	CLUG_00523	4,853
C. lusitaniae chromosome 2	40.5	1,791,934	1,818,051	CLUG_02107	CLUG_02104	CLUG_02105	3,283
C. lusitaniae chromosome 3	38.6	1,146,534	1,178,953	CLUG_02875	CLUG_02872	CLUG_02873	4,375
C. lusitaniae chromosome 4	39.9	121,683	157,316	CLUG_03260	CLUG_03262	CLUG_03263	4,681
C. lusitaniae chromosome 5	44.0	270,454	316,945	CLUG_04242	CLUG_04241	CLUG_04242	3,924
C. lusitaniae chromosome 6	41.3	260,220	298,959	CLUG_04968	CLUG_04966	CLUG_04967	4,919
C. lusitaniae chromosome 7	39.1	358,422	394,057	CLUG_05422	CLUG_05420	CLUG_05421	3,248
C. lusitaniae chromosome 8	38.2	147,355	176,623	CLUG_05668	CLUG_05669	CLUG_05670	5,773

[a] Fifteen-gene window with the lowest average GC3s value on the chromosome. Only genes with *C. albicans* orthologs are considered in these columns.
[b] Longest intergenic region within the 15-gene window. All annotated genes are considered in these columns.
[c] Beginning and end coordinates of the genes at the ends of the window.
[d] Gene with the lowest individual GC3s value within the window.

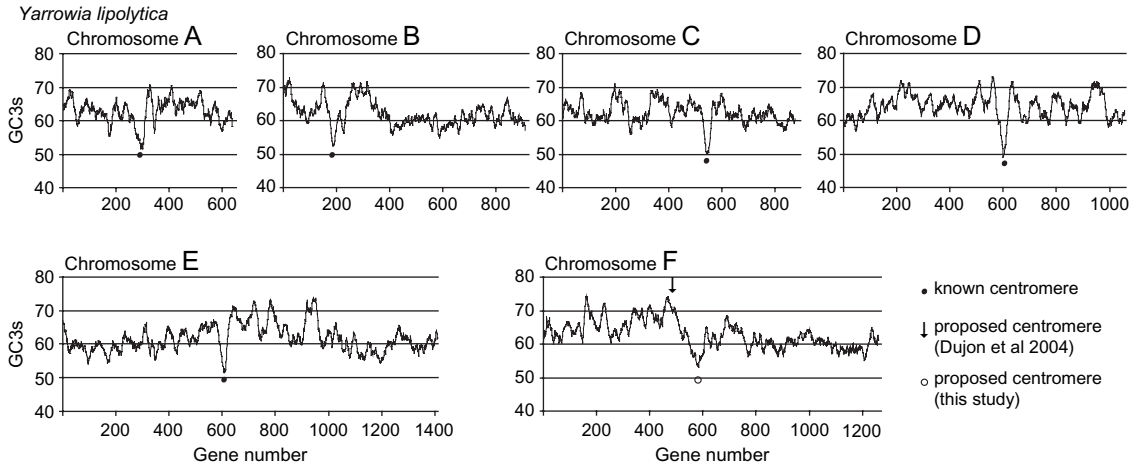

Fig. 6.—GC3s profiles of *Yarrowia lipolytica* chromosomes. Dots mark the positions of the known centromeres on chromosomes A–E (Vernis et al. 2001). On chromosome F, the arrow marks the centromere position proposed by Dujon et al. (2004), and the circle marks the location suggested by our analysis (near gene YALI0F14619g).

