Yeast
Yeast 2006; **23**: 689–699.
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI:** 10.1002/yea.1390

Research Article

A genomic study of the inter-ORF distances in *Saccharomyces cerevisiae*

Vicent Pelechano, José García-Martínez and José E. Pérez-Ortín*
Departamento de Bioquímica y Biología Molecular, Facultad de Ciencias Biológicas, Universitat de València, C/Dr. Moliner 50, E46100 Burjassot, Spain

**Correspondence to: José E. Pérez-Ortín, Departamento de Bioquímica y Biología Molecular, Facultad de Ciencias Biológicas, Universitat de València, C/Dr. Moliner 50, E46100 Burjassot, Spain. E-mail: jose.e.perez@uv.es*

Received: 1 March 2006
Accepted: 23 May 2006

Abstract

The genome of eukaryotic microbes is usually quite compacted. The yeast *Saccharomyces cerevisiae* is one of the best-known examples. Open reading frames (ORFs) occupy about 75% of the total DNA sequence. The existence of other, non-protein coding genes and other genetic elements leaves very little space for gene promoters and terminators. We have performed an *in silico* study of inter-ORF distances that shows that there is a minimum distance between two adjacent ORFs that depends on the relative orientation between them. Our analyses suggest that different kinds of promoters and terminators exist with regard to their length and ability to overlap each other. The experimental testing of some putative exceptions to the minimum length model in tandemly orientated ORF pairs suggests that, in those cases, defects in promoter or terminator functionality exist that provoke transcription of polycistronic mRNAs. Copyright © 2006 John Wiley & Sons, Ltd.

Keywords: transcription interference; ORF; *Saccharomyces cerevisiae*; polycistronic; co-regulation

Introduction

The genome of the yeast *Saccharomyces cerevisiae* is extremely compact. More than 72% of the sequence is organized in open reading frames (ORFs). Thus, there is very little space for non-coding signatures, such as promoters, terminators and other elements. The first calculations made by Dujon (1996) showed that, in yeast, ORFs were organized randomly with regard to their respective orientation. About half of the arrangements were tandemly orientated, 25% convergent and 25% divergent ORFs. Assuming that genes are non-overlapping units, the sizes of the promoters and terminators could be calculated from the distances between divergent and convergent ORFs, respectively (Dujon, 1996). This pioneering work determined that the sizes of the average yeast promoter and terminator were of 309 bp and 163 bp, respectively. Strictly speaking, these numbers correspond to the sizes of the 5′ and 3′ sequence flanking of the ORFs. This estimation does not make an important error for the terminator, because most of the signals for 3′ mRNA generation are within the transcribed region (Van Helden *et al*., 2000). However, the identification of a 5′ region of an ORF as a gene promoter is not correct because there is a space between the first nucleotide of the mRNA and the first nucleotide of the ATG: 15–75 bp (Zhang and Dietrich, 2005). Thus, the promoter size would be, on average, about 30 bp shorter (279 bp). These calculations, however, use a very simplistic approach because the distribution of inter-ORF distances is not a Normal (Gaussian) distribution.

Intergenic sequences have, on the other hand, compositional differences with regard to the adjacent ORFs (Dujon, 1996) and with regard to the ORF orientation (Marín *et al*., 2004), which suggests a functional specialization regarding the ORF orientation. The separation between ORFs and between genes is very important for gene regulation. It is critical that transcription of the adjacent gene does not interfere with the initiation of the

transcription of the gene immediately downstream (tandem or convergent ORFs) or with the space required for the functionality of two gene promoters (divergent ORFs). It has been shown that transcriptional interference (TI) in natural cases is quite low (Puig *et al*., 1999) or even absent (Atkins *et al*., 1994) when two convergent genes coincide in a small region that is, in part, transcribed from both strands. Only when terminators are deleted is a strong TI observed, due to collisions between RNA polymerases (Prescott and Proudfoot, 2002). One reason could be that, in general, the 3′ end formation signals are degenerate, redundant and disperse (Aranda *et al*., 1998a; Van Helden *et al*., 2000), which allows for an easy overlap or even the existence of bidirectional signals (discussed in Aranda *et al*., 1998b). Interestingly, in the case of tandem genes TI seems to be caused by RNA polymerase complexes that initiate transcription from the promoter of the upstream gene and subsequently read through the promoter of the downstream gene. This causes promoter occlusion by disruption of transcription factor binding (Greger *et al*., 2000; Martens *et al*., 2004). The potential existence of TI in these cases has apparently guided the evolution of mechanisms to avoid it (Valerius *et al*., 2002). Strong transcriptional terminator signals, specific factor binding sites and nucleosomal organization have been demonstrated to be required to avoid TI (Greger *et al*., 2000; Aranda *et al*., 1998b; Valerius *et al*., 2002). Closely spaced genes are more prone to TI. These observations prompted us to study the intergenic distances in the entire genome of *S. cerevisiae*.

We first made a statistical study of the inter-ORF separations for the three types of ORF arrangements. Our conclusion is that a minimum distance exists between two consecutive ORFs and that this minimum and the typical distance for each case are specific for each of the three possible arrangements. This supports Dujon's model of promoter and terminator sizes, although our results provide more accurate estimations. From these analyses, it can be also concluded that there are different subgroups of terminator and promoter sizes for yeast genes. Our specific experimental analysis of the case of tandemly arranged ORFs demonstrates that most of the previously suspected exceptions to the 'minimum distance' rule were false ORFs that are now considered dubious (according to SGD: **http://www.yeastgenome.org/**) and are likely to be annotation artefacts. Moreover, for the few cases of very short distances, a potential TI on the downstream ORF exists, specifically, mRNAs from the upstream gene invade its promoter and coding regions, producing polycistronic transcripts.

Materials and Methods

Strains and media

We used the yeast strains S288c and BQS252 (*MAT***a**, *ura3-52*, derived from FY1679) as laboratory strains, and T73 as an alternative strain (Querol *et al*., 1996). Cells were grown in YPD (yeast extract 1%, peptone 2%, glucose 2%) in agitation at 28 °C and recovered by centrifugation at OD 0.5.

PCR and RT-PCR

PCR amplifications were performed with genomic DNA that was phenol extracted using Fast-Prep (Bio101, Inc.) and precipitated with ethanol (Hoffman and Winston, 1987). The oligonucleotides were designed to prime from the 3′ region of the first ORF of the tandem to the 5′ region of the second ORF, in order to amplify the intergenic region (see Table 1 for oligonucleotide sequences).

