Genome-wide distribution of DNA replication origins at AT-rich islands in *Schizosaccharomyces pombe*

Mónica Segurado, Alberto de Luis and Francisco Antequera*

Instituto de Microbiología Bioquímica
CSIC/Universidad de Salamanca
Edificio Departamental
Campus Miguel de Unamuno
37007 Salamanca. Spain

*Corresponding author:
Francisco Antequera
Instituto de Microbiología Bioquímica
CSIC/Universidad de Salamanca
Edificio Departamental
Campus Miguel de Unamuno
37007 Salamanca. Spain

Tel: 34 923 121778
Fax: 34 923 224876
e-mail: CpG@usal.es

Running title: Genome-wide distribution of ORIs in *S. pombe*

ABSTRACT

Genome-wide analysis of replication dynamics requires the prior identification of DNA replication origins (ORIs). However, variability among them makes it difficult to predict their distribution across the genome on the basis of their sequence. We report here that ORIs in *Schizosaccharomyces pombe* coincide with discrete chromosomal AT-rich islands of up to 1 kb long characterized by a distinctive A+T content that clearly differentiates them from the rest of the genome. Genome-wide analysis has allowed us to identify 384 of these regions, which predicts the position of most ORIs in the genome, as shown by functional replication analyses. AT-rich islands occur at the mating locus, centromeres, and subtelomeric regions at a density approximately 4-fold higher than elsewhere in the genome, suggesting a link between the origin recognition complex (ORC) and transcriptional silencing in these regions. The absence of consensus elements in AT-rich islands implies that different sequences can target ORC to different ORIs.

Keywords: DNA replication / Replication origins / AT-rich islands / Genome organization / *S. pombe*

INTRODUCTION

Eukaryotic DNA replication origins (ORIs) have been identified in a variety of organisms including fungi, insects and mammals and they have been well characterised biochemically and genetically in the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. In *S. cerevisiae*, ORI regions span less than 150 bp and include one or several copies of an 11 bp ARS (autonomous replicating sequence) consensus sequence (Broach et al., 1983; Newlon and Theis, 1993) which is essential for binding of the origin recognition complex (ORC). An extended version of this consensus, that improves the identification of ARS elements, has been reported (Theis and Newlon, 1997). In addition, ORIs contain three or four partially redundant auxiliary elements whose sequence and distribution varies between them (Theis and Newlon, 1997; 2001). Two recent approaches, based on chromatin immunoprecipitation and density labelling of replication intermediates, have predicted the distribution of approximately 400 putative ORIs in *S. cerevisiae* (Wyrick *et al.*, 2001; Raghuraman *et al.*, 2001).

S. pombe ORIs require a minimum length of 0.5 kb to 1 kb and do not have recognisable consensus elements. However, functional dissection analyses have identified several A+T rich elements, that frequently contain stretches of asymmetrical adenines or thymines, whose length and number are not conserved between different ORIs. Some of these elements are individually or collectively required for ORI activity in plasmids and in the chromosome (Clyne and Kelly, 1995; Dubey *et al.*. 1996; Okuno *et al.*, 1999; Takahashi *et al.,* 2003). A key advance in our understanding of how ORC is targeted to ORIs in *S. pombe* was the discovery of

the unique structure of its Orc4 subunit (Chuang and Kelly, 1999). The N terminus of the SpOrc4 protein contains 9 AT-hook domains that are essential for ORC binding to ORIs *in vitro* and *in vivo* (Lee *et al.*, 2001; Kong and DePamphilis, 2002). These domains recognize the structure of AT-rich stretches through the minor groove of DNA without the requirement of a specific nucleotide sequence (Reeves and Beckerbauer, 2001). We report here the identification of discrete genomic regions up to 1 kb long with a disctinctively high A+T composition of which approximately 90% colocalise with active ORIs. The properties of these AT-rich islands may account for the specific properties of *S. pombe* ORIs and provide a frame of reference for future analyses of replication dynamics in this yeast.

