Genomic Exploration of the Hemiascomycetous Yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited

Gaëlle Blandin[a], Pascal Durrens[b], Fredj Tekaia[a], Michel Aigle[b], Monique Bolotin-Fukuhara[c], Elisabeth Bon[d], Serge Casarégola[d], Jacky de Montigny[e], Claude Gaillardin[d], Andrée Lépingle[d], Bertrand Llorente[a], Alain Malpertuy[a], Cécile Neuvéglise[d], Odile Ozier-Kalogeropoulos[a], Arnaud Perrin[a], Serge Potier[e], Jean-Luc Souciet[e], Emmanuel Talla[a], Claire Toffano-Nioche[c], Micheline Wésolowski-Louvel[f], Christian Marck[g], Bernard Dujon[a,*]

[a] *Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR 927 Univ. P. M. Curie, Paris), Institut Pasteur, 25, Rue du Dr Roux, F-75724 Paris Cedex 15, France*
[b] *Laboratoire de Biologie Cellulaire de la Levure, IBGC, 1 rue Camille Saint-Säens, F-33077 Bordeaux Cedex, France*
[c] *Institut de Génétique Moléculaire (UMR 8621 CNRS), Bâtiment 400, Université de Paris Sud-Orsay, F-91405 Orsay Cedex, France*
[d] *Collection de Levures d'Intérêt Biotechnologique, Laboratoire de Génétique Moléculaire et Cellulaire (INRA UMR216, CNRS URA1925), INA-PG, BP01, F-78850 Thiverval-Grignon, France*
[e] *Laboratoire de Génétique et Microbiologie, (UPRES-A 7010 ULP/CNRS), Institut de Botanique, 28 rue Goethe, F-67000 Strasbourg Cedex, France*
[f] *Microbiologie et Génétique (ERS2009 CNRS/UCB/INSA), Bâtiment 405 R2, Université Lyon I, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne Cedex, France*
[g] *Service de Biochimie et de Génétique moléculaire, Département de Biologie Cellulaire et Moléculaire, DSV/CEA-Saclay, F-91191 Gif-sur-Yvette, France*

Received 9 November 2000; accepted 11 November 2000

First published online 27 November 2000

Edited by Horst Feldmann

Abstract Since its completion more than 4 years ago, the sequence of *Saccharomyces cerevisiae* has been extensively used and studied. The original sequence has received a few corrections, and the identification of genes has been completed, thanks in particular to transcriptome analyses and to specialized studies on introns, tRNA genes, transposons or multigene families. In order to undertake the extensive comparative sequence analysis of this program, we have entirely revisited the *S. cerevisiae* sequence using the same criteria for all 16 chromosomes and taking into account publicly available annotations for genes and elements that cannot be predicted. Comparison with the other yeast species of this program indicates the existence of 50 novel genes in segments previously considered as 'intergenic' and suggests extensions for 26 of the previously annotated genes.

Key words: Annotation; Pseudogene; Intron; Frameshift; Family

1. Introduction

The complete sequences of the 16 chromosomes of the laboratory strain S288C (or its derivative) of *Saccharomyces cerevisiae* were released from 1992 to 1996 [1–16]. This scientific masterpiece (the first eukaryote ever sequenced) has been subsequently extended by the completion of the sequence of mitochondrial DNA from the same strain [17]. During the last 4 years, the chromosome sequences have been the object of some minor revisions, some of them extending coding regions originally frameshifted, and of a few more significant ones such as the addition of ca. 10 kb of sequences on chromosome IV or of a previously missing telomere on chromosome V. A European program has also been launched to entirely resequence chromosome III, the first chromosome ever sequenced, since its original version was a composite of sequences from several different strains. Edited chromosome sequences are maintained at MIPS (http://www.mips.biochem.mpg.de/proj/yeast/).

In addition to the major publicly available sequence archives such as EMBI and GenBank, the annotated yeast sequence and its interpretation can be accessed from three specialized sites: MIPS which was responsible for the original annotation and systematic nomenclature, SGD (http:// genome-www.stanford.edu/saccharomyces/) which, among other things, maintains the original gene name register, and YPD (http://www.proteome.com/databases/index.html) which is specialized in proteins and maintains a complete literature survey program.

