Repetitive sequence variation and dynamics in the ribosomal DNA array of Saccharomyces cerevisiae as revealed by whole-genome resequencing

Citation: James, S. A. et al. "Repetitive Sequence Variation and Dynamics in the Ribosomal DNA Array of Saccharomyces Cerevisiae as Revealed by Whole-genome Resequencing." Genome Research 19.4 (2009): 626–635. © 2009 by Cold Spring Harbor Laboratory Press

As Published: http://dx.doi.org/10.1101/gr.084517.108

Publisher: Cold Spring Harbor Laboratory Press

Persistent URL: http://hdl.handle.net/1721.1/76761

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

CSH PRESS GENOME RESEARCH

Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing

Stephen A. James, Michael J.T. O'Kelly, David M. Carter, et al.

Genome Res. 2009 19: 626-635 originally published online January 13, 2009
Access the most recent version at doi:10.1101/gr.084517.108

Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing

Stephen A. James,[1,4] Michael J.T. O'Kelly,[2,4,5] David M. Carter,[3,6] Robert P. Davey,[1]
Alexander van Oudenaarden,[2] and Ian N. Roberts[1,7]

[1]*National Collection of Yeast Cultures, Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, United Kingdom;* [2]*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [3]*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, United Kingdom*

Ribosomal DNA (rDNA) plays a key role in ribosome biogenesis, encoding genes for the structural RNA components of this important cellular organelle. These genes are vital for efficient functioning of the cellular protein synthesis machinery and as such are highly conserved and normally present in high copy numbers. In the baker's yeast *Saccharomyces cerevisiae*, there are more than 100 rDNA repeats located at a single locus on chromosome XII. Stability and sequence homogeneity of the rDNA array is essential for function, and this is achieved primarily by the mechanism of gene conversion. Detecting variation within these arrays is extremely problematic due to their large size and repetitive structure. In an attempt to address this, we have analyzed over 35 Mbp of rDNA sequence obtained from whole-genome shotgun sequencing (WGSS) of 34 strains of *S. cerevisiae*. Contrary to expectation, we find significant rDNA sequence variation exists within individual genomes. Many of the detected polymorphisms are not fully resolved. For this type of sequence variation, we introduce the term partial single nucleotide polymorphism, or pSNP. Comparative analysis of the complete data set reveals that different *S. cerevisiae* genomes possess different patterns of rDNA polymorphism, with much of the variation located within the rapidly evolving nontranscribed intergenic spacer (IGS) region. Furthermore, we find that strains known to have either structured or mosaic/hybrid genomes can be distinguished from one another based on rDNA pSNP number, indicating that pSNP dynamics may provide a reliable new measure of genome origin and stability.

[Supplemental material is available online at www.genome.org.]

In the baker's yeast *Saccharomyces cerevisiae*, the rDNA array, which typically comprises of between 150 and 200 tandem repeats depending on the strain, occupies ~60% of chromosome XII (Petes 1979; Kobayashi et al. 1998; Kobayashi 2006). In the case of *S. cerevisiae*, each repeat comprises of four ribosomal RNA (rRNA) genes, the large subunit (LSU) (also referred to as the 26S rRNA), small subunit (SSU) (also referred to as the 18S rRNA), 5.8S and 5S rRNA genes, as well as two internal transcribed spacers (ITS1 and ITS2), two external transcribed spacers (ETS1 and ETS2), and a large intergenic spacer (IGS) (Fig. 1). Due to its large size (~1.5 Mbp), and coupled with its highly repetitive structure, accurate and reliable sequence assembly of the entire rDNA array is impossible by current sequencing methods. Indeed, when the first genome sequence of a eukaryotic organism, namely, *S. cerevisiae* strain S288c, was determined, only the sequences for the terminal left- and right-hand rDNA repeats were published (Goffeau et al. 1996). Both repeats were found to have identical (9.1-kb) sequences and the rest of the array repeats were assumed, by default, to be identical as a result of rapid homogenization predicted by gene conversion (Gangloff et al. 1996). High levels of sequence identity between individual (rDNA) repeats have also been reported for many other eukaryotes (for review, see Eickbush and Eickbush 2007); however, such assumptions overlook the possibility that cryptic variation may exist in large tandem repeat arrays. This type of variation is difficult to detect and even when found can easily be overlooked or dismissed as sequencing errors (O'Donnell and Cigelnik 1997). Evidence to suggest that some degree of sequence variation can exist within yeast rDNA arrays comes from two recent studies. In *Clavispora lusitaniae*, strains have been found that display intragenomic sequence heterogeneity in the D1/D2 variable domains of the large-subunit rDNA (Lachance et al. 2003). Likewise, in a phylogenetic analysis of *Saccharomyces* sensu stricto strains, Montrocher et al. (1998) found that repeated attempts to sequence the PCR-amplified ITS region from the *S. cerevisiae* type strain (CBS 1171^T) failed, indicating that this strain appears to possess more than one type of ITS sequence.

With the advent of whole-genome shotgun sequencing (WGSS), the opportunity has arisen, not only to be able to search for variation within large repeat arrays such as the rDNA but also to quantify such variation should it exist. The advantage of the WGSS approach is that, in theory, all regions of the target genome receive equal coverage. In essence, each repeat is equally likely to contribute to the final data set. This means that WGSS data is well-suited for determining the absolute level of sequence variation within a specific repeat array. This approach was used recently in two separate studies examining rDNA sequence variation, one focusing on *Drosophila* (Stage and Eickbush 2007) and the other on fungi (Ganley and Kobayashi 2007). In the

[4]These authors contributed equally to this work.
Present addresses: [5]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [6]Autonomy Corporation, Cambridge Business Park, Cambridge CB4 0WZ, UK.
[7]Corresponding author.
E-mail ian.roberts@bbsrc.ac.uk; fax 44-(0)1603-458414.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.084517.108.

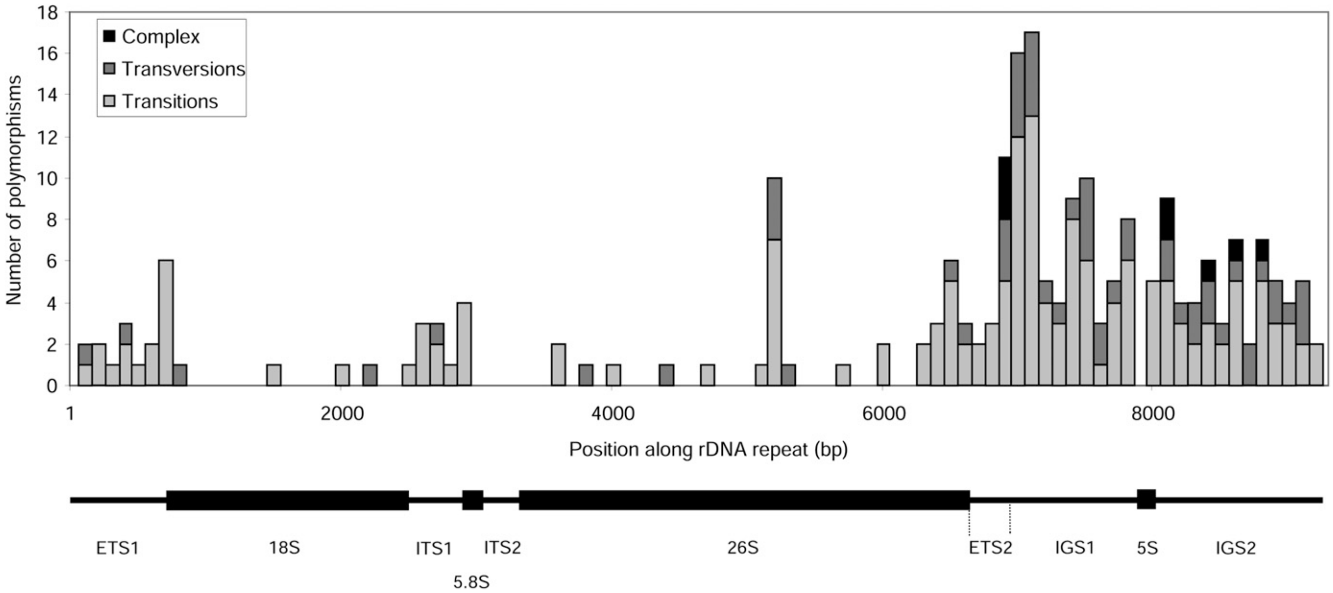

Figure 1. Distribution of (base substitution-only) polymorphisms, occurring in 100-bp bins, in the *S. cerevisiae* rDNA array derived from the combined sequence data of 34 different strains of *S. cerevisiae* (Table 1; Supplemental Table 1).

