A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression

Barak A. Cohen*, Robi D. Mitra*, Jason D. Hughes & George M. Church

*These authors contributed equally to this work.

Chromosome correlation maps display correlations between the expression patterns of genes on the same chromosome. Using these maps, we show here that adjacent pairs of genes, as well as nearby non-adjacent pairs of genes, show correlated expression independent of their orientation. We present specific examples of adjacent pairs with highly correlated expression patterns, in which the promoter of only one of the two genes contains an upstream activating sequence (UAS) known to be associated with that expression pattern. Finally, we show that genes with similar functions tend to occur in adjacent positions along the chromosomes. Our results suggest that, in certain chromosomal expression domains, an UAS can affect the transcription of genes that are not immediately downstream from it.

As a source of expression data, we used a data set in which the expression levels of all ORFs in the *Saccharomyces cerevisiae* genome were measured over the course of two mitotic cell cycles. This data set was chosen because ORFs with periodic expression profiles were often found in adjacent positions on the chromosome, suggesting the presence of position effects[1]. Chromosome correlation maps allow visualization of coexpressed genes along the chromosomes of yeast. A representative map (Fig. 1*a*) is shown for a portion of the left arm of chromosome XIII. Adjacent groups of correlated genes are depicted as blocks of green squares centred on the diagonal.

Groups of correlated adjacent genes appear throughout the genome. To determine whether this 'regional coexpression' was statistically significant, we compared the observed number of adjacent pairs with correlation coefficients greater than 0.7 (r>0.7) with the expected number derived from a control set of non-adjacent genes. This analysis showed that a substantial number of adjacent pairs have correlated expression patterns (Table 1). Correlated triplets, but not quadruplets, were also found to occur more often than expected by chance (Table 1). We obtained similar results using expression data collected during sporulation[2] and in response to the mating pheromone α-factor[3] (Table 1).

The correlation of an adjacent pair in one data set is not predictive of its correlation in other data sets. For example r=0.086 between the distributions of correlation coefficients for adjacent pairs in the cell cycle and pheromone-response data sets. Adjacent pairs that are highly correlated (r>0.7) in one condition, however, are more likely to be highly correlated in other conditions. For example, there were twofold more highly correlated adjacent pairs found in the intersection of cell cycle and pheromone data sets than would be expected (P=10^{-7}). Similar results were obtained when other data sets were compared.

We also found two large groups of correlated adjacent genes in the cell-cycle data. The first spans 26 kb on chromosome VI (Fig. 1*b*) and includes YFL061W, *SNO3*, *SNZ3*, *THI5*, *AAD6*, YFL056C, *AGP3*, YFL054C, *DAK2*, YFL052W, YFL051C and *ALR2*. The second spans 20 kb on chromosome XVI (Fig. 1*c*) and includes