Since its completion, the yeast genome sequence has been examined and used by such a large number of scientists from both academies and industries that its applications are too numerous to be quoted. Yet, a major fundamental question that has remained largely unsolved is the exact number of genes it contains. Even in a genome in which introns are scarce, gene prediction is prone to uncertainties. A number of the open reading frames (ORFs) were mentioned as 'questionable' or 'unlikely' in the original annotations [1–16]. Attempts have subsequently been made to rationalize the distinction between probable and improbable ORFs by using additional criteria [18]. The problem is obviously exacerbated by the fact that a rather large number of yeast ORFs have no obvious structural homologs in other completely sequenced organisms or in general databases, and that a number of such ORFs have no function characterized in *S. cerevisiae*.

*Corresponding author. Fax: +33-1-40613456.
E-mail: bdujon@pasteur.fr

0014-5793/00/$20.00
PII: S0014-5793(00)02275-4

Those, combining both an absence of homology outside of yeast and an absence of known function, were originally termed 'orphans' (see [19]). And the question rapidly arose as to the mere existence of 'orphans'. Some authors using theoretical calculations from the DNA sequences even arrived at the conclusions that orphans do not exist and that, consequently, nearly all yeast genes were already known and similar to that of other species [20,21]. Despite the fact that this extreme conclusion is biologically strange (it is perhaps the most obvious signature of the living world that its members simultaneously show unity for the major processes and diversity for many other aspects), it was based in some cases [21] on an incorrect calculation mixing out-of-frame sequences (internal sequences to larger ORFs in either the same or the opposite frame that were never listed in the original yeast genome annotation) and truly questionable ORFs (often listed as such in the original yeast genome annotation).

Annotation of the yeast genome is also imperfect in that it omits some genes, particularly the shortest ones, more difficult to predict [22–24]. Novel genes have been found by the extensive gene expression studies that have been carried out during the last few years [25,26]. The original SAGE study, for example, cited a few transcribed sequences not listed as genes or ORFs. It is obvious, however, that not all transcripts necessarily correspond to protein-coding genes and that a careful examination of the sequences is needed (for example, some of the SAGE tags correspond to LTR sequences or retrotransposons). But the annotation of the yeast genome generally benefited from the transcript studies.

The present sequencing program of 13 hemiascomycetous yeast species [27] offered a unique chance to re-examine the *S. cerevisiae* genome by extensive sequence comparisons with related species. At the same time, we reasoned that a unique version of that genome was needed in order to perform significant comparisons. In the first step, we, therefore, decided to use a dated release of the 16 chromosomes and to reextract from it all predictable ORFs using uniform and explicit criteria since original ORF prediction criteria were not exactly the same for all chromosomes (for example compare chromosomes I, III, VI, VIII, and XI) and the results of such predictions are carried over in present day databases. In the second step, we have examined our predictions along with the annotations of MIPS and included introns and short genes not predicted by our program. In the last step, when using the annotated *S. cerevisiae* sequences as explained in this article to compare against the 13 other yeast species, we discovered 50 novel genes (three of which with introns) and could propose sequence extensions relative to original annotations for 26 others.

Finally, we also used our list of ORFs to define 'partitions' and gene 'families' that were useful in the annotation of the other yeast species and to compare the redundancy of the *S. cerevisiae* genome with that of the other species [28].

Analysis of our repertoire, by comparison to other yeast species, demonstrates the existence and conservation of many of the previously 'orphan' genes and allows us to propose that the actual protein-coding gene set of *S. cerevisiae* amounts to at least 5600 genes.

2. Materials and methods

2.1. DNA sequences

DNA sequences of each of the 16 *S. cerevisiae* chromosomes were downloaded from MIPS on March 2nd, 1999 (http://www.mips.biochem.mpg.de/proj/yeast/). The resequenced version of chromosome III, available at MIPS since August, 2000 was not used in this study.

2.2. Principles for the definition of genetic elements

Dedicated C Macintosh software has been developed to handle the annotations in the sequences of complete genome sequences of lower eukaryotes. The complete sequence of an entire chromosome is read and interpreted automatically for ORFs and tRNA genes, generating an associated file, which can then be manually annotated for other genetic elements or specific features. Genetic elements predicted, annotated or manually validated or disqualified are classified in different types.

For each novel chromosome sequence, four types of genetic elements are automatically defined: (i) 'Protein', defined as an ORF of a length exceeding a selected limit (in this work, a minimum of 99 codons was demanded) followed by a stop codon and not entirely included in another ORF, (ii) 'Disregarded ORF', defined as the same type of object but entirely included in a longer one (whatever the orientation), (iii) 'tRNA', defined as the entire coding sequence of a tRNA gene including its intron, if present, and (iv) 'Linker', defined as all segments of DNA excluding the above.