that there is one cluster on each chromosome. These clusters are easily visualized in dot matrix plots and typically contain a mixture of intact elements, truncated elements, and solo long terminal repeats (supplementary fig. S3A, Supplementary Material online). We find that the GC-poor troughs in *P. stipitis* coincide with these retrotransposon clusters (fig. 5). The open reading frames within the Tps5 elements are not the cause of the GC-poor troughs that we observe because our GC3s analysis ignored all retroelements. Several subfamilies of Tps5 elements exist, but they are all structurally most similar to Tdh5 of *D. hansenii* and Tca5 of *C. albicans* (Plant et al. 2000; Neuveglise et al. 2002). The Tdh5 elements of *D. hansenii* also form one cluster per chromosome (supplementary fig. S3B, Supplementary Material online), so we suggest that these are possible centromere locations in that species. Comparing these Tdh5 clusters to the GC3s profile plots for *D. hansenii* (supplementary fig. S4, Supplementary Material online) shows that although all of them are located in or near local GC-poor troughs, the troughs do not stand out and are often not the deepest on their chromosome. In the more distantly related species *C. albicans*, there are only two copies of the Tca5 element in the sequenced genome (strain SC5314), and these are not close to the known locations of any of its eight centromeres. Although we hypothesize that the Tps5/Tdh5 clusters in *P. stipitis* and *D. hansenii* are associated with centromeres, we do not suggest that the retrotransposon sequence itself has any role in centromere function. Rather, we suggest that the association is caused by preferential integration of retrotransposons into centromeric chromatin in *P. stipitis* and *D. hansenii* and that this association is a recent one, as it only occurs in these two closely related species.

Using the Candida Gene Order Browser (Fitzpatrick et al. 2010), we found that there is partial conservation of synteny, both among some of the centromere locations we proposed in *P. stipitis*, *D. hansenii*, and *C. lusitaniae* and between these putative centromeres and some of the known centromeres of *C. albicans* and *C. dubliniensis*. For example, nine genes near the known centromere of *C. albicans* chromosome 5 have orthologs in the GC-poorest window on *P. stipitis* chromosome 5, and this window also contains a Tps5 cluster (fig. 7). Four of these genes also have orthologs in the GC-poorest window on *C. lusitaniae* chromosome 3, and four other genes near *C. albicans CEN5* have orthologs located close to the Tdh5 cluster on *D. hansenii* chromosome D. Gene order conservation in these regions is not perfect, but it is known that small inversions frequently scramble the local gene order in *Candida* species (Seoighe et al. 2000). In total, we found synteny relationships involving 4 of the 8 *C. albicans/C. dubliniensis* centromeres, providing connections to GC-poor troughs or retrotransposon clusters on 4 *P. stipitis* chromosomes, 2 *D. hansenii* chromosomes, and 2 *C. lusitaniae* chromosomes (fig. 7 and supplementary fig. S5, Supplementary Material online). These observations provide support for the hypothesis that the GC-poor troughs mark centromeres in some species. However, these four examples were the only ones we could find; the other troughs in *P. stipitis* and *C. lusitaniae* are not in regions of gene order conservation between these two species or with *C. albicans/C. dubliniensis* centromeres.

We also examined the pattern of variation of intergenic G + C content (supplementary fig. S6, Supplementary Material online), as opposed to GC3s in genes. In both *C. lusitaniae* and *P. stipitis*, intergenic G + C varies within a much smaller range (approximately 36–44% G + C) than was seen in GC3s. A sliding-window approach shows that in *C. lusitaniae*, the putative centromeres (longest intergenic regions in table 2) are located within local troughs of