RESULTS

Identification of AT-rich islands at origins of replication in *S. pombe*

All DNA replication origins identified to date in *S. pombe* are located in intergenic regions upstream from genes, although not all of these regions are competent in initiating replication (Dubey *et al.*, 1994; Okuno *et al.*, 1999; Gómez and Antequera, 1999). This observation, together with the fact that the AT-hook domains at the N terminus of the SpOrc4 protein is required for ORC binding to ORIs (Chuang and Kelly, 1999; Lee *et al.*, 2001; Kong and DePamphilis, 2002), raised the possibility that differences in the A+T content might be a determinant of ORI activity. To test this possibility, we compared base composition across 16 regions containing active genomic ORIs previously identified in our laboratory (Gómez and Antequera, 1999; Segurado *et al.*, 2002) with another 16 of similar length, also upstream from genes but devoid of ORI activity (Table 1 of Supplementary Material). Given that the shortest DNA fragments capable of conferring autonomous replication to plasmids in *S. pombe* range between 0.5 kb and 1 kb (Maundrell *et al.*, 1988; Dubey *et al.*, 1994; Clyne and Kelly, 1995), we determined base composition using sliding windows of different sizes within this interval. We found that the highest A+T content for each window was significantly higher for ORI-containing regions than for those that replicated passively, and that the 32 regions analysed could be classified in two distinct and non-overlapping groups (Figure 1A) (see Table 1 of Supplementary Material for the specific value for each window). These differences were also evident when the A+T content was plotted across long genomic regions containing ORIs such as, for example, ORI 12 and ORI tug1 (Gómez and Antequera, 1999; Segurado *et al.*, 2002) (Figure 1B). Based on these observations, we defined AT-rich islands as

regions between 0.5 kb and 1 kb whose A+T content was equal or above the following values for every window size (500 bp: 75 %; 600 bp: 74.5 %; 700 bp: 74 %; 800 bp: 73 %; 900 bp: 72.5 %; 1000 bp: 72 %). When a particular region fullfilled the criterium for all window sizes except for one, a 0.5% reduction in the A+T content for this particular size was allowed.

To further assess the association between AT-rich islands and ORIs, we tested whether another set of 14 ORIs identified by other authors using various different approaches would also co-localise with them (green triangles in Figure 2). Base composition analysis across regions containing *ars3001* at the rRNA gene cluster (Sánchez *et al.*, 1998), *ars3003* and *ars3002* (Dubey *et al.*, 1994), and *ars2004* (Okuno *et al.*, 1997; 1999) showed that this was indeed the case for all of them. AT-rich islands were also found to span four ORIs in centromere II (Smith *et al.*, 1995) and eight autonomous replicating sequences (ARS) identified by Maundrell *et al.*, (1988), at least two of which have been shown to act as chromosomal ORIs (Dalgaard and Klar, 2001; Segurado *et al.*, 2002) (Table 2 of Supplementary Material). Examples corresponding to *ars3003*, *ars3002*, *ars2004* and *ars1* are shown in Figure 1C. This Figure also shows that AT-rich islands overlap with the shortest DNA fragments capable of maintaining full ARS activity, and include the replication initiation point (RIP) in *ars2004* (Okuno *et al.*, 1997) and *ars1* (Gómez and Antequera, 1999). AT-rich islands do not extend across the entire intergenic regions, as illustrated by island 1003 (see below), which spans only a discrete fraction of the 5 kb long intergenic region between two divergent genes (Figure 1C).