For RT-PCR analyses, RNA samples were also purified by phenol extraction as described (García-Martínez *et al*., 2004) and, prior to the cDNA synthesis, were treated with DNase I (RNase free, Roche) at 37 °C for 1 h, phenol-extracted and precipitated with ethanol. The cDNA synthesis was carried with SuperScript II (Invitrogen) for 1 h at 42 °C in the presence of RNase OUT (Invitrogen), using an oligo dT as primer. PCR was done using the same oligonucleotides as for DNA.

Both PCR and RT-PCR analysis were done by using Taq DNA polymerase (Biotools) and with the following cycling conditions: 3 min at 94 °C, 35 cycles of 30 s at 94 °C, 45 s at 48 °C and 60 s at 72 °C and a single step of 5 min at 72 °C.

3′ RACE

We used the 3′ RACE method (Frohman *et al*., 1988) basically with the same conditions for amplification described in RT-PCR. We used the 3′ RACE poly(T) as primer during the cDNA synthesis and 3′ RACE and the gene-specific

DOI: 10.1002/yea

Table 1. List of oligonucleotides used in this work

YPL271W-D	ATG TCT GCC TGG AGG AAA GC
YPL270W-R	GGG TGA CCA AAG ACC GAA AC
YHR130C-D	AAG GAC GCT GAA GAT CAT G
YHR131C-R	CAT CAA TGA TGC CAA TCG
YHR130C-D2	GGA GGA CGA CGG TGA CG
YMR063W-D	TGC TCA ATG GAA CAG TAC TC
YMR064W-R	TAA TGA TCT TTC GAC CGT C
YMR064W-D1	TTT GCG AGA ATA TAC TGG
YMR064W-D2	AGA ATA CGG GCG CTG TAG G
YJL089W-D	AGA GAT AGC AAT TCG GTA G
YJL088W-R	GAA GAA AGA AAC GGT GAC
YCL008C-D	ATC CAG AAC AGC ACG GAC
YCL007C-R	TTG TTC TTG GCC TTA AAC TG
YDR462W-D	ACC TTC TAC AAG AAC TCA GC
YDR463W-R	AGA ACG CGT ATA TCT TGC
YNR057C-D	GGA ACC GAA TGA AGG CAA C
YNR056C-R	GAT ACC GAC CCA TGA GCA AC
YOR077W-D	GAA GAT GGA ATG CTG CGA AG
YOR078W-R	TGC GCT GGC TTG ACT TTC
YOR078W-D	TCA AGA CAT GTA ACT TTC G
YJR120W-D	GCG CAA GGA TAT TCC CAT C
YJR121W-R	ATA GCG GGC AAC TCT GAT TG
YDR082W-D	TGA TCT TGA TCC GAA GAA TGG
YDR083W-R	GGC TAC TGA TCC TTC GGA AG
ACT1-D2	GTA TTT TCA CGC TTA CTG C
ACT1-R	TTG GTC TAC CGA CGA TAG ATG
Oligo dT	TTT TTT TTT TTT TTT (AGC)(AGCT)
3′ RACE poly(T)	GGC CAC GCG TCG ACT AG$(T)_{17}$
3′ RACE	GGC CAC GCG TCG ACT AG

oligonucleotide (see Table 1) during the PCR amplification.

Northern blot analysis

We used a 1% agarose gel with 1× MOPS and 6.4% formaldehyde. The samples were transferred to a nylon membrane (Hybond N^+, Amersham) by capillarity with 6× SSC overnight and UV-cross-linked. Then the filter was hybridized for 16 h at 42 °C in 50% formamide, 5× SSPE, 5% dextran sulphate, 0.5% SDS, 5× Denhardt and salmon sperm DNA 200 µg/ml. We used probes that cover the entire ORF and were ^{33}P random-primer labelled with Ready-to-Go (Amersham). The filters were washed twice during 10 min at 42 °C (2× SSPE and 0.1% SDS) and once during 15 min at 65 °C (1× SSPE and 0.1% SDS) and exposed to an imaging plate (BAS-MP, Fujifilm).

Statistical procedures

We calculated the Neperian logarithm of the distance between ORFs (yellow dots in Figure 1). The data were fitted using the non-linear least squares method to the less complex sum of Gaussians that can explain the data properly (red dotted line in Figure 1).

Results

In silico study of adjacent ORF distances

We analysed the distance between adjacent ORFs by dividing them into three groups: convergent, divergent and tandemly orientated (Figure 1A). The abrupt slope of the left side of each curve reflects the existence of a minimum distance between ORFs. We noticed that they are distributed, approximately, as a log-Normal curve (dots in Figure 1B–D). Thus, the maximum for each curve represents the most frequent distance observed in each case. This typical distance is 236, 490 and 402 bp for convergent, divergent and tandem ORFs, respectively. We observed that divergent ORFs are clearly composed of at least two different populations (lines in Figure 1C). Convergent and tandem ORFs form more symmetrical curves, but they can be split up into simpler Gaussian distributions (Figure 1B, D). The two main populations in divergent ORFs are centred at 290 and 771 bp. Assuming that the distance represents the space required for two promoters, their sizes would be 145 and 386 bp for the two populations, respectively. For each subpopulation, we calculated the shortest distance as the one that leaves 5% of the group below it. This value represents an approximation of the minimum separation allowed for two ORFs. These values are shown in Figure 1B–D. A plausible interpretation for such a minimum distance is that gene promoters and terminators behave, at least in part, as solid entities not overlapping with other entities. The length of the promoter and the length of the space in between two promoters would, in any case, also follow a Gaussian distribution when considering the entire yeast genome. Following this reasoning, the minimum length for a gene promoter would be half of the minimum length between divergent ORFs: 79 and 129 bp for the two subpopulations. The minimum length for a gene terminator would be 18 and 37 bp for the two subpopulations of convergent ORFs. In the case of tandem ORFs, the separation would be a composition of a promoter plus a