Drosophila study, Stage and Eickbush (2007) analyzed sequencing reads obtained from the WGSS projects of 12 *Drosophila* species, and identified variation in more than 3% of the rDNA repeats for 11 of these species, with the majority of detected variation located in the noncoding ETS and IGS regions of the rDNA repeat. In the fungal study, which focused on five different fungi, including three species of yeast (viz. *Cryptococcus neoformans*, *S. cerevisiae*, and *Saccharomyces paradoxus*), Ganley and Kobayashi (2007) found that very little variation appears to exist within the rDNA arrays of these eukaryotic microorganisms. Indeed, in the case of *S. cerevisiae*, only four polymorphisms, all regarded to be of high confidence by the investigators, were detected, evenly distributed across the entire rDNA repeat, of which only two were found to be present on a relatively high number of sequence reads. To investigate this in far greater depth, we have analyzed over 35 Mbp of rDNA sequence data (a total of 44,443 individual sequencing reads) for 34 different strains of *S. cerevisiae*, obtained from the Wellcome Trust Sanger Institute as part of the *Saccharomyces* Genome Resequencing Project (SGRP) (www.sanger.ac.uk/Teams/Team118/sgrp), a population genomics study of *S. cerevisiae* and its close wild relative *S. paradoxus* (Liti et al. 2009). The selected *S. cerevisiae* strains, whose genomes were sequenced to varying depths of coverage, were isolated from a wide variety of sources, including baking, brewing, laboratory, pathogenic, pro-biotic, and environmental, and from numerous locations around the world (Supplemental Table 1; Liti et al. 2009),

In this study, we identified all the rDNA-specific shotgun reads from the complete SGRP WGSS data set for each individual strain, and aligned these against the rDNA consensus sequence of the *S. cerevisiae* reference strain (S288c) to look for base substitution-only type polymorphisms. Stringent quality score filtering was employed to minimize the expected number of false positives, due to either the introduction of non-rDNA sequences or sequencing error (for details, see Methods). The results provide an unprecedented in-depth insight into both the differing levels of variation found to exist within different *S. cerevisiae* rDNA arrays and its distribution. Furthermore, we show how partial single nucleotide polymorphism (pSNP) variation and dynamics can be used as a means to differentiate between strains with structured “clean” genomes and strains with genomes of hybrid or “mosaic” origin (Liti et al.2009).

Results

As previously commented on by Ganley and Kobayashi (2007), the high similarity between individual repeats within the rDNA means it is impossible to establish from which repeat in the array any particular sequence read originates. Instead, the sequence data reduce to a single repeat consensus alignment for each strain with coverage determined by the rDNA copy number, which in *S. cerevisiae* typically ranges from 150–200 copies per haploid genome (Petes 1979; Kobayashi et al. 1998), and the level of genome sequencing coverage. In the case of the 34 strains included in this study, genome coverage ranges from a depth of 0.65 (NCYC 361) to 3.83 for (Y55) (see Table 4, below; Liti et al. 2009). The number of rDNA-specific reads divided by the total number of reads (per strain) was used to determine how much of each genome is made up of rDNA. Using the SGRP estimates of the genome size for each strain, coupled with the known length of the *S. cerevisiae* rDNA repeat (9.1 kb) (Goffeau et al. 1996), an estimate of the rDNA copy number for each strain was calculated. Taken together, rDNA coverage was calculated for each individual strain, with the average coverage determined to be 140.36 reads per nucleotide (see Supplemental Table 1). As discussed by Stage and Eickbush (2007), differentiating between authentic sequence variation and sequencing errors is often complicated by recurring errors (i.e., the same error detected in multiple traces of the same sequence) and is especially compounded in large repeat arrays, where high copy number can give rise to hundreds of reads covering each individual repeat region. In an effort to reduce the likelihood of introducing false positives into the data set due to sequencing errors, we have used a stringent filtering strategy based on *phred* quality (q) scores (for details, see Methods) to identify base substitution-only type polymorphisms. We have mapped their positions in relation to the S288c rDNA repeat consensus sequence (Goffeau et al. 1996) for each individual *S. cerevisiae* rDNA array. Polymorphisms found on single reads (referred to as single read polymorphisms) unique to individual strains were also excluded, as we could not reliably discount the possibility that these had arisen due to sequencing error. In the case of UWOPS83-787-3, this led to a substantial reduction in the total number of polymorphisms identified (from 158 to 76). Many of the UWOPS83-787-3 single read polymorphisms were found to occur in small clusters dispersed across IGS2, indicating further research is required. A list of all the single read polymorphisms excluded from our final analyses (a total of 88 "putative" polymorphisms) is presented in Supplemental Table 2.

Sequence variation within individual rDNA arrays

Accurate and reliable inter- and intra-specific alignment of highly variable regions of DNA, such as the rDNA IGS region that includes a number of long homopolymeric polyA and polyT tracts (Skryabin et al. 1984; Jemtland et al. 1986; Goffeau et al. 1996), can prove problematic. The presence of indels (insertions and deletions), which can often extend over multiple positions, particularly in long homopolymeric tracts, are difficult to count in a consistent and unbiased way. For this reason, we restricted our current study to the detection and comparative analysis of base substitution-only type polymorphisms (i.e., transitions and transversions).

Contrary to expectation and despite the exclusion of indels from our analyses, we nevertheless found the level of variation within individual *S. cerevisiae* rDNA arrays to be remarkably high, differing quite markedly between strains (ranging from 10 polymorphisms in DBVPG 1788 to 76 polymorphisms in UWOPS83_787.3) (Table 1). Figure 1 is based on the combined sequence data for all 34 strains and shows the distribution of base substitution-only polymorphisms, occurring in 100-bp bins, over the entire *S. cerevisiae* rDNA repeat. In total, 227 polymorphic sites were identified: 44 sites in the rRNA-encoding genes (five in the 18S, 35 in the 26S, and three in the 5S rRNA genes), 27 sites in the ETS region, 11 sites in the ITS region, and 146 sites in the non-transcribed IGS region.

With regard to 18S and 26S rRNA genes, we also examined polymorphism distribution in relation to rRNA secondary structure. To do this, we used the SSU and LSU rRNA secondary structure models and the derived variability maps from the European Ribosomal RNA database (Van de Peer et al. 1997; Ben Ali et al. 1999). In each variability map, every single nucleotide residue is categorized according to its level of conservation, which in turn is dependent upon whether it is located within a "core region" or an "expansion region" (domains V1 to V9 in 18S rRNA; domains D1 to D12 in 26S rRNA; Hassouna et al. 1984; De Rijk et al. 1992). Core region sequences evolve much slower than expansion region sequences as they contain the essential sites necessary for function (i.e., active sites, substrate binding sites, and subunit contact points). In our study, many of the detected polymorphisms (29 of 40 sites) are found either in the more rapidly evolving expansion regions or in single-stranded and loop regions located elsewhere in the two rRNA molecules. Between the two genes, the 26S rRNA D7 domain was found to contain by far the most polymorphisms (11 in total). In contrast, far less variation was detected in the core regions, which is unsurprising, as these regions are under far greater functional constraint than the expansion regions. A complete list of all the 18S and 26S rRNA polymorphisms and their secondary structure locations is presented in Supplemental Table 3.