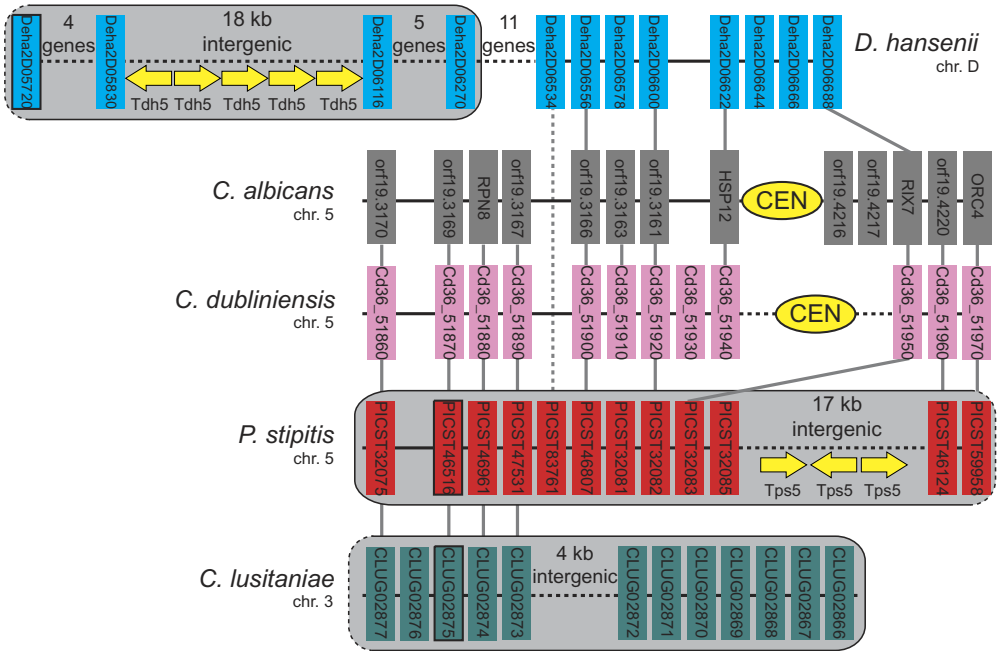

Fig. 7.—Partial synteny conservation between *CEN5* of *Candida albicans* and *C. dubliniensis* and proposed centromeres in *Pichia stipitis*, *C. lusitaniae*, and *Debaryomyces hansenii*. Vertical lines indicate orthologous genes. Gray ovals show the GC-poorest 15-gene window in *P. stipitis*, *D. hansenii*, and *C. lusitaniaei*, and the name of the gene with the lowest individual GC3s value in this window is surrounded by a bold box (some genes at the ends of windows are not shown).

G + C–poor intergenic regions, but these troughs are much less dramatic than those seen for GC3s. The putative centromeres of *C. lusitaniae* are also located in some of the most G + C–poor intergenic spacers in its genome (supplementary fig. S6E, Supplementary Material online). *Pichia stipitis* does not show the same trend, but the G + C content of its putatively centromeric intergenic regions is affected by the presence of the Tps5 arrays.

Discussion

Our results for the *Saccharomyces* species and for *C. albicans* versus *C. dubliniensis* reveal two main attributes of G + C content evolution among closely related yeast species. First, the pattern of variation along the chromosome is conserved. The locations of peaks and troughs among these species tend to coincide, even though large numbers of nucleotide substitutions have occurred at synonymous codon sites. This observation could be consistent with a model in which recombination determines the local G + C content via biased gene conversion (Birdsell 2002; Duret and Galtier 2009) but only if the locations of recombination hot spots are conserved on orthologous chromosomes among the species. Recent analysis of one chromosome shows that recombination hot spots are conserved between *S. cerevisiae* and *S. paradoxus* (Tsai et al. 2010), which suggests that they may be conserved on all chromosomes among all *Saccharomyces* species. The conservation of recombination profiles is surprising (Tsai et al. 2010), as recombination hot spot locations are not even conserved between human and chimpanzee, two mammals that diverged much more recently than the *Saccharomyces* species (Ptak et al. 2005; Winckler et al. 2005) and which probably have identical GC3s profiles because they are colinear and only 1% divergent in sequence (Chimpanzee Sequencing and Analysis Consortium 2005; Elhaik et al. 2009).

Second, even within this framework of conserved peak and trough locations, the actual GC3s values of orthologous genomic regions can shift significantly and systematically between species (figs. 1 and 2). The shifts cause an orchestrated change in GC3s values, so that the GC richness or poorness of each gene relative to other genes is largely unchanged. We think that the most probable cause for these shifts is factors that can change the overall recombination rate (per year) in a species, affecting all hot spots and cold spots uniformly. Such factors could include, but are not limited to, changes in the sequences of proteins involved in recombination (directly altering the recombination rate) or the mismatch repair system (affecting the gene conversion bias parameter *b*; Duret and Galtier 2009) and changes in the effective population size, which will affect the rate of fixation of nucleotide substitutions. The frequency of meiotic recombination in particular is likely to be highly variable among species in the *Candida* clade because some species (*C. lusitaniae*, *P. stipitis*, and *D. hansenii*) have complete