DOI: 10.1002/yea

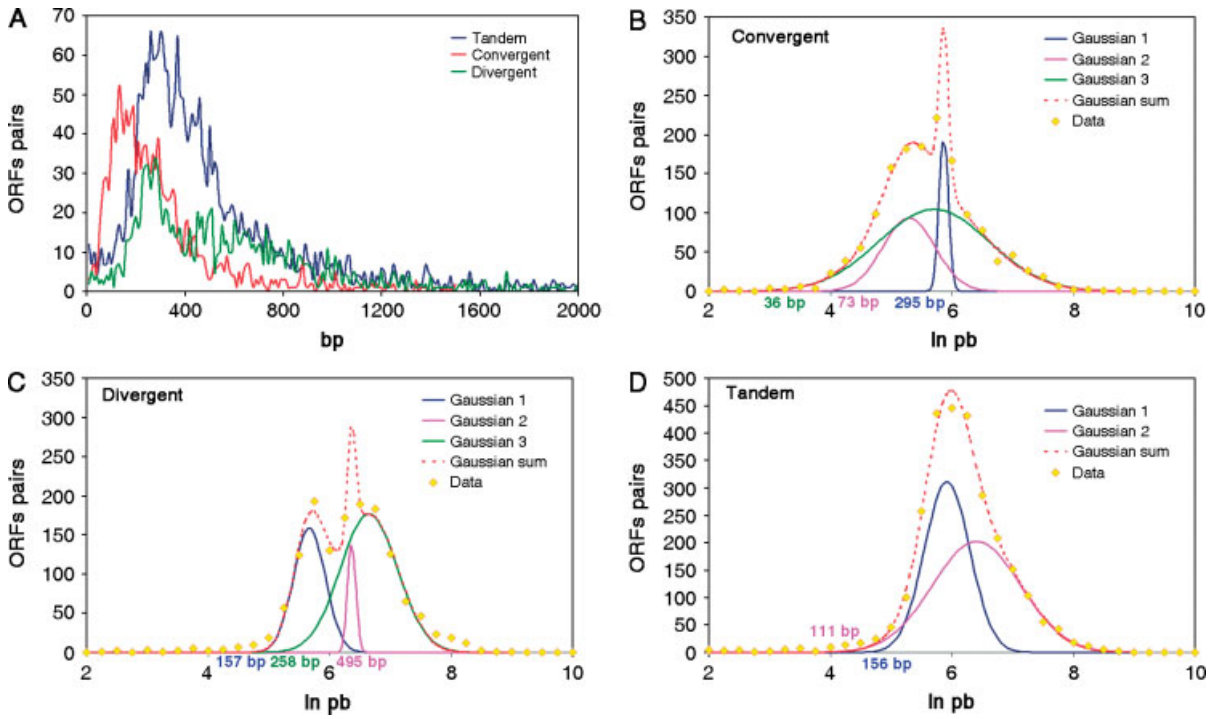

Figure 1. Distances between ORFs in the yeast genome. Size distribution of separations between adjacent ORFs classified depending on the relative ORF orientation (A). Log normal curve deconvolution of the ORF distances for the three types of ORF orientation is shown in (B)–(D). A curve (dotted line) is fitted to the experimental data (yellow dots) using the non-linear least squares method. The curve is deconvoluted into two or three simpler Gaussians that represent ORFs subpopulations. The minimum separation for each subpopulation is calculated as the distance that leaves 5% of the population below it. These minimum distances are written in the same colour as the corresponding curve

terminator. The four possible combinations, using data from convergent and divergent ORFs, are 97, 116, 147 and 166 bp. This estimation is in agreement with the results obtained graphically for two subpopulations: 111 and 156 bp. These distances represent the average of the first two and the last two combinations, respectively.

The short distance observed for the group of genes that are closer to one another suggests that this class of genes may share regulatory elements and, therefore, may have a higher probability of being co-regulated. We have analysed the cosine values for co-regulation as described by Kemmeren *et al*. (2002). Figure 2 shows that all three classes of adjacent ORFs tend to be more co-regulated than the randomly generated pairs, as previously described by other authors (Cohen *et al*., 2000; Spellman and Rubin, 2002). A list of the most probable co-regulated divergent gene pairs is shown in Table 3 (a complete list is available from the authors). With regard to the two subgroups of divergent ORFs, we observed a significant increase in the frequency of co-regulation for the group with the shortest inter-ORF distance, as compared to either the other subgroup or any of the other groups analysed. This result is 99.99% significant using a *t*-test.

In silico analysis of tandem ORFs

If we assume that there is a minimum distance for adjacent tandem ORFs, the cases in which distances are shorter than the minimum should be either exceptions or erroneously annotated ORFs. We considered seven possible reasons for this:

1. There is a single gene, but due to sequencing errors it appears as two consecutive ORFs.
2. There is a single gene in strains other than S288c. In this background it would be a pseudogene containing two ORFs.
3. There is a single promoter and two genes generating a polycistronic mRNA.