Table 1. *S. cerevisiae* rDNA array polymorphism totals

Strain[a]	Type[b]	Polymorphism total[c]
273614N	Clinical	43
322134S	Clinical	33
378604X	Clinical	40
BC187	Fermentation	17
DBVPG 1106	Fermentation	16
DBVPG 1373	Wild	16
DBVPG 1788	**Wild**	**10**
DBVPG 1853	Fermentation	40
DBVPG 6040	Spoilage	44
DBVPG 6044	Fermentation	29
DBVPG 6765	Pro-biotic	27
K11	Fermentation	31
L_1374	Fermentation	18
NCYC 110	Fermentation	23
NCYC 361	Spoilage	31
S288c	Laboratory	23
SK1	Laboratory	35
UWOPS03-461-4	Wild	33
UWOPS05-217-3	Wild	36
UWOPS05-227-2	Wild	33
UWOPS83-787-3	**Wild**	**76**
UWOPS87-2421	Wild	21
W303	Laboratory	23
Y12	Fermentation	27
Y55	Laboratory	40
Y9	Fermentation	29
YIIc17_E5	Fermentation	39
YJM975	Clinical	12
YJM978	Clinical	17
YJM981	Clinical	12
YPS128	Wild	16
YPS606	Wild	21
YS4	Baking	41
YS9	Baking	38

[a] For additional strain details see Liti et al. (2009) and Supplemental Data. Strain W303 was created by multiple crossing (Rothstein 1977; Rothstein et al. 1977).

[b] Strains are subdivided into seven categories according to their source of origin (i.e., baking [2], clinical [6], fermentation [10], laboratory [4], pro-biotic [1], spoilage [2], and wild type [9]).

[c] The rDNA arrays with the lowest and highest number of (base substitution-only) polymorphisms are in bold.

Collectively, our findings contrast with those of Ganley and Kobayashi (2007), who identified only four polymorphisms (or seven, if low-confidence polymorphisms are also considered) in the rDNA array of the single *S. cerevisiae* strain included in their study. These four polymorphisms, all regarded to be of high confidence by the investigators, were found to be evenly distributed across the rDNA repeat, with two located in the large-subunit rRNA gene, one in IGS1, and one in ITS2. None of these polymorphisms were detected in the present study, and it is likely therefore that they are specific to RM11-1a, the *S. cerevisiae* strain sequenced by the Broad Institute as part of the Fungal Genome Initiative (http://www.broad.mit.edu/annotation/fgi/) and included as the sole *S. cerevisiae* representative in the Ganley and Kobayashi (2007) study. At first glance, RM11-1a appears to have a comparable number of polymorphisms to that found in DBVPG 1788 (10), which had the fewest number of polymorphisms in the *S. cerevisiae* strain set analyzed in this study (Table 1). However, it is important to note that one key difference between the two (fungal) studies is that our analyses were restricted to the detection of base substitution-only polymorphisms, whereas Ganley and Kobayashi (2007) measured both base substitutions and indels. Thus, it is quite probable that the amount of variation detected in DBVPG 1788, and indeed in all the *S. cerevisiae* rDNA arrays analyzed in this study, is an underestimate and could, with the inclusion of indels, be significantly higher. Evidence to support this comes from the observation that the rDNA consensus sequences assembled from the SGRP individual strain data sets vary in length between strains, ranging from 9083 bp (UWOPS03-461-4, UWOPS05-217-3, and UWOPS05-227-2) to 9147 bp (273614N) (see Supplemental Table 1). Furthermore, a preliminary visual examination of a subset of 100 DBVPG 1853-specific reads covering the 3′ end of IGS1 has revealed the existence of at least one homopolymeric tract that varies in length between individual repeats within the same rDNA array (see Supplemental Fig. 1). Currently we are developing an accurate alignment method for detecting and quantifying indel variation both between and within individual yeast rDNA arrays.

To examine the rDNA sequence variation further, we also subdivided the polymorphisms into three categories, namely, transitions, transversions, and complex mutations. The latter category was included to accommodate any nucleotide positions at which different *S. cerevisiae* strains exhibit different types of mutation (i.e., a transition in one strain as opposed to a transversion in another). For example, in IGS2 at nucleotide position 8383 (G in the reference strain S288c), YIIc17_E5 has a G→T transversion, whereas strains DBVPG 6040, UWOPS03_461_4, UWOPS05_217_3, UWOPS05_227_2, and UW83_787_3 all have a G→A transition. In total, eight such sites were identified, three in IGS1 (nucleotide positions 6862, 6878, and 6895) and five in IGS2 (nucleotide positions 8021, 8084, 8383, 8581, and 8709) (see Supplemental Table 4). Like Ganley and Kobayashi (2007), we found transitions to be the most abundant form of mutation, representing ~71% (162 of 227 sites) of all base substitutions found in the *S. cerevisiae* rDNA array (Table 2).

Distribution of rDNA polymorphisms

In addition to the fact that many of the strains analyzed in this study have far more variation in their rDNA arrays than was found in *S. cerevisiae* RM11-1a (Ganley and Kobayashi 2007), we also found that the sequence variation is not distributed evenly over the rDNA repeat (Fig. 1). Indeed as discussed in the previous section, the majority of polymorphisms are found in the IGS region: 83 sites in IGS1 and 63 sites in IGS2 (Table 2). To explore this in greater detail, we refined our analyses to focus specifically on the location and distribution of polymorphisms found within this region.

Table 2. **Region-by-region breakdown of rDNA polymorphisms categorized according to mutation type**

	Polymorphisms			
Region	Transitions	Transversions	Complex[a]	Total
ETS1	15	2	0	17
18S	3	2	0	5
ITS1	10	1	0	11
5.8S	0	0	0	0
ITS2	0	0	0	0
26S	27	8	0	35
ETS2	7	3	0	10
IGS1	58	22	3	83
5S	2	1	0	3
IGS2	40	18	5	63
Total	162	57	8	227

[a] Complex mutations are those where transitions and transversions occur at the same nucleotide position but in different *S. cerevisiae* rDNA arrays (see Results).

With the exception of the intervening RNA polymerase III (Pol III) transcribed 5S rRNA gene (Fig. 2), and unlike both the ITS and ETS regions, the IGS is a nontranscribed region of rDNA. Despite this, it nevertheless contains a number of highly conserved *cis*-acting functional elements essential for chromosome function and maintenance of rDNA copy number. These include a replication fork barrier (RFB) site in IGS1 (Brewer et al. 1992; Kobayashi et al. 1992; Kobayashi 2003), and an origin of replication, also referred to as the ribosomal autonomously replicating sequence, or *rARS*, in IGS2 (Brewer and Fangman 1991). In addition to these, in a recent phylogenetic footprinting study of several *Saccharomyces* species, Ganley et al. (2005) identified a highly conserved sequence in IGS1, corresponding to a previously identified bidirectional RNA polymerase II (Pol II) promoter (Santangelo et al. 1988). Subsequent analyses of this sequence, which the investigators named E-pro (for expansion region [EXP] promoter), have shown it to play an important role in rDNA repeat amplification following copy number loss due to unequal sister chromatid recombination (Kobayashi et al. 2001; Kobayashi and Ganley 2005; Kobayashi 2006).

A distribution plot of (base substitution-only) polymorphisms, occurring in 20-bp bins, over the entire 2.28-kb IGS region is presented in Figure 2. The location of the 5S rRNA gene, as well as the three functional elements (RFB site, E-pro, and *rARS*), is indicated. Like Figure 1, this plot is derived from the combined data sets for all 34 strains. In total, 149 polymorphisms were detected in the IGS region, which if evenly distributed would mean that one site would occur, on average, every 15.3 nucleotides (i.e., ~1.3 sites per 20 nucleotides). However, it is evident from Figure 2 that this is not, in fact, the case. For instance, although smaller in size (IGS1, 915 bp; IGS2, 1243 bp), notably more polymorphisms are found in IGS1 (83 sites) compared with IGS2 (63 sites). Indeed, if the respective size of these two regions is taken into consideration, then IGS1 has nearly 1.8 times more polymorphisms than IGS2 (IGS1, ~1.8 sites per 20 nucleotides; IGS2, ~1 site per 20 nucleotides). Furthermore, even within IGS1, polymorphisms are not distributed evenly, with over half of all detected sites (47 of 83 sites) located in the first third of the spacer region. The location of the 5S rRNA gene, nucleotide positions 7774–7894 (corresponding to positions 916–1037 in Fig. 2), is also immediately apparent as only three polymorphic sites, all specific to the laboratory strains S288c and W303, were detected in this highly conserved rRNA gene. In contrast to IGS1, there appears to be a far more even distribution of polymorphisms across IGS2, with 33 sites found in the 5′ half, and 30 sites in the 3′ half of the spacer region.