sexual cycles and others (*C. albicans*) have no known mechanism of meiosis (Tzung et al. 2001; Forche et al. 2008; Butler et al. 2009; Reedy et al. 2009). However, Spo11-dependent recombination in a parasexual cycle has been observed in *C. albicans* (Forche et al. 2008), and even if it has no true meiosis, it could also experience gBGC during mitotic recombination.

We found one deep GC-poor trough per chromosome in three species (*C. lusitaniae*, *P. stipitis*, and *Y. lipolytica*) and propose that they mark the locations of centromeres. This proposal is based on four arguments: 1) colocalization with five experimentally determined centromeres in *Y. lipolytica*; 2) partial synteny conservation with known centromeres of *C. albicans*/*C. dubliniensis*; 3) colocalization of the GC-poor troughs with clusters of retroelements in *P. stipitis*, considering that retroelements cluster near centromeres in many eukaryotes (e.g., Paterson et al. 2009); and 4) the simple fact that no genetic element other than the centromere is known to occur in precisely one copy per chromosome. We also identified putative centromere locations in *D. hansenii* based on the presence of one Tdh5 retroelement cluster per chromosome. In other yeast species, transposable elements are known to have preferences for integrating into particular regions of the genome (Voytas and Boeke 2002). To our knowledge, transposable elements with a preference for integration at centromeres have not been described previously in ascomycete yeasts, although they are known in some filamentous ascomycetes (Cambareri et al. 1998; Galagan and Selker 2004).

What mechanism could form GC-poor troughs around centromeres? Meiotic recombination is known to be suppressed near centromeres across a wide variety of eukaryotes that have been examined (Choo 1998), including *S. cerevisiae* (Lambie and Roeder 1986; Mancera et al. 2008). The molecular mechanism by which this effect occurs remains unknown. The gBGC model proposes that local G + C content is determined by the local recombination rate, so it predicts that G + C content should be low near centromeres.

There are two very different types of centromere structures among the yeasts studied here (Ishii 2009): *S. cerevisiae* and other species in the *Saccharomyces/Kluyveromyces* clade have "point" centromeres (Hieter et al. 1985; Heus et al. 1993; Kitada et al. 1997; Pribylova et al. 2007), whereas *C. albicans*, *C. dubliniensis*, and *Y. lipolytica* have "regional" centromeres (Vernis et al. 2001; Sanyal et al. 2004; Padmanabhan et al. 2008). Point centromeres are genetically determined (their DNA sequence is necessary and sufficient for their function), whereas regional centromeres are epigenetically determined (Ketel et al. 2009; Malik and Henikoff 2009). Point centromeres are small (<200 bp) and have a strongly conserved consensus sequence near which a single Cse4-containing nucleosome binds, whereas regional centromeres lack a consensus and bind several Cse4 nucleosomes over a region of up to 5 kb (Padmanabhan et al. 2008; Malik and Henikoff 2009).

The species in which we observe deep GC-poor troughs are all expected (from their phylogenetic position) to have regional centromeres, but the troughs are much larger than the expected size of the centromeres themselves. For example, the two peaks flanking the trough on *C. lusitaniae* chromosome 3 (fig. 4) are 132 kb apart, and there are 75 annotated genes between them. Even the region in which GC3s is <50% is 36–48 kb long on each *C. lusitaniae* chromosome (supplementary table S2, Supplementary Material online). Similarly, the peaks flanking the trough on *P. stipitis* chromosome 7 (fig. 5) are 244 kb apart, and on *Y. lipolytica* chromosome D (fig. 6), the distance is 301 kb. In *S. cerevisiae*, the zone of recombination suppression around each centromere is only about 10 kb (average value from supplementary table 1, Supplementary Material online, of Mancera et al. 2008). Our results can therefore be interpreted from either of two viewpoints. One is that the troughs are caused by reduction of gBGC at centromeres, in which case recombination around the centromere must be suppressed over a much larger area in *C. lusitaniae*, *P. stipitis*, and *Y. lipolytica* than in *S. cerevisiae*. Under this interpretation, figure 4 also suggests that there is a gradient of recombination rates along each chromosome arm in *C. lusitaniae*. The alternative view is that GC-poor troughs in *C. lusitaniae*, *P. stipitis*, and *Y. lipolytica* are simply too large to have been caused by recombination suppression and gBGC and so must have a different, unidentified, cause.