DOI: 10.1002/yea

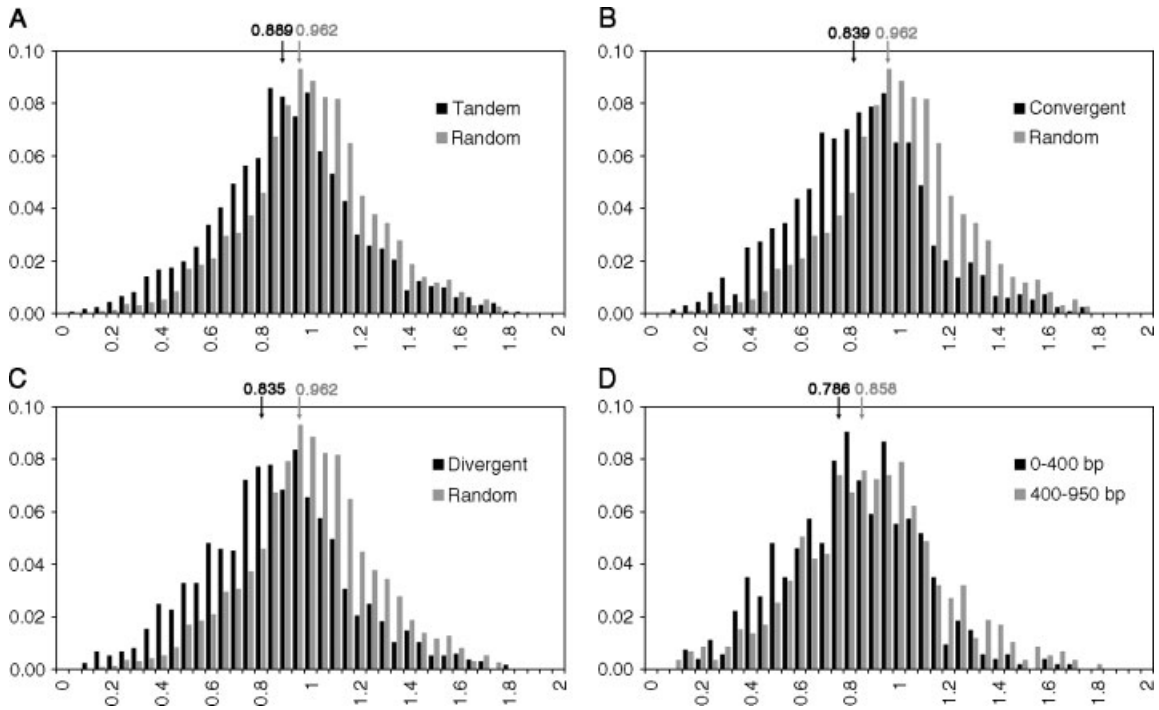

Figure 2. Histograms of the cosine co-regulation values for different adjacent ORF subpopulations. In (A)–(C) the graphs show the analyses of the whole tandem, convergent or divergent ORF classes compared to the random pair distribution. A cosine value of 0 means perfect positive correlation, 1 means no correlation and 2 perfect negative correlation. In (D) the divergent ORFs were divided into two subgroups that correspond approximately to Gaussians 1 and 3 of Figure 1C. The average value for cosine co-regulation is shown for every histogram. All the average pairs are 99.99% different using a *t*-test

4. One of the two ORFs is not a real gene.
5. There is an interrupting intron not annotated that, once spliced, converts them into a single ORF.
6. There is an annotation mistake that affects the ATG of the second ORF.
7. There is a naturally-programmed frameshift during ribosome scanning that generates a single protein.

For convergent and divergent ORFs, reasons 4 and 4/6, respectively, can also apply. We decided to study tandem ORFs because they offer more possibilities.

Using information from the CYGD database (Güldener *et al*., 2005) in July 2002, when this study started, we found 159 ORFs pairs that were separated by 156 bp or less (Table 2). During the course of this study, new genome-wide data were published that compared the genome of *S. cerevisiae* with close relatives (Kellis *et al*., 2003; Cliften *et al*., 2003; Brachat *et al*., 2003; Dujon *et al*., 2004). These studies re-annotated the ATG codons for many ORFs (possibility No. 6), discarded some spurious ORFs (possibility No. 4) and, finally, revealed some cases of pseudogenes (possibility No. 2). Moreover, the work of Namy *et al*. (2003) indicated that some tandem ORFs are part of a single protein because of a ribosome frame shift (possibility No. 7). After these corrections, many of the preselected pairs were no longer included in our list. Of the remaining 34 pairs, only 11 are below the shorter limit of 111 bp described above. Our experimental study included six of the 34 pairs and 4 additional pairs that were selected before the reduction in the list (Table 2).

Experimental testing of the different possibilities

Before the publication of the genome sequence comparisons (Kellis *et al*., 2003; Cliften *et al*., 2003; Brachat *et al*., 2003; Dujon *et al*., 2004), we analysed the sequences of the ORF pairs 1–6 by PCR amplification of the intergenic regions both in the S288c background and in a non-related strain