We also examined the frequency and specific location of polymorphisms found in the three functional elements. In a previous mutational study of the *S. cerevisiae rARS*, Miller and Kowalski (1993) found that the introduction of a double point mutation within only one of the three internal ARS core-consensus sequences, or ACSs, present within this element (Celniker et al. 1984; van Houten and Newlon 1990; Miller and Kowalski 1993), was sufficient to abolish complete origin-of-replication activity. In our study, 37 polymorphisms were found distributed among the three functional elements, 17 in the 118-bp RFB site, 13 in the 141-bp E-pro sequence, and seven in the 107-bp *rARS* (Fig. 2). When polymorphism number was measured against the size of each element, the *rARS* was found to have the fewest polymorphisms

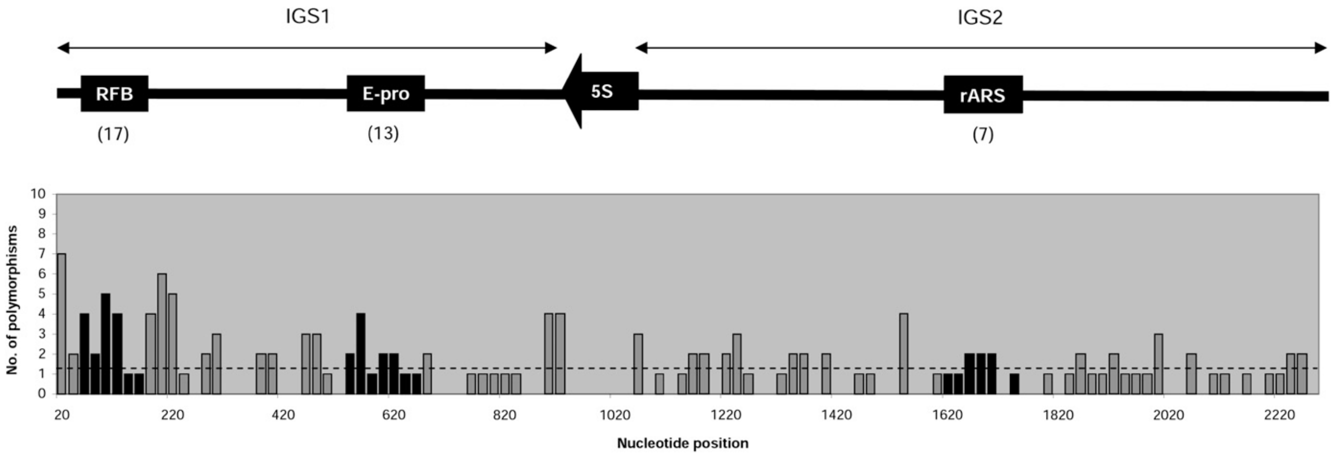

Figure 2. Distribution of (base substitution-only) polymorphisms, occurring in 20-bp bins, across the 2.28-kb IGS region. The approximate locations of the 5S rRNA gene, the RFB site, E-pro, and *rARS* are shown, along with the number of polymorphisms (in parentheses) identified in each *cis*-acting functional element. A trend line (dotted line) has also been added to show the polymorphism distribution if the 149 detected sites were evenly distributed across this region (i.e., at a frequency of ~1.3 sites per 20 nucleotides).

per 100 nucleotides (6.5), while the RFB site had the most (14.4). However, it is worth noting that in the case of the RFB site, 11 of the 17 sites (~65% of sites) are specific to UWOPS83-787-3. Furthermore, two of these elements, namely, the RFB site and *rARS*, are composed of core consensus sequences. These are named RFB1, RFB2, and RFB3 in the RFB site (Ward et al. 2000; Kobayashi 2003) and ACS1, ACS2, and ACS3 in the *rARS* (Miller and Kowalski 1993). In the *rARS*, only two of the seven detected polymorphisms, both T→C transitions, are found in the 11-bp ACSs (at positions 8490 [ACS1] and 8512 [ACS2]), reflecting the functional importance of these core sequences. In comparison, eight of the 17 RFB site polymorphisms are found in the core consensus sequences: three in RFB1, two in RFB2, and three in RFB3. However, of these, five are specific to UWOPS83-787-3.

Partial single nucleotide polymorphisms

Perhaps the most surprising finding of our study is the discovery that so many of the identified rDNA polymorphisms are not fully resolved to a true single nucleotide polymorphism (SNP) in any strain. This indicates that these polymorphisms have either not yet spread to all repeats or have yet to be lost from all repeats within the individual arrays. A similar observation was made by Ganley and Kobayashi (2007), who reported that most of the polymorphisms identified in their study, including two of the four *S. cerevisiae* RM11-1a polymorphisms, were present on only one or a few reads. As these polymorphisms are not present in every single repeat of the rDNA array, they cannot, by definition, be classified as conventional SNPs. Thus we introduce the term partial single nucleotide polymorphism, or pSNP, to describe this type of sequence variation, specific to repeat arrays such as the rDNA.

In total, 156 of the 227 polymorphisms (~69% of all sites) are not fully resolved in any of the 34 strains analyzed and were therefore classified as pSNPs. To examine these sites further, we calculated the frequency of each individual pSNP as a percentage of reads on which a specific base substitution occurs and recorded the highest value (maximum pSNP frequency) for each site. Some pSNPs were found to be strain specific, while others were shared between a number of strains but at differing frequencies. For example, in the ITS1 region, the T→C pSNP at position 2661 is specific to DBVPG 1853, occurring with a frequency of 24.4%, whereas the C→T pSNP at position 2683 is common to strains DBVPG 1853, NCYC 110, SK1, and Y55, occurring with frequencies of 19.4%, 52.7%, 85.5%, and 63.4%, respectively. In the latter example, the SK1 pSNP has the highest frequency (i.e., 85.5%) and so is recorded as the maximum pSNP value for this site. pSNPs were then grouped according to their maximum frequency values, in bins of 10% increments (i.e., <10%, 10%–19%, 20%–29%, etc), and Table 3 shows the resulting region-by-region breakdown. More than half of all pSNPs (~55%) are found in the IGS region (57 in IGS1 and 29 in IGS2) (Table 3), where the majority of rDNA variation occurs (Fig. 1; Table 2). When a distribution of maximum pSNP frequency values was plotted (Fig. 3), it was observed that a significant proportion of pSNPs (94 of 156 [~60%]) occur at low frequency (<10%) and are

Table 3. Region-by-region breakdown of pSNP frequency (occupancy ratio) totals

	pSNP[a] frequency (%)										
Region	<10	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–99	Total
ETS1	8	1	0	0	0	0	1	1	1	1	13
18S	3	0	0	0	0	0	0	0	0	2	5
ITS1	3	0	1	1	1	0	0	1	1	0	8
5.8S	0	0	0	0	0	0	0	0	0	0	0
ITS2	0	0	0	0	0	0	0	0	0	0	0
26S	29	2	1	0	1	0	0	0	0	1	34
ETS2	3	1	0	0	0	0	0	1	0	2	7
IGS1	35	5	1	2	2	2	4	1	1	4	57
5S	0	3	0	0	0	0	0	0	0	0	3
IGS2	13	4	2	1	2	0	1	2	1	3	29
Total	94	16	5	4	6	2	6	6	4	13	156

[a] Base substitution-only pSNPs.