Other elements are manually defined in the associated file. These include protein-coding genes that may be added because they are shorter than the set limit or edited because they contain intron(s). The resulting genes are classified as 'Blessed'. After the definition of 'Introns', the remaining exons that would be automatically predicted in subsequent runs are 'Damned'. Other manually defined elements contain the 'RNA', defined as the sequences encoding known RNAs, 'LTR', defined as the sequences corresponding to known LTRs, 'Centromere', defined as the sequences extending from conserved block I to block III of centromeres, and 'Subtelomeric', defined as the sequences corresponding to the telomeric (CACA1–3) repeats. For convenience, Ty reading frames, which are frameshifted are 'Damned'.

The initial process of automatic definition of genetic elements is then repeated taking into account the new user-defined genetic objects. These objects are all automatically checked by the software before being accepted. For example, a 'Blessed' object is checked for having no in-frame stop codon. The user can inspect all problems identified by the software and either correct the annotation or confirm it. Once all new elements are validated, a final automatic search is performed and all 'Linkers' are recomputed as the sequences left in the intervals of 'Protein', 'Blessed', 'tRNA', 'Centromere', 'RNA', 'LTR' and 'Subtelomeric' elements. Note that 'Damned' and 'Disregarded' sequences are integrated into the 'Linkers'.

This cycle of prediction–validation can be repeated as many times as needed until all elements are satisfactorily defined. Each of the 16 *S. cerevisiae* chromosomes was analyzed and annotated independently in this way and, eventually, the sequence files and the associated annotation files were merged leading to a single sequence file with over 15 000 annotations. All genetic elements defined and validated by this process can be extracted by type, filed into proper folders, and can be used by any other software for further analyses.

2.3. ORFs predicted from the complete sequence of S. cerevisiae

A complete list of the ORFs predicted using the strategy defined above is available at http://cbi.genopole-bordeaux.fr/genolevures. The following ORFs, listed at MIPS, have been ignored in the present work because they correspond to tRNA genes (*YNL017c*, *YNL285w*), overlap LTR or other Ty elements (*YBL107wa*, *YCR018ca*, *YDR034ca*, *YER138wa*, *YGR122ca*, *YHR145c*, *YIL080w*, *YMR046wa*, *YMR158cb*, *YOL013wa*, *YOL106w*, *YPR002ca*) or are part of subtelomeric repeats (*YDR543c*, *YGR296w*, *YHL050c*, *YIL177c*, *YJL225c*, *YLR464w*, *YNL338w*, *YNL339c*, *YPL283c*, *YPR202w*).

2.4. Analysis of intergenes and subtelomeric sequences

In a first step, the 6274 intergenes ('Linkers') were compared with the entire set of RSTs from each of the 13 yeast species, using tblastx. In a second step, comparisons of the same sequences were performed using blastx against GPROTEOME and against the 6213 protein sequences of *S. cerevisiae* [29]. Novel *S. cerevisiae* genes discovered were compared to GenBank (see Table 1).

2.5. Establishing the list of partitions and gene families

Every *S. cerevisiae* ORF product was compared to all other *S. cerevisiae* ORF products using blastp version 2 [30] with the pam250 substitution matrix in order to favor large segment pairs (and hence detect distantly related ORFs) [31], and the seg filter to eliminate compositionally biased regions in the query sequence [32]. The limit of significance of the blastp probability scores was set at 10^{-9}, after the simulation procedure described in [33]. ORFs were classified in 'partitions' and 'families' based on a top-down approach that use both single-linkage clustering (the partitions) and complete linkage clustering (the families).

A set of ORF products in *S. cerevisiae* defines a 'partition' if, and only if, the three following properties are simultaneously verified: (i) each member of the set has at least one significant blastp match with one other member of the same set, (ii) no member of the set has significant blastp matches with members not included in the set, and (iii) the set cannot be partitioned into subsets verifying (i) and (ii) (i.e. the set is minimal). Note that an ORF product that has no significant match fulfills these properties and is, therefore, considered as a 'single member partition' or 'singleton'. Partitions are denoted P*n.m*, where *n* is the size of the partition (the number of distinct ORF products it includes) and *m* is the order of the partition (for example P2.5 denotes the fifth partition of two elements). Whenever possible, partitions were subdivided into 'families'. A family includes the maximum number of 'reciprocally similar' members of the partition (i.e. for each pair A,B: A finds B with a similarity index above the set threshold and B finds A with a similarity index above the set threshold). Such families are denoted f*p.q*, where *p* is the number of elements in the family and *q* its order (for example: f3.2 denotes a subset of three members, which are 'reciprocally similar').