DOI: 10.1002/yea

Table 2. List of tandemly orientated ORF pairs separated by less than 156 bp

pb[$]				pb				pb			
0		YDR504C	YDR505C	65		YBL090W	YBL089W	123		YKL137W	YKL136W
0		YIL168W	YIL167W	66		YKL115C	YKL114C	124		YGL169W	YGL168W
1		YGL165C	YGL164C	67		YIL043C	YIL042C	**124**		**YMR068W**	**YMR069W**
2		YIL171W	YIL170W	67		YLR111W	YLR112W	124		YPR095C	YPR096C
								125	**6**		
4		YBR026C	YBR027C	68		YBR027C	YBR028C	(299)		YDR462W	YDR463W
6		YBR226C	YBR227	73		YBL071C	YBL070C	125		YJL007C	YJL006C
					1	YPL271W	YPL270W		**8**	**YOR077W**	**YOR078W**
6		YLR365W	YLR366W	73		(ATP15)	(MDL2)	**127**		**(RTS2)**	**(BUD21)**
7		YBL112C	YBL111C	74		YMR084W	YMR085W	**128**		**YDR422C**	**YDR423C**
7		YKL067W	YKL066W	**74**		**YPR168W**	**YPR169W**	129		YBR291C	YBR292C
8		YLR101C	YLR102C	76		YDR229W	YDR230W	129		YDL016C	YDL015C
					4	**YJL089W**	**YJL088W**				
8		YML101C-A	YML101C	**77**		**(SIP4)**	**(ARG3)**	**129**		**YGL212W**	**YGL211W**
9		YIL165C	YIL164C	78		YGL133W	YGL132W	129		YGR201C	YGR202C
11		YCR086W	YCR087W	81		YDR431W	YDR432W	129		YPR130C	YPR131C
									9		**YJR121W**
12		YDR024W	YDR025W	81		YKL021C	YKL020C	**130**		**YJR120W**	**(ATP2)**
12		YKL031W	YKL030W	81		YOR300W	YOR301W	**131**		**YOR059C**	**YOR060C**
12		YOL163W	YOL162W	**84**		**YML014W**	**YML013W**	132		YGL244W	YGL243W
13		YLR202C	YLR203C	85		YLR373C	YLR374C	132		YOR282W	YOR283W
16		YLR433C	YLR434C	86		YBL063W	YBL062W	**133**		**YIR014W**	**YIR015W**
16		YNL180C	YNL179C	87		YNL320W	YNL319W	134		YDR487C	YDR486C
24		YDR157W	YDR158W	89		YMR153C-A	YMR154C	134		YGL053W	YGL052W
				90	**10**	**YDR082W**	**YDR083W**				
24		YER119C	YER119C-A	**(147)**		**(STN1)**	**(RRP8)**	134		YKL208W	YKL207W
24		YJR023C	YJR024C	**91**		**YLR384C**	**YLR385C**	135		YCR013C	YCR014C
24		YLR311C	YLR312C	92		YNL286W	YNL285W	136		YDR396W	YDR395W
24		YNR065C	YNR066C	95		YKL022C	YKL021C	137		YDR203W	YDR204
26		YIL025C	YIL024C	95		YOL107W	YOL106W	137		YDR400W	YDR401W
28		YDR290W	YDR291W	95		YOR183W	YOR184W	139		YKL111C	YKL110C
28		YIL086C	YIL085C	96		YBR090C	YBR091C	**140**		**YNL046W**	**YNL045W**
29		YEL034W	YEL033W	99		YKR103W	YKR104W	142		YMR075C-A	YMR076C
29		YFL057C	YFL056C	101		YOR330C	YOR331C	**143**		**YCL059C**	**YCL058C**
31		**YLR393W**	**YLR394W***	102		YAL046C	YAL045C	143		YLR030W	YLR031W
32		YHL006C	YHL005C	104		YGL129C	YGL128C	**143**		**YNL158W**	**YNL157W**
33		YIL087C	YIL086C	105		YKL044W	YKL043W	146		YNL316C	YNL315C
39	**2**	YHR130C	YHR131C**	105		YML095C	YML095C-A	**146**		**YOL025W**	**YOL024W**
41		YGR044C	YGR045C	**105**		**YOL068C**	**YOL067C**	147		YCL009C	YCL008C
					3	**YMR063W**	**YMR064W**				
43		YGR025W	YGR026W	**106**		**(RIM9)**	**(AEP1)**	147		YDR219C	YDR220C
45		YOR014W	YOR015W	107		YOR081C	YOR082C	147		YGL046W	YGL045W
47		YEL046C	YEL045C	**107**		**YPR066W**	**YPR067W**	148		YER134C	YER135C
47		YGR163W	YGR164W	110		YLR281C	YLR282C	**149**		**YML079W**	**YML078W**
48		YOR024W	YOR025W	111		YDR423C	YDR424C	150		YLL031C	YLL030C
49		YPL278C	YPL277C	111		YOL135C	YOL134C	**150**		**YNL149C**	**YNL148C**
52		**YER152C**	**YER153C**	112		YOR125C	YOR126C	**151**		**YGL231C**	**YGL230C**
52	**7**	**YNR056C**	**YNR057C**								
(142)		**(BIO5)**	**(BIO4)**	**114**		**YLR414C**	**YLR415C**	151		YGL035C	YGL034C
54		YBL049W	YBL048W	114		YMR056C	YMR057C	**151**		**YHR110W**	**YHR111W**
53		YGL241W	YGL240W	**115**		**YHL014C**	**YHL013C**	152		YMR151W	YMR152W
55		YGL128C	YGL127C	115		YMR213W	YMR214W	153		YAL045C	YAL044C
56		YJL097W	YJL096W	116		YDL172C	YDL171C	153		YDR204W	YDR205W
56		YJL021C	YJL020C	116		YDR467C	YDR468C	**153**		**YHR057C**	**YHR058C**
56		YML032C	YML031C-A	**119**		**YFR046C**	**YFR047C**	**154**		**YOR022C**	**YOR023C**
					5	**YCL008C**	**YCL007C**				
56		YOR302W	YOR303W	**120**		**(STP22)**	**(CWH36)**	**154**		**YPL068C**	**YPL067C**
57		YDL158C	YDL157C	120		YDR023W	YDR024W	**155**		**YGL005C**	**YGL004C**
57		YER039C	YER039C-A	121		YMR294W-A	YMR294W	155		YGR164W	YGR165W
61		YKL084W	YKL083W	122		YBR032W	YBR033W	155		YMR103C	YMR104C
63		YKR072C	YKR073C	122		YCR085W	YCR086W	156		YNL129W	YNL128W

* The 34 pairs that are separated by less than 156 bp after sequence and ATG correction updates are in bold.

** The experimentally studied cases are highlighted and numbered.

[$] The distances shown are those initially assigned in the CYGD data bank (July 2002). In some cases, the newly assigned distance is shown in brackets.

DOI: 10.1002/yea

Table 3. List of the 75 divergently orientated ORF pairs separated by less than 400 bp that are likely to share regulatory elements

Divergent ORFs*		Cosine correlation**	Divergent ORFs		Cosine correlation	Divergent ORFs		Cosine correlation
YNL062C	YNL061W	0.104	YLR065C	YLR066W	0.356	YHR107C	YHR108W	0.431
YHR065C	YHR066W	0.104	YPL246C	YPL245W	0.359	YLR203C	YLR204W	0.432
YER126C	YER127W	0.127	YFR041C	YFR042W	0.362	YPR085C	YPR086W	0.443
YNL248C	YNL247W	0.129	YLL036C	YLL035W	0.369	YNL334C	YNL333W	0.450
YPL212C	YPL211W	0.155	YFL060C	YFL059W	0.370	YPL094C	YPL093W	0.450
YAL036C	YAL035W	0.220	YML047C	YML046W	0.371	YIL154C	YIL153W	0.451
YKL144C	YKL143W	0.227	YHR069C	YHR070W	0.371	YKL174C	YKL173W	0.453
YER142C	YER143W	0.237	YOR164C	YOR165W	0.375	YDL226C	YDL225W	0.457
YOR167C	YOR168W	0.238	YLR209C	YLR210W	0.375	YDR067C	YDR068W	0.458
YNL002C	YNL001W	0.247	YFR003C	YFR004W	0.376	YDR329C	YDR330W	0.461
YKR024C	YKR025W	0.264	YIL098C	YIL097W	0.392	YGR172C	YGR173W	0.462
YBR141C	YBR142W	0.288	YER018C	YER019W	0.392	YGR078C	YGR079W	0.462
YIL020C	YIL019W	0.293	YKR043C	YKR044W	0.393	YMR159C	YMR160W	0.466
YFL002C	YFL001W	0.310	YBL010C	YBL009W	0.397	YCL033C	YCL032W	0.467
YDR120C	YDR121W	0.310	YPL004C	YPL003W	0.398	YDR196C	YDR197W	0.469
YIL104C	YIL103W	0.314	YGL003C	YGL002W	0.398	YNL074C	YNL073W	0.470
YLL062C	YLL061W	0.325	YGL048C	YGL047W	0.401	YHR024C	YHR025W	0.472
YAL065C	YAL064W-B	0.326	YML023C	YML022W	0.401	YBR264C	YBR265W	0.477
YPR186C	YPR187W	0.330	YOR219C	YOR220W	0.406	YLR014C	YLR015W	0.487
YNL294C	YNL293W	0.334	YIR036C	YIR037W	0.408	YCR002C	YCR003W	0.488
YKR081C	YKR082W	0.335	YKL074C	YKL073W	0.412	YBR245C	YBR246W	0.491
YJR073C	YJR074W	0.337	YJR049C	YJR050W	0.415	YOR288C	YOR289W	0.491
YGL246C	YGL245W	0.344	YJL192C	YJL191W	0.416	YOR035C	YOR036W	0.492
YBR046C	YBR047W	0.351	YBL057C	YBL056W	0.420	YPL201C	YPL200W	0.494
YDR020C	YDR021W	0.354	YOL077C	YOL076W	0.421	YML061C	YML060W	0.498

* Only those ORF pairs that do not include a dubious one (according to SGD).