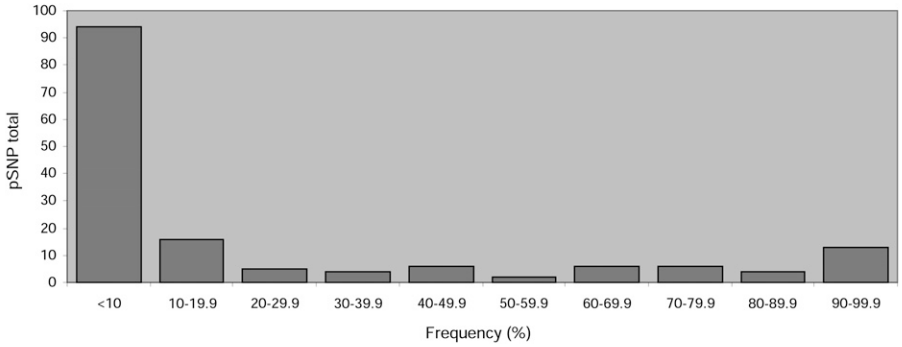

Figure 3. Distribution plot of rDNA pSNP frequency (occupancy ratio) values grouped in bins of 10% increments.

thus present in low copy number in the individual *S. cerevisiae* rDNA arrays. In fact, many of the detected 18S and 26S rDNA polymorphisms (32 of 40 sites [80%]) were found to be low frequency pSNPs, 17 of which have a maximum frequency of 5% or less, undoubtedly reflecting the fact that these genes, in comparison to the transcribed and nontranscribed spacer regions, are under greater functional constraint. Interestingly, when we reexamined the RFB, E-pro, and *rARS* polymorphisms, we found that out of the three functional elements, the RFB site has by far the highest number of pSNPs (14 of 19 polymorphisms [~82% of sites]) (Fig. 4A). In contrast, E-pro has six pSNPs (~46% of sites), and the *rARS* has just two pSNPs (~13% of sites) (Fig. 4B). Furthermore, of the 14 RFB pSNPs, 11 occur at low frequency, with nine of these occurring at a frequency of 5% or less, while the *rARS* has one low frequency pSNP, and E-pro has none.

pSNP number and depth of sequencing coverage

When strains were ordered according to depth of sequencing coverage (see Supplemental Fig. 2), we found no indication to suggest that strains with low coverage (i.e., <1×) possessed significantly fewer pSNPs than strains whose genomes had been sequenced to a greater depth (Pearson's correlation coefficient = −0.03451147; Spearman rank test, correlation coefficient = −0.07392106). For example, although NCYC 361 received considerably less coverage compared with Y55 (0.65× and 3.83×, respectively) (Table 4), both strains were found to possess similar numbers of pSNPs (31 and 25, respectively) (Table 4). However, many of the pSNPs identified in this study (~60%) occur at low frequency and are present on <10% of reads. This means that for some strains such as K11, which not only received

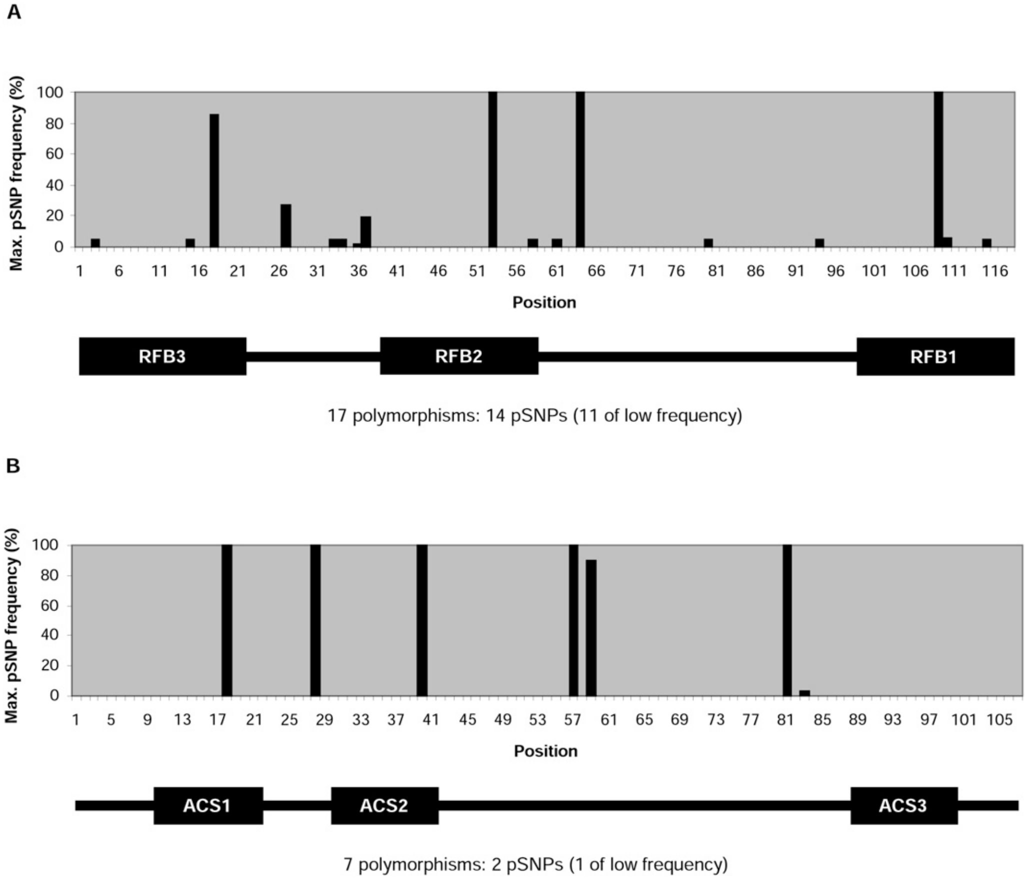

Figure 4. Maximum frequency (occupancy ratio) distribution plots of polymorphisms found in the RFB site, nucleotide positions 6913 (1) to 7030 (118) (*A*), and the *rARS*, nucleotide positions 8473 (1) to 8579 (107) (*B*).

low sequencing coverage (0.88×) (Table 4) but is also estimated to have a relatively small rDNA copy number (54 repeats) (Supplemental Table 1), it is likely that the stringent (quality score) filtering strategy employed in this study underestimates the actual number of low frequency pSNPs present in their rDNA arrays. Furthermore, the fact that strain-specific single read polymorphisms were excluded, for reasons already discussed, would also contribute to an overall underestimate in low frequency pSNP numbers. While it is likely that additional low frequency pSNPs remain to be discovered, the lack of correlation between genome sequencing coverage and pSNP number indicates that the patterns of pSNP variation we have described are unlikely to change significantly with increased coverage.

pSNP number and genome type

As with overall polymorphism totals, we found that the number of pSNPs varies quite markedly between individual strains and ranges from two in YPS128 and DBVPG 1788 to 71 in UWOPS83-787-3. No correlation was observed between either total polymorphism number and rDNA copy number, or pSNP total and rDNA copy number (data not shown). However, when the strains were ordered according to their pSNP totals, we discovered an interesting correlation between genome type and pSNP number (Table 4). As described in the recent population genomics study of *S. cerevisiae* and its wild relative *S. paradoxus*, *S. cerevisiae* strains can be subdivided into two groups according to their genome structure (Liti et al. 2009). One group, which can be further subdivided into five geographical subgroups (i.e., strain sets from Malaysia, North America, West Africa, Wine/Europe, and sake and related fermentations), has "clean" structured genomes, while the other group has mosaic-like genomes of hybrid origin (see Supplemental Table 1). Examples of the latter include the laboratory strains SK1 and Y55, which appear to be derived from crosses between representatives from the West African and Wine/European lineages. When we ordered the strains according to their pSNP totals, we observed that strains with structured genomes have, on average, fewer pSNPs in their rDNA arrays than do strains with mosaic-like genomes (Table 4). With the exception of UWOPS05-217-3 (30 pSNPs), structured genome strains have between two and 17 pSNPs per rDNA array. In contrast, mosaic genome strains, with the exception of UWOPS87-2421 (nine pSNPs), have between 20 and 71 pSNPs per rDNA array. In fact, strains with structured genomes have, on average, 2.9 times fewer pSNPs (structured, 11.1 pSNPs per strain; mosaic, 31.8 pSNPs per strain) (Table 4). Furthermore, although individual strains received different depths of (sequencing) coverage (Table 4), both strain sets were found to have received comparable average depth of coverage (structured, 1.14× per strain; mosaic, 1.43× per strain). Thus, the difference observed between the pSNP totals of mosaic and structured strains cannot be attributed to variation in genome sequencing coverage. Despite the fact that these results are based on the analysis of a limited number of strains (34), this finding nevertheless suggests that rDNA pSNP count might be used as a simple, rapid method for determining the genome type (i.e., structured or of hybrid origin) of *S. cerevisiae* strains, and perhaps of other species as well.

