1816–1835 *Nucleic Acids Research, 2006, Vol. 34, No. 6*
doi:10.1093/nar/gkl085

The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications

Christian Marck*, Rym Kachouri-Lafond[1], Ingrid Lafontaine[2], Eric Westhof[1], Bernard Dujon[2] and Henri Grosjean[3]

Service de Biochimie et de Génétique Moléculaire, Bât 144. CEA/Saclay, 91191 Gif-sur-Yvette, France, [1]Institut de Biologie Moléculaire et Cellulaire, Centre National de la Recherche Scientifique, UPR 9002 'Architecture et Réactivité de l'ARN', Université Louis Pasteur, 15 rue Descartes, 67084 Strasbourg, France, [2]Unité de Génétique Moléculaire des Levures, Institut Pasteur, 25 rue du Dr Roux, 75724 Paris, France and [3]Laboratoire d'Enzymologie et Biochimie Structurales, Bât 34. Centre National de la Recherche Scientifique, 1 av. de la Terrasse, 91198 Gif-sur-Yvette, France

Received December 7, 2005; Revised February 3, 2006; Accepted March 3, 2006

ABSTRACT

We present the first comprehensive analysis of RNA polymerase III (Pol III) transcribed genes in ten yeast genomes. This set includes all tRNA genes (tDNA) and genes coding for SNR6 (U6), SNR52, SCR1 and RPR1 RNA in the nine hemiascomycetes *Saccharomyces cerevisiae*, *Saccharomyces castellii*, *Candida glabrata*, *Kluyveromyces waltii*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Debaryomyces hansenii*, *Candida albicans*, *Yarrowia lipolytica* and the archiascomycete *Schizosaccharomyces pombe*. We systematically analysed sequence specificities of tRNA genes, polymorphism, variability of introns, gene redundancy and gene clustering. Analysis of decoding strategies showed that yeasts close to *S.cerevisiae* use bacterial decoding rules to read the Leu CUN and Arg CGN codons, in contrast to all other known Eukaryotes. In *D.hansenii* and *C.albicans*, we identified a novel tDNA-Leu (AAG), reading the Leu CUU/CUC/CUA codons with an unusual G at position 32. A systematic 'p-distance tree' using the 60 variable positions of the tRNA molecule revealed that most tDNAs cluster into amino acid-specific sub-trees, suggesting that, within hemiascomycetes, orthologous tDNAs are more closely related than paralogs. We finally determined the bipartite A- and B-box sequences recognized by TFIIIC. These minimal sequences are nearly conserved throughout hemiascomycetes and were satisfactorily retrieved at appropriate locations in other Pol III genes.

INTRODUCTION

In eukaryotes, RNA polymerase III (Pol III) transcribes a few hundreds short non-coding RNA genes, the bulk of which are the transfer RNA genes (tDNA) and the 5S RNA genes (1,2). In *Saccharomyces cerevisiae*, a few other non-coding RNAs are also synthesized by Pol III: (i) SNR6 which is the U6 RNA component of the spliceosome (3); (ii) RPR1, the RNA component of ribonuclease P (4) and (iii) SCR1, the RNA component of the signal recognition particle (SRP) (5). Recently, genome wide investigation on Pol III transcription machinery occupancy in *S.cerevisiae* showed that all known Pol III genes are occupied and revealed new potential Pol III genes. These are *SNR52* (6), a C/D snoRNA responsible for the 2′-*O*-methylation of small subunit rRNA at A_{420} (7,8) which was previously considered to be a Pol II product; and *ZOD1*, whose function remains unkwown (9).

The transcription of all Pol III genes is dependent on two general transcription factors: the assembling factor TFIIIC and the recruiting factor TFIIIB (10). The 5S RNA genes require a specific recognition factor, TFIIIA (11), while all other Pol III genes are first recognized by TFIIIC (which also recognizes the 5S RNA-TFIIIA complex). After binding to DNA, TFIIIC directs the upstream binding of TFIIIB, thus forming a

*To whom correspondence should be addressed. Tel: 33 (0)1 69 08 46 20; Fax: 33 (0)1 69 08 47 12; Email: christian.marck@cea.fr

pre-initiation complex able of recruiting Pol III for multiple cycles of transcription (12,13). Among the three eukaryotic RNA polymerases, Pol III is the only one featuring a reinitiation mechanism that makes the production of Pol III transcripts extremely efficient (13,14). Pol III terminates transcription at short tracks of T's (preferentially followed by A or G) in the RNA-like strand (15). Unlike Archaea and Bacteria, the use of sequences upstream of eukaryotic tRNA (and other Pol III) genes as promotor elements is rare, as demonstrated in *S.cerevisiae* (16–18). For review on tRNA genes and Pol III transcription, see also (2,10,12).