3. Results and discussion

3.1. ORFs of the S. cerevisiae genome

Application of the strategy explained in Section 2 results in a catalog of 6213 ORFs. This catalog does not contain ORFs from Ty retroelements or subtelomeric Y elements. Annotation of short ORFs was taken from MIPS after elimination of a series of them, which overlap other elements (usually LTRs) or actually correspond to tRNA gene. Similarly, intron-containing ORFs are usually annotated as in MIPS with exceptions (see Section 2). This list was subsequently crosschecked with YIDB [34]. Otherwise, our catalog is extensive. It contains all ORFs, be they independent or partially overlapping with others. In some extreme cases, the overlap covers the major part of one of the two ORFs (e.g. *YBR089w*, *YDL228c* or *YLR338w*), making it highly questionable. Note, however, that no ORF entirely included in a longer one is listed in our catalog with one exception, *YLR040c* which corresponds to the characterized gene *RPS21* and is the antisense of a longer ORF. This exclusion applies both for out-of-frame ORFs from the same strand as the longer one, and for the antisense.

The exhaustive character of our catalog has the consequence that some of the listed ORFs are unlikely to correspond to actual protein-coding genes. A typical example of such a situation is given by *YLR162w* which, by our criteria, is an ORF (and is listed as such in MIPS and SGD) but corresponds to an antisense sequence of part of the 25S rRNA gene.

The exhaustive character of our catalog also has the consequence that some ORFs, ignored or disregarded from original publications, do not have systematic names and are indicated with the initial working nomenclature used at MIPS during the sequencing program. If a good number of them are questionable genes, others are now demonstrated as being actual genes based on their sequence conservation with other yeast species of this work.

3.2. The tRNA genes of S. cerevisiae

Prediction of tRNA genes from the DNA sequence was made using our specific software described in Section 2. Briefly, this algorithm explores the occurrence of the A and B transcription boxes characteristic of tRNA genes [35] and examines the possibility of forming typical three-dimensional tRNA structures with the neighboring sequences. Using this procedure which is entirely predictive not comparative, a total of 274 tRNA genes are found in the yeast genome plus one in which a Ty element is inserted near the 5′ end. The list of tRNA genes confirms what was previously published by others [36].

From their sequence alignments, the 274 tRNA genes (plus the mutant gene) can be grouped in 52 distinct families in which all members are strictly identical in sequence. Some of the families with only one or two member(s) only differ from a larger family by one of a few nucleotides, suggesting that they represent recent mutations or possible sequencing errors. Considering such situations, a total of 45 distinct tRNA gene families can be recognized, 13 of which have introns. Note that all members of a family either have or do not have an intron, mixed situations are not found. Yet in four cases, different introns may be found in a given family. The entire tRNA set of *S. cerevisiae* contains 41 different anticodons.

The list of 52 distinct tRNA sequences was used to compare the sequences of the 13 other yeast species, and the homologs are listed in [27].

3.3. Novel genes within the 'intergenes' of S. cerevisiae

All intergenes of the *S. cerevisiae* genome (totaling ca. 3 Mb or 24% of the complete sequence) were compared to all RSTs as indicated in Section 2. A minimum of 58 of them show homologies to RST originating from one or several yeast species. In addition, 184 other intergenes show homologies to RSTs of *S. bayanus* var. *uvarum* only. The two sets were carefully inspected in order to determine the nature of the homologous segments. Four cases were found: (i) short genes of *S. cerevisiae* that were previously overlooked, (ii) possible extensions of annotated genes of *S. cerevisiae*, (iii) pseudogenes of *S. cerevisiae* with or without homologs in *S. cerevisiae* itself, (iv) very short alignments or short repeated motifs.

Table 1 lists the 50 genes that were discovered in our 'intergenes' from their homology with one or several other yeast species. Two of them (*YER039ca* and *YCR020wb*) were already described, respectively in SGD and TrEMBL (accession number Q9URQ5), but overlooked in our list. But all 48 others were previously unrecognized and constitute, therefore, novel genes discovered in this work. These genes have been designated according to the systematic nomenclature [37]. All but two genes are short (less than the selected limit of 99 codons), the two others are longer than this limit but contain introns. One of the short genes (*YER074wa*) contains two introns. The occurrence of introns was initially suggested to us from the alignments with the other yeast species studied in this program. Their precise location in the *S. cerevisiae* genes was subsequently defined from the consensus junction and branching point sequences. Three of the four introns (in

YBR255ca and in *YER074ca*) are conserved in the other yeast species in which homologs were found, confirming their existence. The fourth intron (in *YDR381ca*) is absent from *Kluyveromyces lactis* and the corresponding region is not found in *Saccharomyces exiguus*. The novel genes are found on 13 distinct chromosomes (not on chromosomes I, VIII and IX). The reality of the novel genes discovered is confirmed by the fact that they have homologs in distantly related yeast species and, for some of them, even in non-fungal eukaryotes. It is also confirmed for some of them by their biased codon usage (CAI values).