Table 4. *S. cerevisiae* **strain ordering based on rDNA pSNP number**

Strain	Genome type[a]	pSNP total	Coverage[b]
DBVPG 1788	Structured	2	1.18
YPS128	Structured	2	1.28
YJM975	Structured	6	1.03
YJM981	Structured	6	0.84
YPS606	Structured	8	1.59
K11	Structured	9	0.88
UWOPS87-2421	Mosaic	9	0.95
BC187	Structured	10	0.67
DBVPG 1106	Structured	10	0.69
DBVPG 1373	Structured	10	1.28
UWOPS03-461-4	Structured	10	0.99
NCYC 110	Structured	11	0.85
YJM978	Structured	11	1.06
L_1374	Structured	12	1.01
UWOPS05-227-2	Structured	12	1.02
DBVPG 6765	Structured	15	3.28
Y12	Structured	16	0.88
Y9	Structured	17	0.81
DBVPG 6044	Structured	13	1.4
SK1	Mosaic	20	3.58
S288c	Mosaic	23	1.2
W303	Mosaic	23	2.44
Y55	Mosaic	25	3.83
322134S	Mosaic	27	0.98
DBVPG 1853	Mosaic	28	0.97
YS4	Mosaic	28	1.08
UWOPS05-217-3	Structured	30	0.99
NCYC 361	Mosaic	31	0.65
YIIc17_E5	Mosaic	34	1.01
YS9	Mosaic	36	1.03
273614N	Mosaic	38	0.93
378604X	Mosaic	40	1.05
DBVPG 6040	Mosaic	44	0.86
UWOPS83-787-3	Mosaic	71	0.94

No. of strains	Genome type	pSNP total	pSNP average
19	Structured	210	11.1
15	Mosaic	477	31.8

[a]For a more detailed discussion of individual genome structure, see Liti et al. (2009).
[b]Depth of genome sequencing coverage.

Discussion

To our knowledge, this study has presented the first in-depth quantitative analysis of base substitution polymorphisms present within the *S. cerevisiae* rDNA array using a relatively large set of strains isolated from a wide variety of different sources and geographic locations (Liti et al. 2009). Contrary to both expectation and results from a previous study by Ganley and Kobayashi (2007), we found that the level of variation present within individual *S. cerevisiae* rDNA arrays is surprisingly high and can differ by nearly an order of magnitude between individual strains. Indels were excluded from our study, due to computational issues concerning the accurate and unbiased measurement of multiple position indels, particularly when found in long homopolymeric tracts. However, results from a preliminary analysis of a subset of IGS1 sequence reads indicate that the overall levels of variation (i.e., base substitutions plus indels) within individual *S. cerevisiae* rDNA arrays are likely to be much higher. Furthermore, it is also likely that the number of single-copy polymorphisms detected is underestimated, due to the stringent filtering strategy used to minimize the introduction of false positives due to sequencing error. This would especially apply to those strains whose genomes had been sequenced to onefold coverage or less. In such strains, single-copy polymorphisms would be expected to appear, at most, in only a single covering read. Nevertheless, depth of sequencing coverage can be discounted as the main cause of pSNP variation observed.

When the S288c rDNA consensus sequence (Goffeau et al. 1996) was used to map the positions of all the detected polymorphisms in each individual array, we found that the variation is not distributed evenly over the rDNA repeat. In fact, many of the detected polymorphisms are found in the nontranscribed IGS region. This result is perhaps not unexpected, as of all the coding and noncoding regions within the rDNA, the IGS is under the least functional constraint and is, as a consequence, evolving the most rapidly. For instance, in *Drosophila*, the IGS region is composed of subrepeat sequences, and these can differ in number (from one to six), sequence, and length between different species (Tautz et al. 1987; Stage and Eickbush 2007). Likewise, in human rDNA arrays, individual IGS units can vary considerably in length, ranging from 9–72 kb, with an average length of 34.2 ± 5.4 kb (Caburet et al. 2005). Indeed, for this very reason, the rDNA IGS has been specifically used in a number of yeast phylogenetic studies to examine both inter- and intraspecific relationships (Molina et al. 1993; Fell and Blatt 1999; Diaz et al. 2000, 2005). However, even within the IGS, it is apparent that the distribution of polymorphisms is uneven. Polymorphism distribution is clearly influenced by the presence of the intervening and highly conserved 5S rRNA gene, which separates IGS1 from IGS2. Furthermore, although notably smaller in size, a larger number of base substitution polymorphisms are found in IGS1.

Less clear is the precise degree of influence upon polymorphism distribution imparted by the presence of *cis*-acting elements such as the RFB site and the ribosomal autonomous replicating sequence (*rARS*) (Brewer and Fangman 1991; Brewer et al. 1992; Kobayashi et al. 1992; Kobayashi 2003). Despite their functional importance, both for chromosome function and maintenance of rDNA copy number, none of the three elements (RFB site, E-pro, or *rARS*) examined in this study were found to contain significantly less polymorphisms than the individual IGS region (IGS1 or IGS2) in which they are located. Indeed, in the case of the RFB site, this functional element contains notably more polymorphisms (per 20 nucleotides) than the surrounding IGS1. One possible explanation for this, certainly in the case of the *rARS*, is the fact that this element, like the RFB site, is composed of three short core consensus sequences (ACSs) known to be essential for function (Celniker et al. 1984; van Houten and Newlon 1990; Miller and Kowalski 1993). Upon re-examination, slightly less variation is found in the three ACSs compared with the entire *rARS* element, suggesting that not every single nucleotide in this element is necessarily essential and that a certain degree of sequence variation can possibly be tolerated without a loss of overall function. Evidence to support this comes from the mutational study of the *S. cerevisiae rARS* by Miller and Kowalski (1993), who discovered that the introduction of a double point mutation (TT to AA) at the ninth and 10th positions in ACS1 is sufficient to abolish complete origin-of-replication activity. Whereas, when the same double mutation was introduced into either ACS2 or ACS3, at the same positions, overall activity was retained.

Significantly, many of the rDNA array polymorphisms identified in each strain are not present on every covering sequence read of that strain. A comparative analysis of rDNA polymorphism frequency reveals that most, in fact, are present on 10% or less of covering reads (i.e., occur at a relatively low frequency). This indicates that many of the detected base substitutions are some way from either being fully removed or from becoming fixed polymorphisms (SNPs), as they are present on only a small fraction of the total number of repeats within each individual rDNA array. Similar observations were made by Ganley and Kobayashi (2007) in their study of five fungal rDNA arrays, including that of *S. cerevisiae* RM11-1a, and by Stage and Eickbush (2007) in their study of *Drosophila* rDNA arrays. By definition, these polymorphisms cannot be classified as conventional SNPs. Since they appear to be a more widespread phenomenon, being present in both fungal and nonfungal species, we have introduced the term partial single nucleotide polymorphism, or pSNP, to describe this type of sequence variation, which is present in rDNA and which may also exist in other large repeat arrays. It is worth noting that although only four polymorphisms were found in *S. cerevisiae* RM11-1a, none were fully resolved and so all can be described as pSNPs according to our definition. The level of rDNA variation found in RM11-1a, while lower than average, is still comparable to that found in the rDNA arrays of some of the strains analyzed in the present study (e.g., YPS128 and DBVPG 1788).

No correlation was observed between the number of pSNPs in a strain and rDNA copy number (Liti et al. 2009). However a relationship was discovered between pSNP number and hybridization history (i.e., genome type). When strains were ordered according to their pSNP number, all but one of the strains previously identified as having clean structured genomes (Liti et al. 2009), were found to have, on average, significantly fewer pSNPs than strains found to have mosaic genomes, i.e., genomes of hybrid origin. Although the rDNA arrays of further *S. cerevisiae* strains will need to be studied, our finding suggests that rDNA pSNP number might be used more broadly in population genetics studies as a simple indicator of genome structure, perhaps precluding the need for more extensive genome sequencing and analysis.