Why do all yeast species not show GC-poor troughs around centromeres? In fact, reexamination of chromosomal GC3s for other sequenced Saccharomycotina species with point centromeres (which can be identified bioinformatically; Dujon et al. 2004; Souciet et al. 2009) shows that virtually all the point centromeres lie in locally GC-poor regions. This property was not noticed before because for many species with point centromeres, the trough is not the lowest one on the chromosome. The data for *S. cerevisiae* are shown in figure 1 and for other point centromere species (*C. glabrata*, *Z. rouxii*, *E. gossypii*, *K. lactis*, *Lachancea thermotolerans*, and *L. kluyveri*) in supplementary fig. S2 (Supplementary Material online). Therefore, *C. albicans* and its close relative *C. dubliniensis* stand out as unusual, among the species whose centromeres are mapped, because their centromeres are not in local GC-poor troughs (fig. 2), but these are also the two most GC-poor species in our whole study.

It is also interesting to note that there may be a correlation between the presence of deep centromeric GC3s troughs and the absence of one of the two possible pathways for meiotic recombination, the *MSH4*/*MSH5* pathway (Richard et al. 2005; Butler et al. 2009; Reedy et al. 2009). Msh4/Msh5-dependent recombination in *S. cerevisiae* is subject to crossover interference, and deleting the genes reduces

crossing over by 2- to 3-fold (Ross-Macdonald and Roeder 1994; Hollingsworth et al. 1995). If we use a 10% G + C content difference between peak and trough as a rule of thumb to define "deep" troughs, then we find that all the species with deep troughs at their centromeres lack *MSH4* and *MSH5*, both among the regional centromere species (*Y. lipolytica*, *C. lusitaniae*, and *P. stipitis*) and among the point centromere species (*E. gossypii* and *L. thermotolerans*), whereas most species in which centromeric troughs are weaker (such as *K. lactis*) or absent (such as *C. albicans*) retain these two genes. The only exceptions to this correlation among the species we studied are *C. guilliermondii* and *D. hansenii*, which lack *MSH4/MSH5* but do not show deep troughs at centromeres (centromeres are not mapped experimentally in either of these species, but neither of them shows deep troughs anywhere in its genome; supplementary fig. S2K, Supplementary Material online).

As the link between GC3s troughs and centromeres has only been validated experimentally in *Y. lipolytica*, it would be of interest to test the centromere locations we have proposed for other species. We are currently carrying out validation experiments for the centromeres of *C. lusitaniae*. It would also be of interest to investigate whether a depression of GC content around the centromere occurs in other taxonomic groups such as animals and to examine the properties of holocentric organisms.

Supplementary Material

Supplementary tables S1 and S2 and figures S1–S6 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

This study was supported by Science Foundation Ireland and Irish Research Council for Science, Engineering and Technology. We thank David Fitzpatrick for constructing the phylogenetic tree in figure 3.