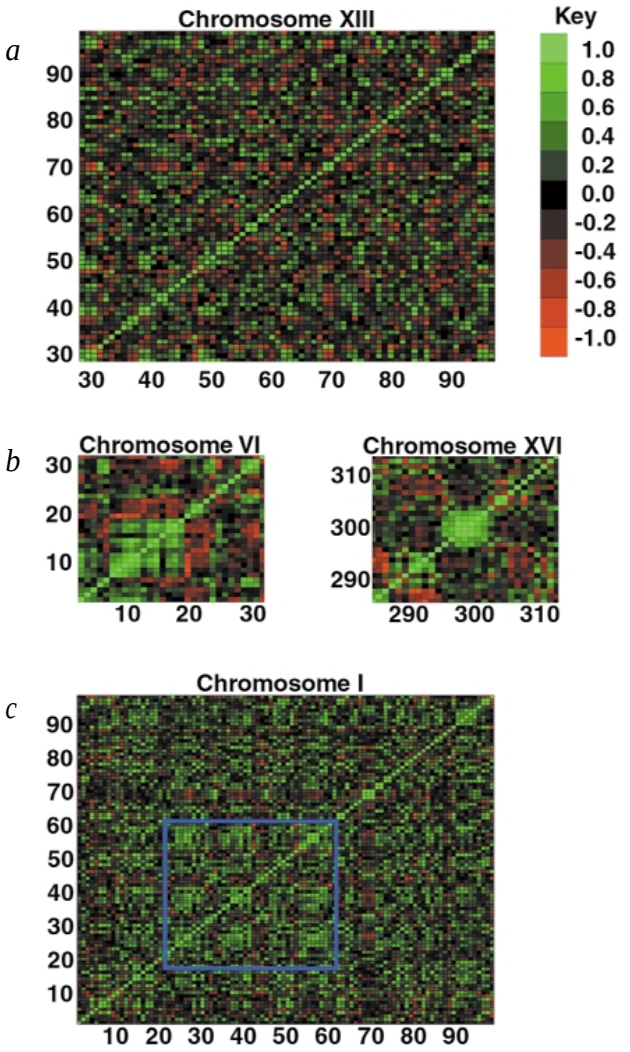

Fig. 1 Correlation maps for yeast chromosomes. In this analysis, we included every ORF in the genome (excluding overlapping ORFs), not just those represented in the filtered list. The numbers along the side of the matrix represent the ORFs along the chromosome. The squares are coloured green for positive correlation and red for anti-correlation. The intensity of the colour at each position represents the degree of correlation or anti-correlation. ***a***, Correlation map for ORFs on the left arm of chromosome XIII between positions 58,939 bp and 212,515 bp. ***b***, Correlation map for ORFs on the left arm of chromosome VI between positions 53 bp and 106,957 bp. The coexpressed group of adjacent ORFs extends from YFL061W at position 9,545 bp to YFL050C at position 35,848 bp. ***c***, Correlation map for ORFs on the right arm of chromosome XVI between positions 592,327 bp and 657,339 bp. The coexpressed group of adjacent ORFs extends from YPR027C at position 620,420 bp to YPR034W at position 640,953 bp. ***d***, Correlation map for all of chromosome I. The box shows a region of the chromosome in which regularly spaced blocks of correlated genes are found.

Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to G.M.C. (e-mail: church@arep.med.harvard.edu).

Table 1 • Analysis of coexpressed groups of adjacent ORFs

Group size	Data set	Observed	Expected	*P* value
pair				
	cell cycle	394	98	2.3×10^{-114}
	sporulation	647	323	2.7×10^{-60}
	pheromone	489	257	9.0×10^{-39}
triplet				
	cell cycle	50	30	5.5×10^{-4}
	sporulation	103	80	6.7×10^{-3}
	pheromone	63	43	2.1×10^{-3}
quadruplet				
	cell cycle	7	6.5	0.47
	sporulation	16	17	0.65
	pheromone	11	8.5	0.23

For each size group, the expected number of coexpressed groups is shown along with the observed number of coexpressed groups and the *P* value for obtaining such a result by chance calculated using the cumulative binomial distribution.

YPR027C, *YIP2*, *APL4*, *CSR2*, YPR031W, *SRO7*, *HTS1* and *ARP7*. We searched for common promoter elements in these groups using AlignACE (ref. 4) and by looking for known UASs catalogued in the *S. cerevisiae* Protein Database[5]. We also looked for the presence of sequence motifs previously shown to be associated with the expression of these genes[6]. No sites were present in most of the intergenic regions within these groups. This suggests that the observed regional coexpression was not due to the presence of similar UASs in the intergenic regions between the ORFs in these groups.

In addition to adjacent correlated genes, correlation maps often reveal regularly spaced groups of correlated genes along the chromosomes that may be indicative of higher-order chromosome structure (Fig. 1*d*).

We next examined the possibility that evolution takes advantage of regional coexpression by keeping genes with similar functions in adjacent positions. Using the Munich Information Center for Protein Sequences (MIPS) classifications[7], we determined how often adjacent genes fall into the same functional category. Of the 2,081 adjacent pairs examined, 387 fell into the same functional category, significantly more than would be expected by chance ($P=10^{-8}$).

There are several examples in yeast of co-regulated genes that are divergently transcribed from the same intergenic region[8–11]. To determine whether the increased correlation of adjacent pairs was due to divergently transcribed promoters, we compared the distributions of correlation coefficients for divergent, convergent and tandem pairs of adjacent genes with a control set of randomly picked, non-adjacent pairs (Fig. 2*a–c*). There was a significant difference ($\chi^2=357$, d.f.=9, $P=10^{-71}$) between the distributions for divergent and random pairs. There was also a significant difference between the distributions of convergent and random pairs ($\chi^2=133$, d.f.=9, $P=10^{-24}$), and tandem and random pairs ($\chi^2=221$, d.f.=9, $P=10^{-42}$). We found more divergent, convergent and tandem pairs with correlation coefficients above 0.7 than randomly picked, non-adjacent pairs (Fig. 2*d*). Similar results were observed using the sporulation and pheromone-response data sets (Fig. 2*e–f*). These results are not a consequence of recent duplications in the yeast genome or crosshybridization of UTRs that overlap adjacent ORFs (see Methods). These results demonstrate that adjacent genes, in any orientation, are more likely to be coexpressed than non-adjacent genes.