Table 2 lists the cases of possible extensions of previously annotated *S. cerevisiae* genes. A total of 19 extensions occur upstream of the originally proposed initiator codon. In some cases, a novel upstream initiator codon can be found if one considers the existence of an intron. In other cases, a putative upstream initiator codon, defined from the amino acid alignments, can be found in the same frame upstream of a stop

Table 1
List of novel *S. cerevisiae* genes identified from this program

Novel gene	Coordinates	Size	CAI	Hemiascomycetes
YBL029ca	164 450–164 734	94	0.125	Ss, Sk, Yl
YBL071wa[2,3,4,5]	89 973–90 221	82	0.155	Ss, Km, Pa
YBL108ca	7 605–7 733	42	0.531	Sb
YBR103ca	44 9279–44 9422	47	0.097	Sb
YBR191wa	607 107–607v181	24	0.088	Sb
YBR233wa	684 935–685 219	94	0.105	Zr, Kt
YBR255ca *	726 576–727 032	120	0.171	Sb, Se
YCL001wb	113 382–113 636	84	0.088	Ss, Zr
YCR020wb	155 035–155 271	78	0.121	Sb
YCR038wa	198 032–198 157	41	0.060	Sb
YCR097wa	293 172–293 438	88	0.088	Sb
YDL114wa	254 934–255 047	37	0.167	Sb
YDL159wa	172 183–172 314	43	0.145	Sb
YDL185ca	126 609–126 836	75	0.145	Sb
YDL240ca	22 471–22 608	45	0.162	Sb
YDL247wa	3 762–3 836	24	0.047	Sb
YDR079ca[2,5]	603 587–603 805	72	0.119	Sb, Kt, Dh, Yl
YDR320ca	1 108 272–1 108 490	72	0.091	Km
YDR379ca[4]	1 233 268–1 233 507	79	0.161	Kl, Pa
YDR381ca *	1 238 302–1 238 840	114	0.128	Se, Kl
YER039ca	229 262–229 480	72	0.107	Sb, Se, Ss
*YER074wa***[2,6]	307 649–308 119	85	0.089	Sb, Km
YFR012wa	169 215–169 301	28	0.179	Sb
YGL258wa[1]	9 162–9 395	77	0.102	Sb
YGR161wa	810 221–810 499	92	0.087	Sb
YGR271ca	1 037 796–1 037 987	63	0.142	Sb
YJL012ca	410 923–411 120	65	0.183	Sb
YJL047ca	348 668–348 802	44	0.087	Sb
YJL052ca	337 583–337 699	39	0.179	Sb
YJL062wa	316 419–316 676	85	0.105	Sb, Km, Dh
YJL127wa	179 892–180 008	38	0.047	Sb
YJL156wa	126 301–126 522	73	0.072	Sb
YKL003wa	437 416–437 535	39	0.072	Sb
YKL018ca	403 218–403 517	99	0.050	Sb, Se
YKL106ca	236 790–236 909	39	0.049	Sb
YKL165ca	135 792–136 025	77	0.084	Sb
YLR099wa	341 326–341 589	87	0.072	Sb
YLR149ca	440 371–440 457	28	0.236	Sb
YLR363wa	853 461–853 718	85	0.282	Ss, Zr
YML007ca	253 162–253 272	36	0.100	Sb
YMR013wa	298 310–298 390	26	0.086	Sb
YNL067wa	498 535–498 681	48	0.119	Sb
YNL162wa	330 326–330 544	72	0.129	Km
YNR001wa	631 260–631 478	72	0.082	Sb
YOL086wa	159 172–159 444	90	0.154	Sb
YOL159ca	15 232–15 504	90	0.123	Sb
YOR008wb	343 928–344 029	33	0.045	Sb
YOR314wa	904 450–904 560	36	0.117	Sb
YPR016wa	593 091–593 351	86	0.112	Sb
YPR074wa	695 013–695 183	56	0.074	Sb