Genome-wide distribution of AT-rich islands and replication origins

The fact that the 30 previously known ORIs (Tables 1 and 2 of Supplementary Material) were associated with AT-rich islands raised the possibility of predicting the genome-wide distribution of ORIs by localizing the position of the islands on the basis of their base composition. Therefore, we used the Artemis and EMBOSS software packages (see the Methods section) to serch the *S. pombe* genome sequence for regions between 0.5 kb and 1 kb long whose A+T content was higher than the boundaries described above and identified 384 that qualified as AT-rich islands. Their distribution along the three *S. pombe* chromosomes is represented in Figure 2, and their genomic localization is indicated in Table 3 of Supplementary Material. Their average genomic frequency was one every 33 kb and, with no exceptions, they mapped at intergenic regions. AT-rich islands were overrepresented between divergent transcription units [52% observed (O), 26.6% expected (E)] and underrepresented between colinear (37.7% O, 46.8% E) and convergent (10.3% O, 26.6% E) transcription units. These percentages excluded the 25 AT-rich islands clustered in the three centromeres. A similar or even more pronounced bias has previously been reported using a smaller number of cases (Gómez and Antequera, 1999). That was probably due to the relatively large size of the intergenic regions analyzed, which made it more likely that they would contain an AT-rich island. Divergent intergenic regions are longer than the average intergenic distance in *S. pombe* (Wood *et al.*, 2002) and this fact, perhaps in combination with the proximity of two promoters, could contribute to the overrepresentation of AT-rich islands in these regions.

AT-rich islands reliably predict the localization of genomic ORIs

To evaluate the reliability of AT-rich islands in predicting the localization of ORIs, we selected 20 of them at random (yellow triangles in Figure 2) and monitored their replication pattern by neutral two-dimensional gel electrophoresis (Brewer *et al.*, 1988; Huberman, 1993). Figure 3A shows that 18 out of the 20 islands tested (90%) co-localized with active ORIs, as evidenced by the presence of intermediates containing initiation bubbles. This predicts the existence of approximately 345 ORIs associated with AT-rich islands in the entire genome (in addition to those at the rRNA gene clusters). A spike of recombination intermediates is present in regions containing a replication origin (white arrow in Figure 3A) as we have previously described (Segurado et al. 2002) although the hybridization signal is very weak in some cases. This could probably be improved by using synchronous cultures (Segurado et al. 2002). It is also possible that a small proportion of ORIs, that could be active only under certain circumstances or at certain genomic localization, might not be associated with recombination intermediates.

A conspicuous feature of the distribution of AT-rich islands was their high frequency in the three centromeres, the subtelomeric regions of chromosomes I and II, and the mating type locus (Figure 2). Four active ORIs had been previously described in the K and L repeats of centromere II (Smith *et al.*, 1995) that turned out to be coincident with islands 2053, 2054, 2057 and 2058. We tested an additional island in centromere II (2055) and two more in the subtelomeric region of chromosome I (1002 and 1003) and in all three cases, they also co-localised with active ORIs (Figure 3A). Island 2068 is immediately adjacent to the *mat1* gene and coincides with the pARS756 sequence (Maundrell *et al.*, 1988) and with an active

genomic ORI (Dalgaard and Klar, 2001). The 20 kb region encompassing the *mat2* and *mat3* loci includes AT-rich islands 2069, 2070 and 2071 plus another one in a 12 kb region missing from the sequence available at the Sanger Centre but included in the NCBI database (labeled with an asterisk in Figure 2 and not included in Table 3). Altogether, the average density of AT-rich islands and ORIs in centromeric, subtelomeric and mating-type regions is about 4-fold higher than the genome average.

Since we had defined the AT-rich islands on the basis of a statistical average, we wondered how strict the boundary we had established was. In other words, how much lower the A+T content of an intergenic region could be relative to the lower limit we have used to define the islands and still act as an ORI. We addressed this point by analysing the replication pattern of 18 promoter-containing regions of comparable size to those shown in Figure 3A but with an A+T content in the 500 bp to 1 kb windows halfway between the intervals defined by ORI and non-ORI regions in Figure 1A. Analysis of 1 Mb of each chromosome showed that there is an average of 19 intergenic regions per Mb with this base composition, which predicts approximately a total of 240 in the entire genome. We tested 18 of these regions for replication and found 3 of them (16.6%) associated with active ORIs. Bubble arcs were not detected in the other 15, even in overexposed autoradiographs (Figure 3B and data not shown). The implication is that there could be approximately 40 ORIs in addition to those located at AT-rich islands, which represents approximately 10% of all ORIs. This also indicates that a small decrease in the A+T content relative to the criterion used to define AT-rich islands reduces dramatically (from 90% to