Although adjacent genes in all orientations tended to be coexpressed, divergent pairs showed the greatest deviation from the control set of non-adjacent pairs (Fig. 2*a*,*d*). One explanation for this is that divergent genes share an upstream intergenic region, whereas convergent and tandem pairs do not. Alternatively, the increased correlation of divergent pairs may result from the fact that the promoters of divergent pairs tend to be closer together than the promoters of convergent and tandem pairs (Fig. 3*a*). To distinguish between these two possibilities, we compared the distributions of correlation coefficients for divergent and tandem pairs whose start sites are equally far apart (Fig. 3*b*). We found no significant difference ($\chi^2=5.8$, d.f.=9, $P=0.72$) between the distributions. There was also no significant difference between the distributions for convergent and tandem pairs whose start sites are equally far apart ($\chi^2=9.0$, d.f.=9, $P=0.44$). These results suggest that the distance between adjacent genes, apart from their orientation, is important in determining their coexpression.

The occurrence of high coexpression declined with the distance between adjacent ORFs (Fig. 4), but remained above the occurrence of high coexpression observed for pairs picked at random, even at the largest distances examined. Therefore, the distance between ORFs is not in itself predictive of increased correlation. The closer together two ORFs occur, however, the more likely it is that they will be coexpressed.

One model that might explain regional coexpression is that each ORF in a correlated pair has a similar UAS in its upstream intergenic region. Contrary to this model, we found several examples of convergent and tandem pairs with highly correlated expression patterns in which the promoter of only one of the

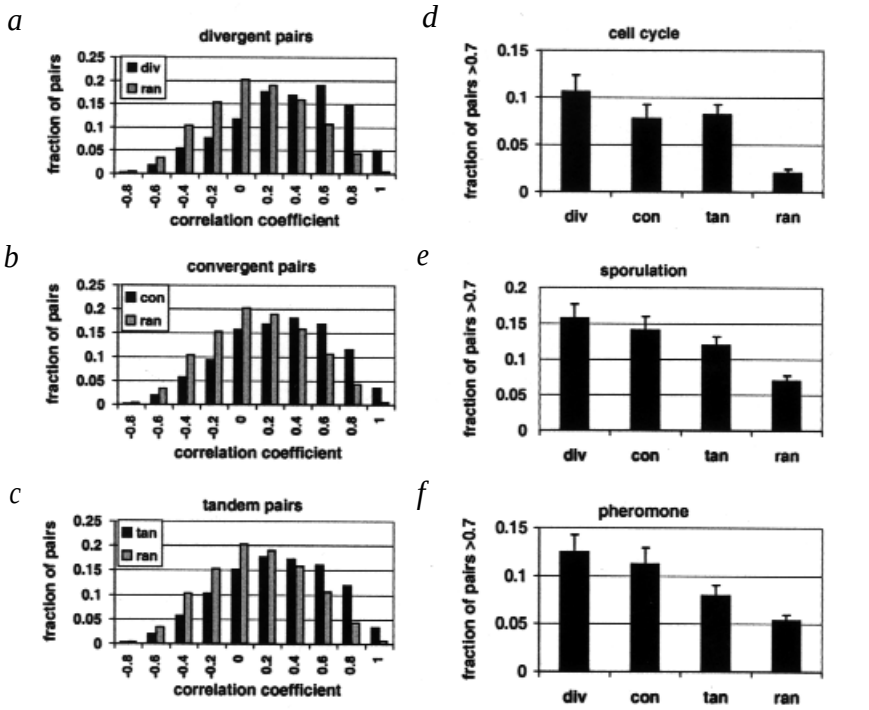

Fig. 2 Histograms of the distributions of correlation coefficients describing the expression of divergent (***a***), convergent (***b***) and tandem (***c***) adjacent pairs in the cell cycle are shown along with the distribution of correlation coefficients for a control set of randomly picked, non-adjacent pairs of ORFs. The fraction of adjacent pairs with correlation coefficients above 0.7 in each orientation is plotted for the cell cycle (***d***), sporulation (***e***) and pheromone-response (***f***) data sets.