Literature Cited

Bernardi G, et al. 1985. The mosaic genome of warm-blooded vertebrates. Science 228:953–958.
Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol. 19:1181–1197.
Bradnam KR, Seoighe C, Sharp PM, Wolfe KH. 1999. G+C content variation along and among *Saccharomyces cerevisiae* chromosomes. Mol Biol Evol. 16:666–675.
Butler G, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature 459:657–662.
Cambareri EB, Aisner R, Carbon J. 1998. Structure of the chromosome VII centromere region in *Neurospora crassa*: degenerate transposons and simple repeats. Mol Cell Biol. 18:5465–5477.
Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.
Choo KHA. 1998. Why is the centromere so cold? Genome Res. 8:81–82.
Cliften PF, et al. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. Genome Res. 11:1175–1186.
Dietrich FS, et al. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science 304:304–307.
Dujon B. 1996. The yeast genome project: what did we learn? Trends Genet. 12:263–270.
Dujon B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet. 22:375–387.
Dujon B, et al. 2004. Genome evolution in yeasts. Nature 430:35–44.
Duret L, Eyre-Walker A, Galtier N. 2006. A new perspective on isochore evolution. Gene 385:71–74.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 10:285–311.
Elhaik E, Landan G, Graur D. 2009. Can GC content at third-codon positions be used as a proxy for isochore composition? Mol Biol Evol. 26:1829–1833.
Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. Nat Rev Genet. 2:549–555.
Fischer G, James SA, Roberts IN, Oliver SG, Louis EJ. 2000. Chromosomal evolution in *Saccharomyces*. Nature 405:451–454.
Fitzpatrick DA, O'Gaora P, Byrne KP, Butler G. 2010. Analysis of gene evolution and metabolic pathways using the *Candida* Gene Order Browser. BMC Genomics. 11:290.
Forche A, et al. 2008. The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains. PLoS Biol. 6:e110.
Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends Genet. 20:417–423.
Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911.
Gerton JL, et al. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A. 97:11383–11390.
Goffeau A, et al. 1996. Life with 6000 genes. Science 274:546, 563–567.
Heus JJ, Zonneveld BJ, Steensma HY, van den Berg JA. 1993. The consensus sequence of *Kluyveromyces lactis* centromeres shows homology to functional centromeric DNA from *Saccharomyces cerevisiae*. Mol Gen Genet. 236:355–362.
Hieter P, et al. 1985. Functional selection and analysis of yeast centromeric DNA. Cell 42:913–921.
Hollingsworth NM, Ponte L, Halsey C. 1995. *MSH5*, a novel MutS homolog, facilitates meiotic reciprocal recombination between homologs in *Saccharomyces cerevisiae* but not mismatch repair. Genes Dev. 9:1728–1739.
Huberman JA, Pridmore RD, Jäger D, Zonneveld B, Philippsen P. 1986. Centromeric DNA from *Saccharomyces uvarum* is functional in *Saccharomyces cerevisiae*. Chromosoma 94:162–168.
Ishii K. 2009. Conservation and divergence of centromere specification in yeast. Curr Opin Microbiol. 12:616–622.
Jackson AP, et al. 2009. Comparative genomics of the fungal pathogens *Candida dubliniensis* and *C. albicans*. Genome Res. 19:2231–2244.