Irrespective of genome type (i.e., structured or mosaic), polymorphisms continually appear and disappear within the rDNA array as a result of the ongoing processes of mutation and recombination (Gangloff et al. 1996). However, in the case of mosaic strains (i.e., strains of hybrid origin), additional variation can be introduced by fixed differences (SNPs) present in the respective parental rDNA arrays. Since the *S. cerevisiae* rDNA array is ~1.5 Mbp in size (Petes 1979; Kobayashi et al. 1998; Kobayashi 2006), some degree of crossover (due to recombination) between the parental rDNA arrays is inevitable. Hence mating events between strains from different "geographical" lineages provide an immediate means by which additional pSNPs can arise within the array, other than by point mutation. In such a cross, the resulting hybrid would be expected to possess a larger number of pSNPs in its rDNA array than either of its parents. This expanded population of pSNPs would then be expected to persist in the hybrid rDNA array until all, or at least the majority of repeats had undergone homogenization as a result of concerted evolution (Zimmer et al. 1980). In the case of the laboratory strains SK1 and Y55, this certainly appears to be true. Both are believed to be derived from crosses between West African and European/Wine strains (Liti et al. 2009), and both have more pSNPs (in their rDNA arrays) than any of the West African and European/Wine strains included in this study. Indeed, an analysis of SNP-only rDNA polymorphisms suggests that the parental West African strain of both SK1 and Y55 may be more like DBVPG 6044 (isolated from bili wine) than NCYC 110 (isolated from ginger beer) (data not shown). In the case of UWOPS87-2421, this mosaic strain is somewhat atypical as it has a pSNP number (nine) similar to that of structured strains (two to 13). One conceivable explanation for this could be that, in contrast to other mosaic strains (e.g., SK1 and Y55), UWOPS87-2421 is derived from a cross between strains of the same geographical lineage (which would likely have fewer fixed differences in their respective rDNA arrays). Indeed, in their population genomics study, Liti et al. (2009) found UWOPS87-2421, isolated from a cactus (*Opuntia megacantha*) in Hawaii, to display a loose (phylogenetic) affinity to the North American strain lineage (as represented by YPS128 and YPS606).

In *S. cerevisiae*, a number of studies have shown that although this yeast possesses more than 100 copies of the rDNA repeat only half are actually actively transcribed (Dammann et al. 1993; French et al. 2003). In fact, Kobayashi and colleagues have created a strain of *S. cerevisiae*, which possesses only one seventh of the wild-type rDNA copy number and yet still retains normal viability in the laboratory (Takeuchi et al. 2003). This suggests that the *S. cerevisiae* rDNA array may be able to accommodate/tolerate a certain degree of sequence variation between its individual repeats without suffering any overall deleterious loss of function. It is therefore notable that many of the pSNPs identified in this study are not only found in the nontranscribed regions of the rDNA but also occur at a relatively low frequency (i.e., in less than 10% of covering reads), presumably below a threshold level that would otherwise compromise fitness. The fact that the pSNPs are not uniformly distributed across the repeat also indicates that those that arise in coding regions (e.g., the 18S and 26S rRNA genes) are eliminated much more rapidly that those that appear in regions such as the IGS where sequence variation is less critical to cellular function. While beyond the scope of the present study, we can nevertheless speculate that the repeats in which many of these pSNPs are found could reside in inactive (nucleosomal) regions of the (rDNA) array (Dammann et al. 1993; Ganley and Kobayashi 2007; Kobayashi 2008). If so, then this would suggest that chromatin structure, as well as regulating rRNA transcription (Dammann et al. 1993), may also, in some way, influence the spread of polymorphisms in the rDNA array.

In conclusion, as more genomes are sequenced or resequenced, analyses of WGSS data, similar to that carried out in this study and those by Ganley and Kobayashi (2007) and Stage and Eickbush (2007) will undoubtedly help to establish whether the rDNA arrays of other species display similar levels of cryptic variation. Indeed extension of this approach to other repeat families will help to elucidate whether this type of sequence variation is specific to rDNA arrays or a more widespread phenomenon. Furthermore, examination of pSNPs may also provide a simple measure of genome mosaicism and hybrid history in population genetics studies of other species, both fungal and nonfungal. Analysis of shared and private pSNPs among different lineages could provide a novel measure of rates of recombination and gene conversion across lineages, and allow rigorous testing of hypotheses of rDNA function in cellular ageing and maintenance of genome integrity (Kobayashi 2008).

Methods

Identifying rDNA polymorphisms

BLAST (Altschul et al. 1997) was used to identify and tentatively align reads containing rDNA sequences; MUSCLE (Edgar 2004), to perform multiple alignment of the results across all strains. Script parameters and quality score filters were selected to minimize the introduction of non-rDNA or erroneous sequences into the final multiple alignments.

Raw sequencing reads were processed in three stages. In stage one, all reads containing rDNA sequences were identified by BLASTing the complete SGRP *S. cerevisiae* WGSS database. Using the rDNA consensus sequence derived from the reference strain S288c (Goffeau et al. 1996), a series of 100-bp (rDNA) query sequences were selected at 20-bp sliding intervals. These sequences were used for gapped BLAST queries against the *S. cerevisiae* (WGSS) database to identify all rDNA sequence reads. At this stage, to ensure that non-rDNA sequences were not selected, only alignments spanning at least 70 bp of the 100-bp query and having no more than 30 mismatches were accepted. All accepted reads from stage one were assembled into a new rDNA-only database and formatted for BLAST searching.

In stage two, using the newly created rDNA-only database, less stringent BLAST searches were performed to find alignments that might be quite divergent from the S288c consensus sequence. False positives, due to sequencing error, were accepted at this stage to ensure comprehensive sampling of variability. In order to generate such alignments, minimal values were used (i.e., for gap opening [−3], gap extension [−1], mismatch [−1], and extension penalty [1]). The same series of rDNA query sequences as in stage 1 were used. However, in this instance, the central 20 bp of each query sequence was considered the target for polymorphism analysis. The flanking 40 bp on either side of the central 20-bp target were buffer sequences used to ensure the specificity of the BLAST searches. Complete coverage of the entire rDNA repeat was ensured by this sliding window technique. In stage two, BLAST alignments were only accepted if they spanned at least 62 bp of the query sequence, and these were collected for subsequent multiple alignment.

In stage three, multiple alignments were performed using MUSCLE (Edgar 2004), with the default parameters. To improve overall efficiency, all redundant reads were excluded from the alignments at this stage. Although very few reads are identical in their entirety, the majority of 100-bp substrings for a particular alignment window of interest are identical. The original BLAST window query was added to the stack of read fragments to be aligned so the output alignments could be compared to the original S288c rDNA repeat consensus sequence.

Finally, in order to distinguish authentic sequence variation from variation due to sequencing error, *phred* quality (q) scores (Ewing and Green 1998; Ewing et al. 1998), published along with the original SGRP WGSS database, were extracted and analyzed. To minimize the expected number of false positives, due to sequencing error, to less than one, stringent quality score filtering was employed. Candidate polymorphisms (base substitutions only) were accepted only if they appeared on covering reads with a quality score of 40 or above. Furthermore, polymorphisms found on single reads (single read polymorphisms) were only accepted if the polymorphism was shared by two or more strains. All single read polymorphisms unique to a single strain were excluded, as we could not reliably discount the possibility that these had arisen due to sequencing errors. Using this conservative filtering strategy, rDNA polymorphisms were identified and their positions mapped (in relation to the S288c rDNA repeat consensus sequence) for each individual *S. cerevisiae* strain. In addition, the percentage frequency for each individual polymorphism was calculated (i.e., number of accepted reads carrying the polymorphism divided by the total number of accepted covering reads).

Acknowledgments

We thank Ed Louis and Gianni Liti (Institute of Genetics, University of Nottingham) for invaluable discussion and comments on this manuscript. S.A.J., R.P.D., and I.N.R. were supported by the BBSRC. D.M.C. was supported by the Wellcome Trust. M.J.T.O. and A.v.O. were supported by NSF, NIH, and the Hertz Foundation.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipmann, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Ben Ali, A., Wuyts, J., De Wachter, R., Meyer, A., and Van de Peer, Y. 1999. Construction of a variability map for eukaryotic large subunit ribosomal RNA. *Nucleic Acids Res.* **27:** 2825–2831.

Brewer, B.J. and Fangman, W.L. 1991. Mapping replication origins in yeast chromosomes. *Bioessays* **13:** 317–322.

Brewer, B.J., Lockshon, D., and Fangman, W.L. 1992. The arrest of replication forks in the rDNA of yeast occurs independently of transcription. *Cell* **71:** 267–276.

Caburet, S., Conti, C., Schurra, C., Lebofsky, R., Edelstein, S.J., and Bensimon, A. 2005. Human ribosomal RNA gene arrays display a broad range of palindromic structures. *Genome Res.* **15:** 1079–1085.

Celniker, S.E., Sweder, K., Srienc, F., Bailey, J.E., and Campbell, J.L. 1984. Deletion mutations affecting autonomously replicating sequence *ARS1* of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **4:** 2455–2466.