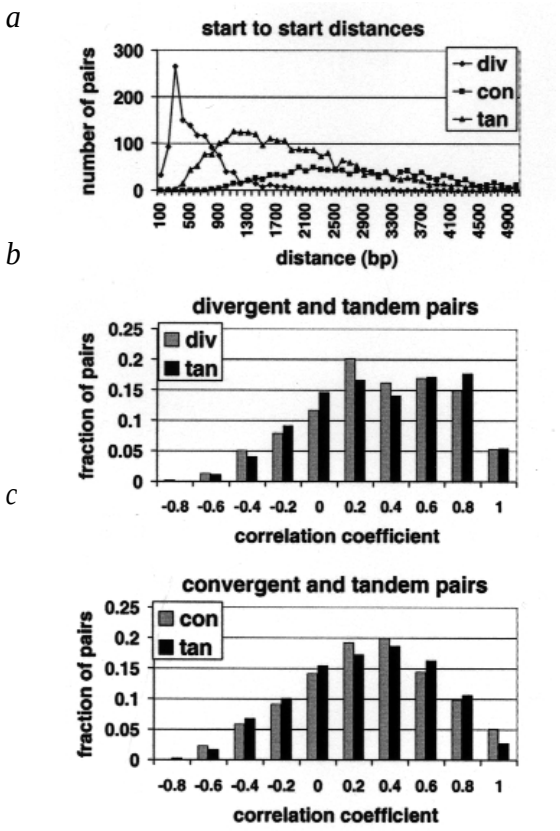

Fig. 3 Comparison of adjacent pairs with different orientations. ***a***, Histograms showing the distances between the predicted start sites for divergent, convergent and tandem pairs of adjacent ORFs. ***b***, Histograms showing the distributions of correlation coefficients for divergent and tandem pairs of ORFs whose predicted start sites are 400–1,000 bp apart. ***c***, Histograms showing the distributions of correlation coefficients for convergent and tandem pairs of ORFs whose predicted start sites are 1,600–3,600 bp apart.

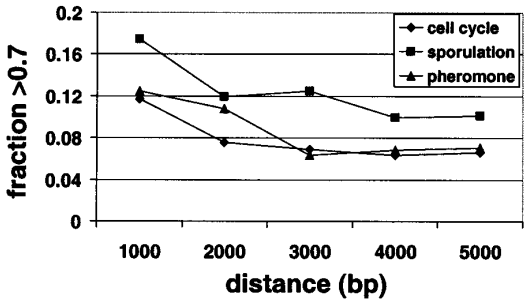

Fig. 4 Relationship of the distance between the predicted start sites of adjacent ORFs and their correlation. Each point represents the fraction of pairs of adjacent ORFs that have a correlation coefficient above 0.7 and whose predicted start sites are at a given distance apart. Background levels of correlation were determined from Fig. 2, and taken to be x for cell cycle, y for sporulation and z for pheromone response.

two genes contained a UAS known to be associated with that expression pattern (Table 2). For example, YPL162C and *SVS1* are an adjacent pair of tandem genes with highly correlated expression. The expression of both genes peaks in the G1 phase of the cell cycle, a pattern that has been shown to be associated with the SCB Box motif[6,12,13]. Only *SVS1*, however, contains this motif in its upstream intergenic region. This and other examples demonstrate that regional coexpression cannot be explained by the presence of similar UASs in the promoters of adjacent genes.

UASs may be capable of long-range interactions when they occur within areas of open chromatin structure[14]. Visualizing chromosomes as correlation maps will help uncover these areas in which genes are regionally coexpressed. Comparing chromosome correlation maps that use different expression data sets will reveal how expression domains change in different conditions and mutants.

Methods

Data processing. Our analysis was performed using whole-genome mRNA expression data generated for the mitotic cell cycle (http://genomics.stanford.edu/yeast/cellcycle.html). Briefly, transcript levels were quantified during the cell cycle of *S. cerevisiae* by synchronizing *cdc28-13* cells and collecting mRNA from cells at 10-min intervals over the course of 2 cell cycles. The abundance of each mRNA species in the yeast genome was quantitated by hybridization to oligonucleotide microarrays (Affymetrix). We also analysed expression data generated on spotted cDNA microarrays, including a study of the changes in gene expression during sporulation (http://cmgm.stanford.edu/pbrown/sporulation) and in response to mating pheromone (http://www.rii.com).