Jeffries TW, et al. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. Nat Biotechnol. 25:319–326.
Kaback DB, Guacci V, Barber D, Mahon JW. 1992. Chromosome size-dependent control of meiotic recombination. Science 256:228–232.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254.
Ketel C, et al. 2009. Neocentromeres form efficiently at multiple possible loci in Candida albicans. PLoS Genet. 5:e1000400.
Kitada K, Yamaguchi E, Hamada K, Arisawa M. 1997. Structural analysis of a *Candida glabrata* centromere and its functional homology to the *Saccharomyces cerevisiae* centromere. Curr Genet. 31:122–127.
Kurtzman CP. 2003. Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. FEMS Yeast Res. 4:233–245.
Kurtzman CP, Robnett CJ. 2003. Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. FEMS Yeast Res. 3:417–432.
Lambie EJ, Roeder GS. 1986. Repression of meiotic crossing over by a centromere (*CEN3*) in *Saccharomyces cerevisiae*. Genetics 114:769–789.
Malik HS, Henikoff S. 2009. Major evolutionary transitions in centromere complexity. Cell 138:1067–1082.
Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454:479–485.
Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19:330–338.
Marsolier-Kergoat MC, Yeramian E. 2009. GC content and recombination: reassessing the causal effects for the *Saccharomyces cerevisiae* genome. Genetics 183:31–38.
Mouchiroud D, Gautier C, Bernardi G. 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. J Mol Evol. 27:311–320.
Neuveglise C, Feldmann H, Bon E, Gaillardin C, Casaregola S. 2002. Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. Genome Res. 12:930–943.
Padmanabhan S, Thakur J, Siddharthan R, Sanyal K. 2008. Rapid evolution of Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, *Candida albicans* and *Candida dubliniensis*. Proc Natl Acad Sci U S A. 105:19797–19802.
Paterson AH, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556.
Payen C, et al. 2009. Unusual composition of a yeast chromosome arm is associated with its delayed replication. Genome Res. 19:1710–1721.
Petes TD, Merker JD. 2002. Context dependence of meiotic recombination hotspots in yeast: the relationship between recombination activity of a reporter construct and base composition. Genetics 162:2049–2052.
Plant EP, Goodwin TJ, Poulter RT. 2000. Tca5, a Ty5-like retrotransposon from *Candida albicans*. Yeast 16:1509–1518.
Pribylova L, Straub M-L, Sychrova H, de Montigny J. 2007. Characterisation of *Zygosaccharomyces rouxii* centromeres and construction of first *Z. rouxii* centromeric vectors. Chromosome Res. 15:439–445.
Ptak SE, et al. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet. 37:429–434.
Reedy JL, Floyd AM, Heitman J. 2009. Mechanistic plasticity of sexual reproduction and meiosis in the Candida pathogenic species complex. Curr Biol. 19:891–899.
Richard GF, Kerrest A, Lafontaine I, Dujon B. 2005. Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination. Mol Biol Evol. 22: 1011–1023.
Ross-Macdonald P, Roeder GS. 1994. Mutation of a meiosis-specific MutS homolog decreases crossing over but not mismatch correction. Cell 79:1069–1080.
Sanyal K, Baum M, Carbon J. 2004. Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. Proc Natl Acad Sci U S A. 101:11374–11379.
Seoighe C, et al. 2000. Prevalence of small inversions in yeast gene order evolution. Proc Natl Acad Sci U S A. 97:14433–14437.
Sharp PM, Lloyd AT. 1993. Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. Nucleic Acids Res. 21:179–183.
Souciet JL, et al. 2009. Comparative genomics of protoploid Saccharomycetaceae. Genome Res. 19:1696–1709.
Tsai IJ, Burt A, Koufopanou V. 2010. Conservation of recombination hotspots in yeast. Proc Natl Acad Sci U S A. 107:7847–7852.
Tzung KW, et al. 2001. Genomic evidence for a complete sexual cycle in *Candida albicans*. Proc Natl Acad Sci U S A. 98:3249–3253.
van het Hoog M, et al. 2007. Assembly of the *Candida albicans* genome into sixteen supercontigs aligned on the eight chromosomes. Genome Biol. 8:R52.
Vernis L, et al. 1997. An origin of replication and a centromere are both needed to establish a replicative plasmid in the yeast *Yarrowia lipolytica*. Mol Cell Biol. 17:1995–2004.
Vernis L, et al. 2001. Only centromeres can supply the partition system required for *ARS* function in the yeast *Yarrowia lipolytica*. J Mol Biol. 305:203–217.
Voytas DF, Boeke JD. 2002. Ty1 and Ty5 of *Saccharomyces cerevisiae*. In: NL Craig, R Craigie, M Gellert, and AM Lambowitz, editors. Mobile DNA II. Washington DC: ASM Press. p. 631–662.
Wang HC, Singer GA, Hickey DA. 2004. Mutational bias affects protein evolution in flowering plants. Mol Biol Evol. 21:90–96.
Winckler W, et al. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308:107–111.
Yamane S, H Karashima, H Matsuzaki, T Hatano, S Fukui. 1999. Isolation of centromeric DNA from *Saccharomyces bayanus*. J Gen Appl Microbiol. 45:89–92.

Associate editor: Dmitri Petrov