16.6%) the reliability in predicting ORIs. Regions with an even lower A+T content probably contain very few ORIs, if any. This is suggested by the fact that none of the 16 regions with an A+T content close to the intergenic genomic average, was associated with ORIs (Figure 1A, white circles) and by the localization of all the previously identified ORIs with AT-rich islands. Altogether, our results indicate that AT-rich islands are very reliable predictors of ORIs in *S. pombe* and that the proportion of ORIs not associated with them is likely to be small.

DISCUSSION

This work establishes AT-rich islands as discrete genomic regions that can account for the distinctive properties of the *S. pombe* ORIs. Recent studies have suggested that several ORC binding sites are collectively required for efficient ORI firing (Kong and DePamphilis, 2002; Takahashi *et al.*, 2003). Cooperation would require a minimal length of AT-rich DNA to attain the critical concentration of ORC complexes to trigger replication, which is consistent with the size of the AT-rich islands and with the 0.5-1 kb long size of *S. pombe* ORIs. Cooperativity and a certain degree of redundancy are also suggested by the fact that progressive shortening of several ORI-containing regions results in a gradual decline in replication efficiency, rather than in an all-or-none effect (Dubey *et al.*, 1994; Clyne and Kelly, 1995; Okuno *et al.*, 1999). Redundancy can also explain that removal of a 330 bp long region encompassing the replication initiation point in *ars1* in the chromosome diminishes, but does not prevent, ORI activity (Gómez and Antequera, 1999). One of the best characterized ORIs in *S. pombe* is *ars2004* (Okuno et al., 1997, 1999; Takahashi et al., 2003) where three specific regions have been shown to be collectively essential for

ORI activity in the chromosome (Takahashi et al., 2003). Region I includes a tract of poly-adenine 19 bp long and region III contains 11 repeats of the AAAAT sequence. These elements, however, are not present in many other *S. pombe* ORIs, suggesting that –although important for ars2004- they are not a general requirement. For example, SpOrc4 binds *in vitro* to several A+T-rich regions of *ars1*, none of which has a poly A tract longer than 6 bp (Lee et al., 2001). Also, only three non-contiguous adenine stretches 4 bp long are present in the sequence to which SpOrc4 binds with higher affinity *in vivo* (Kong and DePamphilis, 2002). The lack of conserved sequence elements between AT-rich islands suggests that the SpOrc4 protein can bind, via its AT-hooks, AT-rich sequences of different composition at different ORIs. Our data predict the existence of approximately 345 ORIs located at AT-rich islands and 40 additional ORIs at regions with a slightly lower A+T content (Figure 3). Although it is possible that there could be a few additional ORIs at other genomic regions, our estimate is very close to the recent prediction of approximately 400 ORIs in *S. cerevisiae* (Raghuraman *et al.*, 2001; Wyrick *et al.*, 2001). However, no correlation between ORIs and a distinctive base composition has been found in *S. cerevisiae* (Raghuraman *et al.*, 2001). The distribution of AT-rich islands in *S. pombe* is relatively homogeneous except at centromeres, subtelomeric regions and the mating type locus, where the average frequency of AT-rich islands is one every 8 kb. A high density of ORIs has also been found in many subtelomeric regions in *S. cerevisiae* (Wyrick *et al.*, 2001). This suggests a possible role for ORC (or for other components of the replication initiation machinery) in establishing or maintaining the chromatin organisation and transcriptional silence in these regions in *S. pombe*, as has been described in *S. cerevisiae* (Fox *et al.*, 1995; Palacios DeBeer *et al.*, 2003) and *Drosophila*