Before proceeding with our analysis, we excluded several data points. We deleted the 90- and 100-min time points from the cell-cycle data set, as the mRNA from these time points was not efficiently labelled[6]. We also ignored ORFs that displayed a mean intensity of less than 20 units in the cell-cycle data set because experimental noise obscured meaningful quantitation of transcript abundance below this cutoff.

As we began to analyse the data, we discovered that there are a number of annotated ORFs whose sequences physically overlap in the genome. Overlapping ORFs are problematic because cDNA from both ORFs may hybridize to the probes corresponding to a single ORF, resulting in an artificially high correlation between overlapping ORFs. To eliminate this artefact, we identified all pairs of overlapping ORFs in the genome and removed the smaller of the two ORFs from the data set.

The filtered data sets therefore consisted of the non-overlapping ORFs (with average intensities above 20 units in the cell-cycle data set). The final lists contained 5,531 ORFs for the cell cycle, 5,622 ORFs for the sporulation data and 5,797 ORFs for the pheromone-response data. All experiments use these filtered lists unless otherwise stated.

For experiments analysing adjacent ORFs, we considered two ORFs to be adjacent if they occurred on the same chromosome, and if there were no other ORFs, Ty elements, long terminal repeats, centromeres, tRNAs,

Table 2 • Convergent and tandem pairs of genes showing regional coexepression and their associated motifs[6]

				Upstream motifs associated with expression pattern	
ORF 1	ORF 2	Orientation	Correlation	ORF 1	ORF 2
YRA1	RPP2B	tan	0.96	none	M1a (×2)
YPL162C	SVS1	tan	0.96	none	SCB Box (×3)
RPL35A	ARF1	tan	0.95	Rap1 Site (×3), M1a	none
YDL025C	YDL027C	tan	0.82	STRE (×2)	none
YJL163C	SRA3	tan	0.88	none	STRE (×4)
RPP1A	RPL13A	con	0.9	none	Rap1 site (×2)
YAR003W	RFA1	con	0.83	none	MCB box (×4)

rRNAs or snRNAs between them. Triplets and quadruplets were defined as series of adjacent pairs.

Coexpression of groups of adjacent genes. Adjacent pairs were considered to be coexpressed if they had a correlation coefficient greater than 0.7. Triplets and quadruplets were considered to be coexpressed if their component adjacent pairs all had correlation coefficients greater than 0.7.

To determine if the observed number of correlated adjacent pairs was significant, we used the cumulative binomial distribution, given by the formula,

$$P(n \geq n_0) = \sum_{n=n_0}^{N} p^n (1-p)^{N-n} \left[\frac{N!}{n!(N-n)!} \right]$$

where N is the total number of adjacent pairs sampled, n_0 is the observed number of correlated adjacent pairs, and p is the observed probability of two randomly picked non-adjacent genes having a correlation above the cutoff.

To determine the significance of correlated triplets, we used the same formula, except where N is the total number of triplets sampled, n_0 is the observed number of triplets with a correlation above the cutoff, and $P = (p_{pair})^2$, where p_{pair} is the observed probability of adjacent pairs having a correlation above the cutoff.

To determine the significance of correlated quadruplets, we used the same formula, except where N is the total number of quadruplets sampled, n_0 is the observed number of quadruplets with a correlation above the cutoff, and $P = p_{pair} (p_{pair-1})^2$ where p_{pair-1} is the observed probability of an adjacent pair having a correlation above the cutoff given that the previous adjacent pair has a correlation above the cutoff.

Comparing the correlation of divergent, convergent and tandem ORF pairs with randomly selected non-adjacent pairs of genes. Adjacent pairs of ORFs present in the filtered data sets were grouped according to their orientations. Control pairs were derived by randomly picking two non-adjacent ORFs and grouping them together as a pair. The correlation coefficients for all pairs were computed, and a histogram showing the fraction of pairs in each correlation bin was plotted for each group. The significance of the differences between the control group and the different classes of adjacent pairs was determined using the χ^2 test[15]. For the cell-cycle analysis 1,161 divergent, 1,239 convergent, 2,284 tandem and 4,542 random pairs were included in the analysis. The corresponding numbers were 1,225, 1,257, 2,358 and 4,840 for the sporulation data, and 1,318, 1,292, 2,479 and 4,963 for the pheromone response data.