For all Pol III genes, but the 5S RNA gene, the primary recognition by TFIIIC relies on two short promoter sequences, which are generally internal to the genes. In tRNA, the nucleotides implicated are those that make up the universally conserved tertiary base pairs bridging the D- and T-loops. At the genomic DNA level, these nucleotides are the two variably distant linear promoter regions recognized by TFIIIC (traditionally referred to as A- and B-boxes) (19,20). Early definitions of the A- and B-box consensus sequences [TGGCnnAGTGG and GGTTCGAnnCC, respectively (19)] appear now too restrictive as more sequences become available. Updated consensus have been later proposed for the A-box; all terminate with (or extend beyond) the two universally conserved nucleotides G_{18} and G_{19} [see, e.g. (5)]. The A- and B-promoter sequences must also be present in other Pol III genes, but no accurate definition and genome wide compilation of these promoter sequences were presented yet.

A recent work, based on the comparative analysis of genomes showed that tRNA genes from Eukaryotes, Archaea and Bacteria display both common and domain-specific features (21). However, only two yeasts (*S.cerevisiae* and *Schizosaccharomyces pombe*) among seven eukaryotes were available. The large number of hemiascomycetous genomes now sequenced (22,23) offers the opportunity to perform a detailed comparative genomics of Pol III genes. With compact genomes (less than 20 Mb) these organisms give access to a wide evolutionary range, even larger than that of Chordates if one considers the phylogenetic distance between *S.cerevisiae* and *Yarrowia lipolytica* (24). Pol III genes were analysed in nine yeast species across the evolutionary tree of hemiascomycetes: *S.cerevisiae* (25), *Saccharomyces castellii* (26) which is now placed in the *Naumovia* clade (27) close to the *Saccharomyces*, *Candida glabrata* (24), *Kluyveromyces waltii* (28), *Kluyveromyces lactis* (24), *Eremothecium gossypii* (29), *Debaryomyces hansenii* (24), *Candida albicans* (30) and *Yarrowia lipolytica* (24). The archiascomycete *S.pombe* (31) was used as an outgroup.

Over 2300 Pol III genes were extracted from these ten yeast genomes. The majority of them are the tRNA genes (a detailed list of the 2335 tRNA genes is available as Supplementary Data). Whether these tDNAs from the ten yeast genomes obey the rules previously defined for eukaryotic tDNA was tested. Several sequence deviations to the cloverleaf tRNA model that may possibly affect the tertiary structure of some tRNAs were discovered. Peculiarities in the decoding of leucine and arginine codons, previously seen in *S.cerevisiae* only, are extended to related yeasts. Eight of the genomes harbour head-to-tail tDNA pairs, with a maximum of 17 cases in *D.hansenii*. We also have performed a global distance analysis over all tDNA sequences. Despite the short length of tDNA (hence presumed low informational content), the results suggest a common phylogenetic origin inside each amino acid-specific family, confirm a case of tRNA capture and suggest a novel one. Finally, from the compilation of all tDNA data, the sequences TRGYnnAnnnG (11 nt) and GWTCRAnnC (9 nt) were derived as the hemiascomycetous signatures of the Pol III transcriptional promoters for tDNA. These identity elements are also found at appropriate locations in the other Pol III RNA genes *SNR6*, *SNR52*, *RPR1* and *SCR1*.