(Pak *et al.*, 1997). The chromoprotein Swi6 could be involved in this link given that it binds to silent chromatin in centromeres, telomeres and the mating-type locus and interacts directly with DNA polymerase α. In addition, temperature-sensitive mutants of DNA polymerase α show delocalization of Swi6 and defects in transcriptional silencing (Ahmed et al., 2001; Nakayama et al., 2001). A lower ratio in the bubble to Y arcs in ORI 1002 and 2055 at subtelomeric and centromeric regions (Figure 3A) relative to other ORIs could be due to interference, given the high density of predicted ORIs in these regions, as has been observed at the closely spaced ORIs *ars3002*, *ars3003* and *ars3004* (Dubey et al. 1994). Alternatively, a lower efficiency of these ORIs could correlate with a higher efficiency in the binding of ORC and MCM proteins, as it has been shown at the *HM* and telomeric regions in *S. cerevisiae* (Wyrick et al., 2001; Palacios DeBeer et al., 2003).

SPECULATION

S. pombe ORIs are similar to those of mammalian cells in terms of size, lack of consensus sequences, and in their preference to localize in intergenic regions close to promoters. A striking difference, however, is that many replication origins in mammals are associated with CG-rich islands (Delgado *et al.*, 1998), which are regions with a G+C content higher than the genome average. Sequence instability in these regions is suggested by their association with the most expandable loci containing trinucleotide repeats in the human genome (Brock *et al.*, 1999) and we have suggested that CG-rich islands might have originated from a bias in the errors made by DNA polymerases, or in their repair, during the replication initiation event (Antequera and Bird, 1999). On the other hand, the activation of *S. pombe* ORIs is

associated with a high level of mitotic recombination (Segurado *et al.*, 2002), which could make these regions prone to genetic instability (Strathern *et al.*, 1995). It is possible that replication initiation is associated with some kind of genetic instability that, along the evolution of some organisms, could have shifted the base composition of ORI regions in two alternative directions to generate either CG-rich or AT-rich islands.

METHODS

Base composition analysis

This study was performed taking as a reference the *S. pombe* genome sequence available at the Sanger Centre. The telomeric repeats and the 1.1 Mb rRNA gene clusters were not included in the analysis. We scanned the entire sequence with the base composition tool of the Artemis software package (http://www.sanger.ac.uk/Software/Artemis/) using windows of 500 bp to 1 kb and a step of 1 bp. Those whose A+T content was above the limits indicated in the text_were considered AT-rich islands. Their position in the genome is indicated in Table 3 of Supplementary Material. For analysis of specific regions (Figures 1B and 1C) the corresponding sequences were downloaded and scanned with the FREAK programme of the EMBOSS software package (http://www.emboss.org/) using a 500 bp window and a step of 100 bp.

Culture conditions and two-dimensional gel electrophoresis analysis

Cultures of *S. pombe* h⁻ 972 grown in rich medium were used for all the experiments. DNA from 500 ml of an exponentially growing culture (A_{595} = 0.8) was isolated and used for each gel as previously described (Segurado *et al.*, 2002). Replication intermediates were separated by two-dimensional neutral gel electrophoresis under conditions described by Brewer *et al.*, (1988) and Huberman (1993). Specific genomic regions were selected for two-dimensional gel electrophoresis depending on the availability of restriction sites that would give rise to a restriction fragment between 3 and 6 kb long, whis is the optimal size range required for this analyses. The position of restriction fragments and probes used for the 38 regions tested (Figure 3) are available upon request.

Acknowledgements

We are grateful to Andrés Aguilera, Josep Casadesús, María Gómez, Pablo Hernández and Mercedes Tamame for many suggestions and for helpful comments on the manuscript. We also thank Val Wood for the data on the distribution of intergenic regions. M. S. and A. de L. were supported by postgraduate fellowships from the Ministerio de Ciencia y Tecnología. This work was funded by grant BMC2002-03591 from the Ministerio de Ciencia y Tecnología.