The table indicates, for each novel gene identified, its proposed designation according to the systematic nomenclature (column 1), its coordinates (column 2), the size of its translation product (column 3) and its codon adaptation index (column 4). The list of yeast species having homologs to this gene is indicated under the heading 'Hemiascomycetes' and designated as follows: Ct: *Candida tropicalis*; Dh: *Debaryomyces hansenii* var. *hansenii*; Kl: *K. lactis*; Km: *Kluyveromyces marxianus* var. marxianus; Kt: *Kluyveromyces thermotolerans*; Pa: *Pichia angusta*; Ps: *Pichia sorbitophila*; Sb: *Saccharomyces bayanus* var. *uvarum*; Se: *S. exiguus*; Sk: *Saccharomyces kluyveri*; Ss: *Saccharomyces servazzii*; Yl: *Yarrowia lypolitica* and Zr: *Zygosaccharomyces rouxii*. Note that two of the 'novel' genes identified were already listed in SGD (http://genome-www.stanford.edu/cgi-bin/dbrun/SacchDB) (*YER039ca*) and TrEMBL (*YCR020wb*). All others are new from this work. Some of the genes have one (*) or two (**) intron(s). Some genes have additional orthologs in *Candida albicans* (1), *Schizosaccharomyces pombe* (2), *Arabidopsis thaliana* (3), *C. elegans* (4), *Drosophila melanogaster* (5) or *Homo sapiens* (6). CAI values were calculated according to [47].

Table 2
List of possible extensions to previously identified *S. cerevisiae* genes

Extension of the gene	Coordinates	Length	Paralog	Hemiascomycetes
*YBL091ca**(N)	47 174	76		Sb, Km
YBL104c(N)	22 255	62		Kt
YBR041w(C)	320 236	46		Sk
YCL069w(N)	9 498	69	*YPR198w*	Ct
YDR179wa(N)	819 428	194		Sk
YDR475c(C), *YDR474c*(N)	1 407 454–1 410 082	165	*YOR019w*	Sb, Ss, Kl
YER066w(N)	289 640	301	*YFL009w*	Sb, Kl
YGL059w(C)	393 697	46		Sr, Kl
YGL183c(N)	157 206	45		Kl
YGL196w(N)–(C)	130 135–131 172	139–140		Sk, Zr
*YHR079ca**(C)	262 193	45		Sb, Kl
YHR176w(C)	455 527	60		Sb
YJL160c(C)	117 957	107	*YJL158w*	Sb, Ss, Zr, Sk, Kl, Pa, Ps
YJL213w(N)	31 821	114		Sb
YJR013w(N)	460 069	98		Pa, Ps
YKR058w(N)	552 421	131	*YJL137c*	Sb, Kt
YLR054c(N)	250 751	165		Sb, Kt
YMR207c(N)	683 686	41	*YNR016c*	Kt
YMR269w(N)	804 455	70		Zr, Kl
YNL083w(C)	473 008	50		Sk, Dh
YNR062c(N)	745 778	147		Kl
YOL048c(N)	241 126	201		Sb, Zr
YOL163w(N)	9 249	115	*YIL166c*	Sb,Pa,Ct
YOR298ca(N)	877 678	92		Se
*YPR098c**(N)	729 525	53		Sk

The table indicates, for each possible gene extension identified (column 1), the location of the extension (N: N-terminal extension upstream of the previously annotated initiator codon; C: C-terminal extension downstream of the stop codon), the coordinate of the maximal extension (column 2) and its corresponding amino acid length (column 3). The existence of paralogs in *S. cerevisiae* is indicated in column 4 and the list of yeast species having homologs to this gene is indicated in column 5 (same abbreviations as in Table 1). *Presence of an intron.

codon or in one of the two other frames, suggesting a possible sequencing error or a natural in-frame stop codon or frameshift. There remained a few cases in which no upstream initiator codon could be proposed for lack of homology in the relevant segment. A total of eight extensions occur downstream of the originally proposed stop codon of the annotated *S. cerevisiae* gene. In those cases, amino acid homology is found if one assumes a frameshift in the *S. cerevisiae* sequence bypassing the previously defined stop codon. Note that, in some cases, the proposed extension of the *S. cerevisiae* gene is only tentative because that gene is a member of a family whose other members are longer and also possess homology with the other yeast species. Finally, we observed one case in which two neighboring genes (*YDR475c* and *YDR474c*) are extended such as to produce a single fused gene.

We also observed a few cases of *S. cerevisiae* sequences not previously annotated as ORFs but showing three or more segments of homology with other yeast species. Such sequences may correspond to pseudogenes in *S. cerevisiae* containing two or more frameshifts. Some of them have homologs in *S. cerevisiae* itself. For all other intergenic sequences analyzed, alignments are too short to be conclusive or correspond to simple motifs tandemly repeated leading to apparent homology.