** Defined in Materials and methods. Note that the lower the value of cosine, the higher the correlation.

(T73). In all cases, we detected no differences in DNA sequence that could convert the ORF pair into a single ORF (not shown). This discarded possibilities Nos. 1 and 2 for these ORF pairs.

We performed RT-PCR on total RNA from the S288c strain using the same oligonucleotides (priming inside the two consecutive ORFs) to check for the existence of an intron. In all cases, we obtained a fragment with the same size as that obtained from genomic DNA (see Figure 3B for two examples). This result allowed us to discard possibility No. 5, but demonstrates that an mRNA extends over the region of the two ORFs, supporting possibility No. 3.

In order to check for the existence of polycistronic mRNAs in the 10 ORF pairs studied, we used the 3′ RACE method to map the ends of the mRNAs. Figure 3A and Table 4 show the 3′ ends of the mRNAs. As shown, we observed that in most cases mRNAs have alternative 3′ ends. In two cases, ORF pairs 7 and 10, the mRNA does not enter the next ORF, although it should overlap with the 5′ end of the next ORF mRNA. In the rest of the cases, however, the mRNA enters extensively

Table 4. List of mRNA overlaps between experimentally studied tandemly orientated ORF pairs

Pair**	Number of nucleotides that the mRNA extends beyond the stop codon of the first gene of the pair
1	Approximately 260 and 210 nt
2	Approximately 180 and 280 nt
3	988, 1109, 1789 and 1879 nt* (the last two mRNA completely cover both ORFs)
4	Approximately 90, 140, 410 nt
5	Could not be determined in this study
6	A population of mRNA that extends approximately between 160 and 210 nt
7	89 and 113 nt*
8	841 nt* (completely covers both ORFs)
9	Could not be determined in this study
10	147 nt* (match exactly with the second ORF start)

* Confirmed by sequencing.

** Numbers as in Table 2.

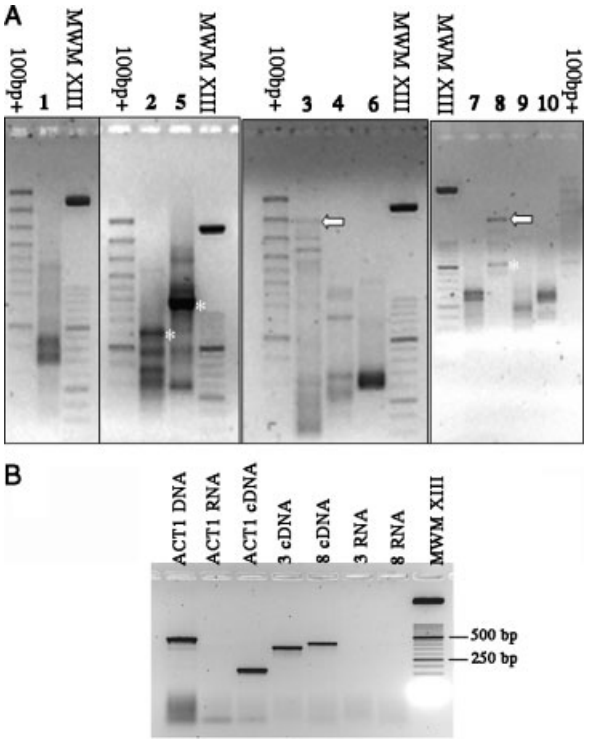

Figure 3. Identification of polycistronic mRNAs. (A) 3′ RACE of the mRNAs of several ORF pairs. The agarose gel electrophoresis shows the result of RACE bands of the candidates, numbered as in Table 4. The white arrows mark the sequenced fragments that correspond to mRNAs completely covering the adjacent next ORF. The asterisks indicate unspecific bands discarded after sequencing. Molecular weight markers used were MWM XIII from Roche (50 bp ladder from 50 to 750 bp) and Gene Ruler 100 bp+ ladder from MBI Fermentas. (B) RT-PCR analysis of intergenic region between ORF pairs 3 and 8. In both loci, a band (obtained by RT-PCR from oligonucleotides marked in Figure 4B) that spans both ORFs is seen. *ACT1* is used as a control because it contains an intron. DNA contamination in the RNA sample would give a 480 bp band (lane ACT1DNA). This band is not present in the RNA sample either before (ACT1RNA) or after RT (ACT1cDNA). A shorter band of 172 bp due to intron splicing is seen instead in the cDNA sample, which demonstrates that the bands of 367 bp and 418 bp seen in samples 3cDNA and 8cDNA are not due to DNA contamination