REFERENCES

Ahmed, S., Saini, S., Arora, S. and Singh, J. (2001). Chromodomain protein Swi6-mediated role of DNA polymerase a in establishment of silencing in fission yeast. *J. Biol. Chem.*, **276**: 47814-47821

Antequera, F. and Bird, A. CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.* **9**, R661-R667 (1999)

Brewer, B. J., Sena, E. P. and Fangman, W. L. (1988). Analysis of replication intermediates by two-dimensional agarose gel electrophoresis. In Cancer Cells, 6. *Eukaryotic DNA Replication*. Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press, pp. 229-234

Broach. J. R., Li, Y.Y., Feldman, J., Jayaram, M., Abraham, J., Nasmyth, K. A. and Hicks, J. B. (1983). Localization and sequence analysis of yeast origins of DNA replication. *Cold Spring Harbor Symp. Quant. Biol.* **47**: 1165-1173

Brock, G. J., Anderson, N. H. and Monckton, D. G. Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG. *Hum. Mol. Genet.* **8**, 1061-1067 (1999)

Chuang, R. Y. and Kelly, T. J. (1999). The fission yeast homologue of Orc4p binds to replication origin DNA via multiple AT-hooks. *Proc. Natl. Acad. Sci. USA*, **96**: 2656-2661

Clyne, R. K. and Kelly, T. J. (1995). Genetic analysis of an ARS element from the fission yeast *Schizosaccharomyces pombe. EMBO J.*, **14**: 6348-6357

Dalgaard, J. Z. and Klar, A. J. (2001). A DNA replication-arrest site RTS1 regulates imprinting by determining the direction of replication at mat1 in *S. pombe*. *Genes Devel.*, **15**: 2060-2068

Delgado, S., Gómez, M., Bird, A. & Antequera, F. Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J.* **17**, 2426-2435 (1998)

Dubey, D. D., Zhu, J., Carlson, D. L. Sharma, K. and Huberman, J. A. (1994). Three ARS elements contribute to the ura4 replication origin region in the fission yeast *Schizosaccharomyces pombe*. *EMBO J.*, **13**: 3638-3647

Dubey, D. D., Kim, S. M., Todorov, I. T. and Huberman, J. A. (1996). Large, complex modular structure of a fission yeast DNA replication origin. *Curr. Biol.,* **6**: 467-473

Fox, C. A., Loo, S., Dillin, A. and Rine, J. (1995). The origin recognition complex has essential functions in transcriptional silencing and chromosomal replication. *Genes Devel.*, **9**: 911-924

Gómez, M. and Antequera, F. (1999). Organization of DNA replication origins in the fission yeast genome. *EMBO J.*, **18**: 5683-5690

Huberman, J. A. (1993). Analysis of DNA replication origins and directions by two-dimensional gel electrophoresis. In P. Fantes and R. Brooks (eds.), *The Cell Cycle. A Practical Approach*. Oxford University Press, Oxford, UK, pp. 213-234

Kong, D. and DePamphilis, M. L. (2002). Site-specific ORC binding, pre-replication complex assembly and DNA synthesis at *Schizosaccharomyces pombe* replication origins. *EMBO J.*, **21**: 5567-5576

Lee, J. K., Moon, K. Y., Jiang, Y. and Hurwitz, J. (2001). The *Schizosaccharomyces pombe* origin recognition complex interacts with multiple AT-rich regions of the replication origin DNA by means of the AT-hook domains of the spOrc4 protein. *Proc. Natl. Acad. Sci. USA*, **98**: 13589-13594

Maundrell, K., Hutchinson, A. and Shall. S. (1988). Sequence analysis of ARS elements in fission yeast. *EMBO J.*, **7**: 2203-2209

Nakayama, J., Allshire, R. C., Klar, A. J. and Grewal. S. I. A role for DNA polymerase α in epigenetic control of transcriptional silencing in fission yeast. *EMBO J.*, **20**: 2857-2866 (2001)

Newlon, C. S. and Theis, J. F. (1993). The structure and function of yeast ARS elements. *Curr. Op. Gen. Devel.* **3**: 752-758