Dammann, R., Lucchini, R., Koller, T., and Sogo, J.M. 1993. Chromatin structures and transcription of rDNA in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **21:** 2331–2338.

De Rijk, P., Neefs, J.-M., Van de Peer, Y., and De Wachter, R. 1992. Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res.* **20:** 2075–2089.

Diaz, M.R., Fell, J.W., Boekhout, T., and Theelen, B. 2000. Molecular sequence analyses of the intergenic spacer (IGS) associated with rDNA of the two varieties of the pathogenic yeast, *Cryptococcus neoformans*. *Syst. Appl. Microbiol.* **23:** 535–545.

Diaz, M.R., Boekhout, T., Kiesling, T., and Fell, J.W. 2005. Comparative analysis of the intergenic spacer regions and population structure of the species complex of the pathogenic yeast *Cryptococcus neoformans*. *FEMS Yeast Res.* **5:** 1129–1140.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.

Eickbush, T.H. and Eickbush, D.G. 2007. Finely orchestrated movements: Evolution of the ribosomal RNA genes. *Genetics* **175:** 477–485.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Fell, J.W. and Blatt, G.M. 1999. Separation of strains of the yeasts *Xantophyllomyces dendrorhous* and *Phaffia rhodozyma* based on the rDNA IGS and ITS sequence analysis. *J. Ind. Microbiol. Biotechnol.* **23:** 677–681.

French, S.L., Osheim, Y.N., Cioci, F., Nomura, M., and Beyer, A.L. 2003. In exponentially growing *Saccharomyces cerevisiae* cells, rRNA synthesis is determined by the summed RNA polymerase I loading rate rather than by the number of active genes. *Mol. Cell. Biol.* **23:** 1558–1568.

Gangloff, S., Zou, H., and Rothstein, R. 1996. Gene conversion plays the major role in controlling the stability of large tandem repeats in yeast. *EMBO J.* **15:** 1715–1725.

Ganley, A.R.D. and Kobayashi, T. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res.* **17:** 184–191.

Ganley, A.R.D., Hayashi, K., Horiuchi, T., and Kobayashi, T. 2005. Identifying gene-independent non-coding functional elements in the yeast ribosomal DNA by phylogenetic footprinting. *Proc. Natl. Acad. Sci.* **102:** 11787–11792.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546–567.

Hassouna, N., Michot, B., and Bachellerie, J.P. 1984. The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Res.* **12:** 3563–3583.

Jemtland, R., Maehlum, E., Gabrielsen, O.S., and Øyen, T.B. 1986. Regular distribution of length heterogeneities within non-transcribed spacer regions of cloned and genomic rDNA of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **14:** 5145–5158.

Kobayashi, T. 2003. The replication fork barrier site forms a unique structure with Fob1p and inhibits the replication fork. *Mol. Cell. Biol.* **23:** 9178–9188.

Kobayashi, T. 2006. Strategies to maintain the stability of the ribosomal RNA gene repeats. *Genes Genet. Syst.* **81:** 155–161.

Kobayashi, T. 2008. A new role of the rDNA and nucleolus in the nucleus-rDNA instability maintains genome integrity. *Bioessays* **30:** 267–272.

Kobayashi, T. and Ganley, A.R.D. 2005. Recombination regulation by transcription-induced cohesion dissociation in rDNA repeats. *Science* **309:** 1581–1584.

Kobayashi, T., Hidaka, M., Nishizawa, M., and Horiuchi, T. 1992. Identification of a site required for DNA replication fork blocking activity in the rRNA gene cluster in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.* **233:** 355–362.

Kobayashi, T., Heck, D.K., Nomura, M., and Horiuchi, T. 1998. Expansion and contraction of ribosomal DNA repeats in *Saccharomyces cerevisiae*: Requirement of replication fork blocking (Fob1) protein and the role of RNA polymerase I. *Genes & Dev.* **12:** 3821–3830.

Kobayashi, T., Nomura, M., and Horiuchi, T. 2001. Identification of DNA *cis* elements essential for expansion of ribosomal DNA repeats in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **21:** 136–147.

Lachance, M.A., Daniel, H.M., Meyer, W., Prasad, G.S., Gautam, S.P., and Boundy-Mills, K. 2003. The D1/D2 domain of the large-subunit rDNA of the yeast species *Clavispora lusitaniae* is unusually polymorphic. *FEMS Yeast Res.* **4:** 253–258.

Liti, G., Carter, D.M., Moses, A.M., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Blomberg, A., Warringer, J., Burt, A., et al. 2009. Population genomics of domestic and wild yeasts. *Nature* doi: 10.1038/nature07743.

Miller, C.A. and Kowalski, D. 1993. *cis*-Acting components in the replication origin from ribosomal DNA of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13:** 5360–5369.

Molina, F.I., Shung-Chang, J., and Huffman, J.L. 1993. PCR amplification of the 3′ external transcribed and intergenic spacers of the ribosomal DNA repeat unit in the three species of *Saccharomyces*. *FEMS Microbiol. Lett.* **108:** 259–264.

Montrocher, R., Verner, M.-C., Briolay, J., Gautier, C., and Marmeisse, R. 1998. Phylogenetic analysis of the *Saccharomyces cerevisiae* group based on polymorphisms of rDNA spacer sequences. *Int. J. Syst. Bacteriol.* **48:** 295–303.

O'Donnell, K. and Cigelnik, E. 1997. Two divergent intragenomic rDNA ITS2 types within a monophyletic lineage of the fungus *Fusarium* are nonorthologous. *Mol. Phylogenet. Evol.* **7:** 103–116.

Petes, T.D. 1979. Yeast ribosomal DNA genes are located on chromosome XII. *Proc. Natl. Acad. Sci.* **76:** 410–414.

Rothstein, R.J. 1977. A genetic fine structure analysis of the suppressor 3 locus in *Saccharomyces*. *Genetics* **85:** 55–64.

Rothstein, R.J., Esposito, R.E., and Esposito, M.S. 1977. The effect of ochre suppression on meiosis and ascospore formation in *Saccharomyces*. *Genetics* **85:** 35–54.

Santangelo, G.M., Tornow, J., McLaughlin, C.S., and Moldave, K. 1988. Properties of promoters cloned randomly from the *Saccharomyces cerevisiae* genome. *Mol. Cell. Biol.* **8:** 4217–4224.

Skryabin, K.G., Eldarov, M.A., Larionov, V.L., Bayev, A.A., Klootwijk, J., de Regt, V.C.H.F., Veldman, G.M., Planta, R.J., Georgiev, O.I., and Hadjiolov, A.A. 1984. Structure and function of the nontranscribed spacer regions of yeast rDNA. *Nucleic Acids Res.* **12:** 2955–2968.

Stage, D.E. and Eickbush, T.H. 2007. Sequence variation within the rRNA loci of 12 *Drosophila* species. *Genome Res.* **17:** 1888–1897.

Takeuchi, Y., Horiuchi, T., and Kobayashi, T. 2003. Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes & Dev.* **17:** 1497–1506.

Tautz, D., Tautz, C., Webb, D., and Dover, G.A. 1987. Evolutionary divergence of promoters and spacers in the rDNA family of four *Drosophila* species. *J. Mol. Biol.* **195:** 525–542.

Van de Peer, Y., Jansen, J., De Rijk, P., and De Wachter, R. 1997. Database on the structure of small ribosomal subunit RNA. *Nucleic Acids Res.* **17:** 111–116.

van Houten, J.V. and Newlon, C.S. 1990. Mutational analysis of the consensus sequence of a replication origin from yeast chromosome III. *Mol. Cell. Biol.* **10:** 3917–3925.

Ward, T.R., Hoang, M.L., Prusty, R., Lau, C.K., Keil, R.L., Fangman, W.L., and Brewer, B.J. 2000. Ribosomal DNA replication fork barrier and *HOT1* recombination hot spot: Shared sequences but independent activities. *Mol. Cell. Biol.* **20:** 4948–4957.

Zimmer, E.A., Martin, S.L., Beverley, S.M., Kan, Y.W., and Wilson, A.C. 1980. Rapid duplication and loss of genes coding for the α chains of haemoglobin. *Proc. Natl. Acad. Sci.* **77:** 2158–2162.

Received August 15, 2008; accepted in revised form December 23, 2008.