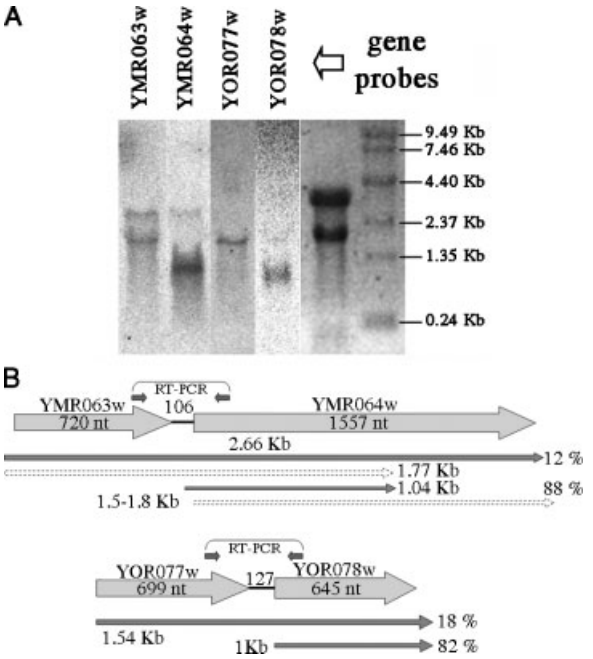

Figure 4. Transcription scheme of two selected loci. (A) Northern blot for confirmation of the mRNA sizes of the ORF pairs 3 and 8. Lanes 5, 6 on the right show ethidium bromide stain of the sample and an RNA size marker. Lanes 1–4 show the hybridization of lane 5 successively with the gene probes. (B) Model for the transcripts encoded by the two ORF pairs analysed. The regions covered by the different mRNAs are depicted. Results were obtained from the 3′ RACE (Figure 3A) and Northern analyses. The oligonucleotides used for RT-PCR (Figure 3B) are shown. The left one was also used for 3′ RACE (Figure 3A). A previously described 1.5–1.8 kb mRNA (Payne *et al*., 1993) that we did not detect and a 1.77 kb mRNA only detected with YMR063w probe are shown as a dotted line. The relative amount of each mRNA is marked

into the next ORF. In two cases, ORF pairs 3 and 8, the sizes of the mRNAs detected (Figure 3A) and the RT-PCR amplification of the intergenic region (Figure 3B) suggests the existence of complete polycistronic mRNAs. In order to precisely map the 3′ ends of these mRNAs, we sequenced the RACE bands for pairs 3 and 8. The analysis of pair 3 indicated that an mRNA from YMR063W (*RIM9*) enters YMR064W (*AEP1*) to positions 1708, 1829, 2509 and 2599 bp from the ATG of the *RIM9*. The longest mRNA completely covers the *AEP1* ORF (see Figure 4B). For ORF pair 8, there is an mRNA which extends to position 1540 from the ATG that also covers the second ORF, YOR078W (*BUD21*) (see Figure 4B). In order to further confirm the RACE results, to evaluate the relative abundance of the mRNA species and to position the 5′ ends of the mRNAs, we performed a series of Northern blots and hybridizations with probes from the two ORFs from each pair (Figure 4A). The interpretation of the results from the 3′ RACE and Northern analysis is shown in Figure 4B. Our results

DOI: 10.1002/yea

on YMR064W coincide in part with those published by Payne *et al*. (1993), who found several mRNAs, including those of 2.66 and 1.77 kb. With regard to the YOR077–078W pair, our results are compatible with those of Hurowitz *et al*. (2003), who found, apart from the expected 1 kb mRNA for *BUD21*, a minor long mRNA that they considered a false positive. We did not detect a short mRNA for YOR077W (*RTS2*) in our experimental conditions.

Discussion

The representation of inter-ORF distances in yeast appears to correspond to a log-Normal distribution. A simple explanation for such a distribution is that it is caused by a limit in size in the left part (the minimum distance allowed) and a random distribution on the other side (there is no maximum distance in between ORFs). The deconvolution of the curves into simpler Gaussians allows us to determine that all three populations seem to be composed of at least two or three groups of inter ORF distances.

Using the distributions, we can calculate the typical and the minimum size for each of the groups. Using the reasoning of Dujon (1996), the distance in between two divergent ORFs corresponds to the sizes of two adjacent promoters and the distance in between two convergent ORFs corresponds to the sizes of two adjacent terminators. The sizes calculated, thus, for the two main groups of gene promoters would be 145 and 386 bp with minima of 79 and 129 bp, respectively. What could be the reason for the existence of at least two different populations of gene promoters? One plausible explanation is the allowance of a partial overlap between promoters. The smaller size group would correspond to overlapping promoters in which some non-essential elements could be interspersed, or even shared between the two genes. If this were true, the genes belonging to this group would have a higher probability of being co-regulated. Our analysis (Figure 2) demonstrates that, in fact, this is the case. One can imagine that a gene promoter should have essential elements, such as TATA and Inr, which cannot be shared because of their unidirectionality. These elements tend to be very close to the ORF's ATG. Other non-essential elements, however, can be either bidirectional or, because of their non-essential role, can be placed alternatively within the region in between two divergent ORFs. The first option would give rise to co-regulated genes and would explain our results. Several examples of co-regulated divergent yeast genes are already known (e.g. *GAL1–GAL10*, or the cases listed in Wade *et al*., 2006). Although this hypothesis could not be proved in a previous analysis made by Cohen *et al*. (2000), we think that our approach, which differentiates between subgroups of divergent genes, strongly supports the hypothesis that one of the main reasons for co-regulation is the use of partially shared promoters. We postulate that many others will be discovered in the future and that the probability of such cases will be higher for genes belonging to the closest distance subgroup. A list of the most likely cases is shown in Table 3. Similar arguments can be used for terminators but, in that case, co-regulation will not be a biological consequence of overlapping. For tandem genes, combinations of overlapping and non-overlapping promoters and terminators can exist as well.

The hypothesis of the existence of a minimum distance necessary for the separation of non-overlapping elements raises the question of why there are some exceptions for the three cases. Are they real exceptions or do they correspond to erroneously annotated ORFs? We selected the case of tandem ORFs to experimentally address these questions. The fact that 79% of the 159 initially selected ORFs pairs in July 2002 were discarded by several studies (Kellis *et al*., 2003; Cliften *et al*., 2003; Brachat *et al*., 2003; Dujon *et al*., 2004; Namy *et al*., 2003) suggests that, for the vast majority of cases, annotation errors are the cause for the short distance. Our results on a sample of the rest of the cases also support this hypothesis (see below).

Our experimental study on a sample of the tandem ORF pairs concludes that, in most cases, the close proximity between two ORFs leads to invasion of the upstream mRNA into the region covered by the downstream ORF. This situation may be not neutral for the second gene. It has been demonstrated previously that TI can occur when an RNA polymerase enters the promoter of the downstream gene. TI can occur when mutations occur in the terminator of the first gene (Greger and Proudfoot, 1998; Valerius *et al*., 2002). However, it is more interesting to note that there are cases in

DOI: 10.1002/yea

which wild-type sequences allow TI on the downstream gene (Greger and Proudfoot, 1998; Greger *et al*., 2000; Martens *et al*., 2004). In these cases, RNA polymerases invade the promoter, causing promoter occlusion by affecting transcription factor binding (Greger *et al*., 2000; Martens *et al*., 2004) and they can even transcribe the downstream ORF, producing bicistronic mRNAs (Gerger and Proudfoot, 1998). This has been shown for tandem ORFs that are separated by 191 bp (*FBP1–PSY3*), 417 bp (*ARO4–HIS7*) or 600 bp (*GAL10–GAL7*) (Springer *et al*., 1997; Aranda *et al*., 1998b; Greger *et al*., 2000). In some cases, however, no large effects on downstream transcription by TI were observed (Aranda *et al*., 1998b). Because changes in efficiency of both the promoter and terminator affects TI (Springer *et al*., 1997; Greger *et al*., 1998), it seems likely that a fine tuning of the 3′ end formation signal to the promoter strength of the upstream mRNA is necessary to prevent TI from the adjacent gene.