Okuno, Y., Okazaki, T. and Masukata, H. (1997). Identification of a predominant replication origin in fission yeast. *Nucleic Acids Res.*, **25**: 530-536

Okuno, Y., Satoh, H., Sekiguchi, M. and Masukata, H. (1999). Clustered adenine/thymine stretches are essential for function of a fission yeast replication origin. *Mol. Cell. Biol.*, **19**: 6699-6709

Pak, D. T., Pflumm, M., Chesnokov, I., Huang, D. W., Kellum, R., Marr, J., Romanowski, P. and Botchan, M. R. (1997). Association of the origin recognition complex with heterochromatin and HP1 in higher eukaryotes. *Cell*, **91**: 311-323

Palacios DeBeer, M. A., Müller, U. and Fox, C. A. (2003). Differential DNA affinity specifies roles for the origin recognition complex in budding yeast heterochromatin. *Genes Devel.* **17**: 1817-1822

Raghuraman, M. K., Winzeler, E. A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D. J., Davis, R. W., Brewer, B. and Fangman, W. L. (2001). Replication dynamics of the yeast genome. *Science*, **294**: 115-121

Reeves, R. and Beckerbauer, L. (2001). HMGI/Y proteins: flexible regulators of transcription and chromatin structure. *Biochim. Biophys. Acta*, **1519**: 13-29

Sanchez, A., Kim, S. M. & Huberman, J. Ribosomal DNA replication in the fission yeast *Schizosaccharomyces pombe*. *Exp. Cell Res.* **238**, 220-230 (1998)

Sanchez, A., Kim, S. M. and Huberman, J. (1998). Ribosomal DNA replication in the fission yeast *Schizosaccharomyces pombe. Exp. Cell Res.* **238**: 220-230

Segurado, M., Gómez, M. and Antequera, F. (2002). Increased recombination intermediates and homologous integration hot spots at DNA replication origins. *Mol. Cell*, **10**: 907-916

Smith, J. G., Caddle, M. S., Bulboaca, G. H., Wohlgemuth, J. G., Baum, M., Clarke, L. and Calos, M. P. (1995). Replication of centromere II of *Schizosaccharomyces pombe. Mol. Cell. Biol.*, **15**: 5165-5172

Strathern, J. N., Shafer, B. K. & McGill, C. B. DNA synthesis errors associated with double-strand-break repair. *Genetics*, **140**: 965-972 (1995)

Takahashi, T., Ohara, E., Nishitani, H. and Masukata, H. (2003). Multiple ORC binding sites are required for efficient MCM loading and origin firing in fission yeast. *EMBO J.*, **22**: 964-974

Theis, J. F. and Newlon, C. S. (1997). The ARS309 chromosomal replicator *of Saccharomyces cerevisiae* depends on an exceptional ARS consensus sequence. *PNAS*, **94**: 10786-10791

Theis, J. F. and Newlon, C. S. (2001). Two compound replication origins in *Saccharomyces cerevisiae* contain redundant origin recognition complex binding sites. *Mol. Cell. Biol.*, *21*: 2790-2801

Wood *et al.*, (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**: 871-880

Wyrick, J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P. and Aparicio, O. M. (2001). Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: High-resolution mapping of replication origins. *Science*, **294**: 2357-2360

LEGENDS TO FIGURES

Figure 1. Identification of AT-rich islands at origins of replication in *S. pombe.*

A, The highest A+T content of 16 ORI and 16 non-ORI regions was determined using sliding windows of 500 to 1000 bp and a step of 1 bp. Their average and standard deviation is indicated. Values for ORI-containing regions (black circles) are: 500 bp: 76.9 % +/- 1.5; 600 bp: 76 % +/- 1.3; 700 bp: 75.5 +/- 1.4; 800 bp: 74.6 +/- 1.2; 900 bp: 73.9 +/- 1.4 and 1000 bp: 73.3 +/- 1.2. Values for non-ORI regions (white circles) are: 500 bp: 70.6 % +/- 2.4; 600 bp: 69.8 % +/- 2.3; 700 bp: 69.2 +/- 2.0; 800 bp: 68.8 +/- 1.9; 900 bp: 68.8 +/- 2.0 and 1000 bp: 68.4 +/- 1.7. Dashed lines indicate the total genomic and intergenic A+T average content (64% and 70%, respectively).