Over-representation of adjacent pairs in MIPS functional categories. We scored adjacent pairs of ORFs as having similar function if they were in the same MIPS functional group. If a MIPS category had more than 537 genes, and contained subgroups, we used only its subgroups. The 'unclassified' MIPS category was excluded. This left a total of 105 MIPS categories for the analysis. Of 2,081 adjacent pairs, 387 were found in the same functional category. Of 27,923 randomly picked pairs, 3,900 were found in the same functional category. Using the cumulative binomial distribution, we computed the probability of the adjacent pairs showing this level of functional similarity by chance to be 4.9×10^{-8}.

Analysis of adjacent pairs with respect to homologous sequences and overlapping UTRs. Less than 5% of the most highly correlated adjacent pairs consisted of homologous sequences that could be identified using the BLAST algorithm[16]. Also, when 1,331 poly(A)-primed ESTs representing 844 different genes[17] were mapped to the yeast genome using BLAST, no 3′ UTRs were found that overlapped an adjacent ORF. This result suggests that overlapping transcripts are rare in yeast and makes it unlikely that regional coexpression can be explained by adjacent genes with overlapping transcripts. Moreover, we continued to observe high correlation between adjacent genes when the most closely packed pairs of genes (and therefore the most likely to overlap) were not considered (http://arep.med.harvard.edu/adjacent/supplement.html).

Acknowledgements

We thank J. Aach, W. Rindone, S. Tavazoie, J. Graber, K. Struhl and F. Winston for advice, suggestions and data files; and P. Sudarsanam, A. Dudley, T. Pilpel, A. Derti, P. Estep, M. Steffen, V. Badarinarayana, T. Wu and M. Bulyk for discussions and critical readings of the manuscript. B.A.C. was supported by a postdoctoral fellowship from the American Cancer Society (PF-98-159-01-MBC). This work was supported by the US Department of Energy (DE-FG02-87-ER60565), the Office of Naval Research and DARPA (N00014-97-1-0865), the Lipper Foundation and Hoechst Marion Roussel.

Received 10 March; accepted 10 August 2000.

1. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
2. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
3. Roberts, C.J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
4. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.* **296**, 1205–1214 (2000).
5. Zhu, J. & Zhang, M.Q. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15**, 607–611 (1999).
6. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
7. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
8. Kraakman, L.S., Mager, W.H., Maurer, K.T., Nieuwint, R.T. & Planta, R.J. The divergently transcribed genes encoding yeast ribosomal proteins L46 and S24 are activated by shared RPG-boxes. *Nucleic Acids Res.* **17**, 9693–9706 (1989).
9. Nakao, J., Miyanohara, A., Toh-e, A. & Matsubsara, K. Saccharomyces cerevisiae PHO5 promoter region: location and function of the upstream activation site. *Mol. Cell. Biol.* **6**, 2613–2623 (1986).
10. Osley, M.A., Gould, J., Kim, S., Kane, M.Y. & Hereford, L. Identification of sequences in a yeast histone promoter involved in periodic transcription. *Cell* **45**, 537–544 (1986).
11. West, R.W. Jr, Yocum, R.R. & Ptashne, M. Saccharomyces cerevisiae GAL1-GAL10 divergent promoter region: location and function of the upstream activating sequence UASG. *Mol. Cell. Biol.* **4**, 2467–2478 (1984).
12. Nasmyth, K.A. repetitive DNA sequence that confers cell-cycle START (CDC28)-dependent transcription of the HO gene in yeast. *Cell* **42**, 225–235 (1985).
13. Koch, C., Schleiffer, A., Ammerer, G. & Nasmyth, K. Switching transcription on and off during the yeast cell cycle: Cln/Cdc28 kinases activate bound transcription factor SBF (Swi4/Swi6) at start, whereas Clb/Cdc28 kinases displace it from the promoter in G2. *Genes Dev.* **10**, 129–141 (1996).
14. Felsenfeld, G. Chromatin unfolds. *Cell* **86**, 13–19 (1996).
15. Moore, D.S. & McCabe, G.P. *Introduction to the Practice of Statistics* (W.H. Freeman, New York, 1998).
16. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
17. Graber, J.H., Cantor, C.R., Mohr, S.C. & Smith, T.F. In silico detection of control signals: mRNA 3′-end-processing sequences in diverse species. *Proc. Natl Acad. Sci. USA* **96**, 14055–14060 (1999).