3.4. Partitions and gene families in S. cerevisiae

Gene redundancy is a common feature to nearly all organisms sequenced so far, including those with the shortest genomes. The short genomes of bacteria such as *Mycoplasma genitalium* [38] or *Rickettsia prowazeckii* [39] tend to be globally less redundant than those of *Bacillus subtilis* [40], *Escherichia coli* [41] or *Mycobacterium tuberculosis* [42], but still contain families of paralogous genes. In eukaryotes, the genome of the multicellular organism, *Caenorhabditis elegans* [43], is more redundant than that of the yeast *S. cerevisiae* which tends to be globally less redundant than that of some bacteria with large genomes. The conservation or the evolution of gene redundancy among the hemiascomycetous yeasts will be examined elsewhere [28]. In this article, we only report the classification of the *S. cerevisiae* genes that was used throughout this program. This classification can be found at http://cbi.genopole-bordeaux.fr/genolevures.

A total of 2458 ORFs are included. All other ORFs are unique in the *S. cerevisiae* genome. The list includes 457 partitions of two members, 124 of three members, 53 of four members, 27 of five members, 13 of six members, 9 of seven members, 7 of eight members, 4 of nine members, and 28 partitions of 10 members or more (11 partitions include over 20 members and one partition includes 108 members). This classification is based upon sequence comparisons of the predicted gene products, as explained in Section 2. It simultaneously involves the single-linkage clustering method and the complete linkage clustering method. The first is used to generate 'partitions', the second to identify 'families' within partitions. Note, however, that many 'partitions', especially those with few members, contain a single 'family' because all of their members show complete sequence relationship (complete linkage). An example of such a case is P5.24.f5.1, a partition of five members with complete internal linkage relationship (i.e. each of the five members shows, when used as query sequence, significant homology with all four members). In other cases, the partition may be composed of several families. For example P5.21, a partition of five genes, is composed of a first subset of three genes (P5.21.f3.1, containing *YGL014w*, *YGL178w* and *YLL013c*) and a second subset of two genes (P5.21.f2.1, containing *YJR091c* and *YPR042c*). In yet other

cases, one or a few ORF(s) within a partition shows homology (sometimes of borderline significance) to only a subset of the members of a family. Such cases are denoted as 'weak links'.

As all blast-based clustering methods, our classification becomes problematic for gene products showing a degree of similarity close to the selected threshold. Thus the exact limit of some partitions, families, or especially weak links may be subject to criticism. Yet, the classification used has the advantage that, in general, our partitions include all ORF products having a common ancestry in a given organism. But the single-linkage clustering on which they are based may result in a partition containing several ORFs of unrelated ancestral origin if, for example, gene fusion has occurred between at least two of the members, creating a complex gene with domains from different ancestry. An example of such a situation is given by the hydroxy-methyl-pyrimidine phosphate kinase gene family recently studied by some of us [44]. This family contains two active genes *YOL055c* and *YPL258c* and a so far an inactive one, *YPR121w*, all three of them encoding proteins with two functional domains based on direct experimental evidence as well as on sequence comparisons with bacterial gene products. In *S. cerevisiae*, the N-terminal domain, which contains the enzymatic activity, is found in the products of the three genes mentioned, while the C-terminal domain is also represented in the product of a fourth independent gene, *YCR020c* (*PET18*), considered below the limit of significance. A number of clear-cut cases of ancestral gene fusions are documented in *S. cerevisiae* ([45] and B. Labedan, personal communication).

Using our classification, partitions in *S. cerevisiae* range from two members (the most frequent class) to 108 members. A total of 722 partitions containing two or more members are found. These include 2458 ORFs or nearly 40% of the ORFs listed in http://cbi.genopole-bordeaux.fr/genolevures. A number of the two-member partitions correspond to genes encoding ribosomal protein or other translation factors. Some large-size partitions correspond to transporter genes or to genes from subtelomeric regions. Some small-size partitions correspond to the ancestral chromosome duplication blocks or include members from such blocks [46].

4. Conclusions

Using the strategy described in this article, which combines explicit analysis of the *S. cerevisiae* genome sequence and comparisons with the other yeast species, we have discovered 50 novel genes of *S. cerevisiae* that, apparently, had escaped attention, and updated the list of predicted ORFs that are, on the contrary, unlikely ones. The revisited *S. cerevisiae* genome contains at least 5651 protein-coding genes.