B, A+T content across 30 kb of cosmids SPAC1296 and SPBC32F12 with ORI 12 and ORI tug1, measured with a 500 bp window and a step of 100 bp. Black or white rectangles represent genes transcribed towards the left or the right, respectively. Brackets indicate the position of restriction fragments containing the ORIs. Dashed lines indicate the intergenic average A+T content. Scale bar represents 5 kb.

C, A+T content across 6 kb long regions containing *ars3003*, *ars3002*, *ars2004*, *ars1* and AT-rich island 1003, measured as in B. Black bars indicate the shortest DNA fragment capable of maintaining full ARS activity in each case. RIP indicates the replication initiation point in *ars1* and *ars2004*. Rectangles represent genes and arrows indicate the direction of transcription. Scale bar represents 1 kb.

Figure 2. Genome-wide distribution of AT-rich islands.

Vertical bars indicate the position of the AT-rich islands across the *S. pombe* chromosomes numbered using a four-digit code starting with 1001, 2001 and 3001 for chromosomes I, II and III, respectively. The 16 ORIs used to define the properties of AT-rich islands (Figure 1A) are indicated by red triangles. The 14 ORIs

and ARS elements described by other authors (green triangles) and the 20 islands tested for replication in Figure 3 (yellow triangles) are also indicated. Only one replication origin at the rRNA cluster is shown (*ars3001*), which coincides with AT-rich island 3001. Centromeres are represented by a black box, and the AT-rich islands included in them are labelled red. The *mat* locus in chromosome II is represented by a red box. The asterisk indicates the position of an AT-rich island in a region missing in the Sanger Centre sequence.

Figure 3. AT-rich islands reliably predict the localization of genomic ORIs.
A, Twenty AT-rich islands (yellow triangles in Figure 2) were tested for replication by two-dimensional gel electrophoresis. Their identification number in the chromosomes is indicated. Intermediates containing replication bubbles are shown by arrows. A white arrow indicates recombination intermediates.
B, Replication analysis of 10 genomic regions (1-10) with an A+T content between the intervals defined by the ORI- and non-ORI-containing regions in Figure 1A.

Legend for Table 1 of Supplementary Material
Columms indicate the name of the 16 ORI-containing and –non containing regions tested for replication; the highest A+T content of every window size and the reference where the two-dimensional gel analysis was published. The average and standard deviation for every window is also shown. References: a: Gómez *et al.*, (1999). b: Segurado *et al.*, (2002). The position in the chromosomes of the ORI-containing regions is given in Table 3.

Legend for Table 2 of Supplementary Material

Columms show the name of the regions tested for replication by two-dimensional gel analysis or by plasmid ARS assay (pARS); the highest A+T content of every window size and the reference where the data were published. The average and standard deviation for every window is also shown. The position in the chromosomes of these regions is given in Table 3.

Legend for Table 3 of Supplementary Material

Columms indicate the four digit code assigned to every AT-rich island; the position of the intergenic regions where they are located in the sequence available at the Sanger Centre; the name of the flanking ORFs or genes; their divergent (< >), colinear (>> or <<) or convergent (> <) direction of transcription relative to one another and the base composition of the 500 bp with the highest A+T content in the region indicated in the second columm. Coordinates at centromeres indicate the position of the nucleotide in the middle of the 500 bp window with highest A+T content. The last two columms indicate the ORI activity of the 20 islands tested in this work (Figure 2, yellow triangles) the 16 genomic ORIs used to define the properties of the islands (Figure 2, red triangles) and the 14 genomic ORIs and ARS elements previously reported by other authors (Figure 2, green triangles).