In conclusion, given that even tandem ORFs separated by 600 bp are susceptible to TI, it is conceivable that most of the members of this group of ORFs are potentially affected by TI (74% of the tandem ORFs are under this distance). The cell has developed mechanisms to avoid such perturbations, such as efficient terminators, DNA-binding factors or/and nucleosomal organization (Valerius *et al*., 2002). Whatever the mechanisms, the shorter the distance between ORFs, the more difficult their use. This may be one of the reasons for the existence of a minimum separation distance. Another reason may be that, although the promoter and the terminator of two successive genes can be interspersed (Aranda *et al*., 1998b; Springer *et al*., 1997), this is again more difficult for shorter distances. We believe that in some cases (as in pairs 3 and 8 shown in this study) TI is not efficiently avoided and this results in minor polycistronic mRNAs. Whether or not these mRNAs are functional, especially for the translation of the second ORF, is a question for further investigation.

Acknowledgements

This work was supported by Grants GEN2001-4707-C08-07 and BMC2003-07072-C03-02 from the Ministerio de Educación y Ciencia (to J.E.P-O) and Grupos 03/096 from the Generalitat Valènciana to Vicente Tordera. V.P. is a fellowship holder of the Consellería de Educació i Ciència de la Generalitat Valènciana.

References

Aranda A, Pérez-Ortín JE, Moore C, del Olmo M. 1998a. Transcription termination downstream of the *Saccharomyces cerevisiae FBP1* poly(A) site does not depend on efficient 3′ end processing. *RNA* **4**: 303–318.

Aranda A, Pérez-Ortín JE, Moore C, del Olmo M. 1998b. The yeast *FBP1* poly(A) signal functions in both orientations and overlaps with a gene promoter. *Nucleic Acids Res* **26**: 4588–4596.

Atkins D, Arndt GM, Izant JG. 1994. Antisense gene expression in yeast. *Biol Chem Hoppe-Seyler* **375**: 721–729.

Brachat S, Dietrich FS, Voegeli S, *et al*. 2003. Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biol* **4**: R45.

Cliften P, Sudarsanam P, Desikan A, *et al*. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**: 71–76.

Cohen BA, Mitra RD, Hughes JD, Church GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26**: 183–186.

Dujon B. 1996. The yeast genome project: what did we learn? *Trends Genet* **12**: 263–270.

Dujon B, Sherman D, Fischer G, *et al*. 2004. Genome evolution in yeasts. *Nature* **430**: 35–44.

Frohman MA, Dush MK, Martin GR. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci USA* **85**: 8998–9002.

Eggermont J, Proudfoot NJ. 1993. Poly(A) signals and transcriptional pause sites combine to prevent interference between RNA polymerase II promoters. *EMBO J* **12**: 2539–2548.

García-Martínez J, Aranda A, Pérez-Ortín JE. 2004. Genomic Run-On evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Mol Cell* **15**: 303–313.

Greger IH, Aranda A, Proudfoot N. 2000. Balancing transcriptional interference and initiation on the *GAL7* promoter of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **97**: 8415–8420.

Greger IH, Proudfoot NJ. 1998. Poly(A) signals control both transcriptional termination and initiation between the tandem *GAL10* and *GAL7* genes of *Saccharomyces cerevisiae*. *EMBO J* **16**: 4771–4779.

Güldener U, Münsterkotter M, Kastenmuller G, *et al*. 2005. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res* **33**: D364–D368.

van Helden J, del Olmo M, Pérez-Ortín JE. 2000. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res* **28**: 1000–1010.

Hoffman CS, Winston F. 1987. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. *Gene* **57**: 267–272.

Hurowitz EH, Brown PO. 2003. Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*. *Genome Biol* **5**: R2.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–254.

Kemmeren P, van Berkum NL, Vilo J, *et al*. 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9**: 1133–1143.

Marín A, Wang M, Gutiérrez G. 2004. Short-range compositional correlation in the yeast genome depends on transcriptional orientation. *Gene* **333**: 151–155.

Martens JA, Laprade L, Winston F. 2004. Intergenic transcription is required to repress the *Saccharomyces cerevisiae SER3* gene. *Nature* **429**: 571–574.

Namy O, Duchateau-Nguyen G, Hatin I, *et al*. 2003. Identification of stop codon readthrough genes in *Saccharomyces cerevisiae*. *Nucleic Acids Res* **9**: 2289–2296.

Prescott EM, Proudfoot NJ. 2002. Transcriptional collision between convergent genes in budding yeast. *Proc Natl Acad Sci USA* **99**: 8796–8801.

Puig S, Pérez-Ortín JE, Matallana E. 1999. Transcriptional and structural study of a region of two convergent overlapping yeast genes. *Curr Microbiol* **39**: 369–373.

Querol A, Ramón D. 1996. The application of molecular techniques in wine microbiology. *Trends Food Sci Technol* **7**: 73–78.

Spellman PT, Rubin GM. 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J Biol* **1**: 5.

Springer C, Valerius O, Strittmatter A, Braus GH. 1997. The adjacent yeast genes *ARO4* and *HIS7* carry no intergenic region. *J Biol Chem* **272**: 26318–26324.

Valerius O, Brendel C, Duvel K, Braus GH. 2002. Multiple factors prevent transcriptional interference at the yeast *ARO4-HIS7* locus. *J Biol Chem* **24**: 21440–21445.

Wade C, Umbarger M, McAlear M. 2006. The budding yeast rRNA and ribosome biosynthesis (RRB) regulon contains over 200 genes. *Yeast* **23**: 293–306.

Zhang Z, Dietrich FS. 2005. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. *Nucleic Acids Res* **33**: 2838–2851.