Acknowledgements: We thank the members of the Unité de Génétique moléculaire des levures for fruitful discussions and H. Feldmann and A. Goffeau for careful reading of the manuscript. B.D. is a member of Institut Universitaire de France.

References

[1] Oliver, S. et al. (1992) Nature 357, 38–46.
[2] Dujon, B. et al. (1994) Nature 369, 371–378.
[3] Johnston, M. et al. (1994) Science 265, 2077–2082.
[4] Feldmann, H. et al. (1994) EMBO J. 13, 5795–5809.
[5] Muramaki, Y. et al. (1995) Nature Genet. 10, 261–268.
[6] Bussey, H. et al. (1995) Proc. Natl. Acad. Sci. USA 92, 3809–3813.
[7] Galibert, F. et al. (1996) EMBO J. 15, 2031–2049.
[8] Jacq, C. (1997) Nature 387 (Suppl.), 75–78.
[9] Dietrich, F.S. et al. (1997) Nature 387 (Suppl.), 75–81.
[10] Tettelin, H. et al. (1997) Nature 387 (Suppl.), 81–84.
[11] Churcher, C. et al. (1997) Nature 387 (Suppl.), 84–87.
[12] Johnston, M. et al. (1997) Nature 387 (Suppl.), 87–90.
[13] Bowman, S. et al. (1997) Nature 387 (Suppl.), 90–93.
[14] Philippsen, P. et al. (1997) Nature 387 (Suppl.), 93–98.
[15] Dujon, B. et al. (1997) Nature 387 (Suppl.), 98–102.
[16] Bussey, H. et al. (1997) Nature 387 (Suppl.), 103–105.
[17] Foury, F., Roganti, T., Lecrenier, N. and Purnelle, B. (1998) FEBS Lett. 325, 325–331.
[18] Kalogeropoulos, A. (1995) Yeast 11, 555–565.
[19] Dujon, B. (1996) Trends Genet. 12, 263–270.
[20] Fisher, D. and Eisenberg, D. (1999) Bioinformatics 15, 759–762.
[21] Mackiewicz, P., Kowalczuk, M., Gierlik, A., Cudek, M.R. and Cebrat, S. (1999) Nucleic Acids Res. 27, 3503–3509.
[22] Andreade, M.A. et al. (1997) Yeast 13, 1363–1374.
[23] Olivas, W., Muhlad, D. and Parker, R. (1997) Nucleic Acids Res. 25, 4619–4625.
[24] Basrai, M., Hieter, P. and Boeke, J.D. (1997) Genome Res. 7, 768–771.
[25] Velculescu, V.E. et al. (1997) Cell 88, 243–251.
[26] Chu, S. et al. (1998) Science 282, 699–705.
[27] Souciet, J.L. et al. 487, 3–12 (this issue).
[28] Llorente, B. et al. 487, 71–75 (this issue).
[29] Tekaia, F. et al. 487, 17–30 (this issue).
[30] Altschul, S.F. et al. (1997) Nucleic Acids Res. 25, 3389–3402.
[31] Altschul, S.F. (1991) J. Mol. Biol. 219, 555–565.
[32] Wootton, J.C. and Federhen, S. (1993) Comput. Chem. 17, 149–163.
[33] Tekaia, F. and Dujon, B. (1999) J. Mol. Evol. 49, 591–600.
[34] Lopez, P.J. and Seraphin, B. (2000) Nucleic Acids Res. 28, 85–86.
[35] Schultz, P. et al. (1989) EMBO J. 8, 3815–3824.
[36] Hani, J. and Feldmann, H. (1998) Nucleic Acids Res. 26, 689–696.
[37] Mewes, W. et al. (1997) Nature 387 (Suppl.), 7–65.
[38] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D. and Clayton, R.A. et al. (1995) Science 270, 397–403.
[39] Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T. and Alsmark, U.C. et al. (1998) Nature 396, 133–140.
[40] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M. and Alloni, G. et al. (1997) Nature 390, 249–256.
[41] Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T. and Burland, V. et al. (1997) Science 277, 1453–1462.
[42] Cole, S.T., Brosch, R., Parkhill, J., Garnier, T. and Churcher, C. et al. (1998) Nature 393, 537–544.
[43] The *C. elegans* Sequencing Consortium, (1998) Science 282, 2012–2018.
[44] Llorente, B., Fairhead, C. and Dujon, B. (1999) Mol. Microbiol. 32, 1140–1152.
[45] Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Nature 402, 86–90.
[46] Llorente, B. et al. 487, 101–112 (this issue).
[47] Sharp, P.M. and Li, W.H. (1987) Nucleic Acid Research 15, 1281–1295.