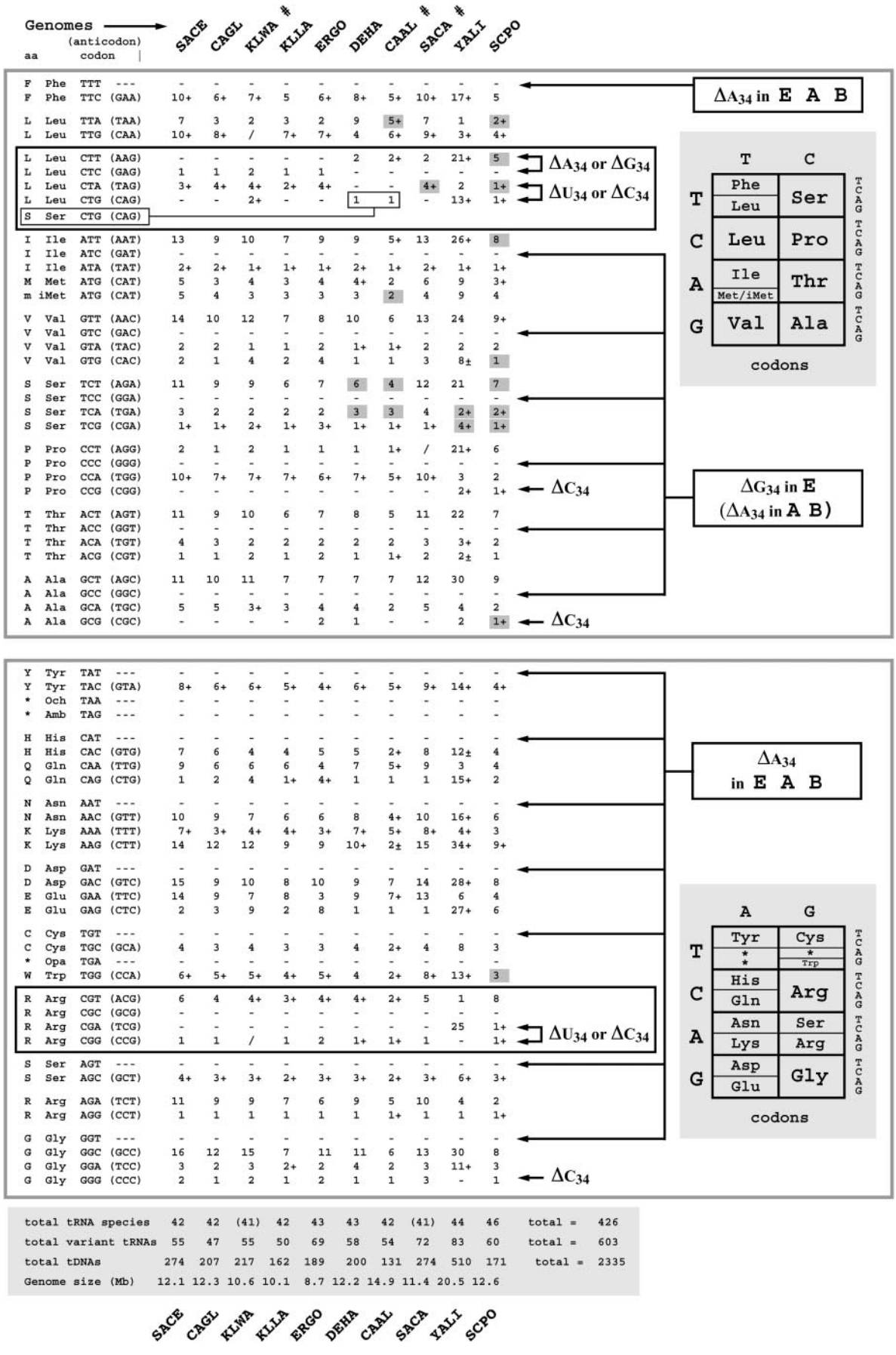
MATERIALS AND METHODS

The ten genomes investigated are listed in Supplementary Table 1; all but the archiascomycete *S.pombe* belong to hemiascomycetes. Genomes are also referred to with a four-letter acronym made of the two first letters of the gender name followed by the two first letters of the species name (e.g. SACE stands for *Saccharomyces cerevisiae*). The tRNA genes and the tRNA are designated as in this example: 'tDNA-Leu (TAG)' and 'tRNA-Leu (UAG)', respectively. The anticodons are always written between brackets with nt 34, indicated or not, in the first position. The conventional IUB/IUPAC degenerate DNA alphabet (32) and special symbols used for base pairings combinations are defined in the legend to Figure 2; 'n' is often used, instead of 'N', for clarity. The universal conventional numbering system for tRNA positions is that adopted in the tRNA database (33). The sequences of all tDNA identified in the ten genomes are given in the Supplementary Table 4.

Search for tRNA genes

The full set of nuclear tRNA genes were searched in each genome using the procedure described earlier (21). This search method is based on the detection in a given genome of the nucleotide sequences corresponding to the eukaryotic-type conserved nuclear tRNA cloverleaf structure [Figure 2, see also Supplementary Table 4 in (21)]. However, in the case of *Y.lipolytica* (acronym YALI), this procedure failed to reveal a number of tDNA, otherwise correctly detected by tRNAscan-SE (34). These tDNA contained an unexpected number of GT pairs (in tDNA, GU in tRNA stems) and/or Watson–Crick mismatched pairs within the stems of the cloverleaf structure. Our initial search parameters were therefore adapted (for this particular genome only) as follows: number of GT pairs allowed in the anticodon stem: three (instead of two); total number of mismatches in the four stem: three (instead of two); total number of GT and mismatched pairs: six (instead of five) (Figure 2).

Two possible pseudogenes were identified in *Y.lipolytica*: one encoding tRNA-Ala (anticodon AGC) (cove score 61.36) which differs from the other 29 copies by a T instead of a G at position 63, thus creating a second mismatched pair in the T-stem; the second encoding tRNA-Leu (AAG) (cove score 51.86) which, among 21 copies, has a T instead of G at position 19, thus creating a mismatch in place of the usual $G_{19}C_{56}$ tertiary base pair. The functionality of these two gene products is therefore questionable. In the genome of *K.waltii* (KLWA), a mitochondrial origin was suspected for 13 single copy tDNAs for the following reasons: (i) these tDNAs were located


Genomes
(anticodon)
aa codon
SACE CAGL KLWA # KLLA ERGO DEHA CAAL # SACA # YALI SCPO
ΔA34 in E A B
ΔA34 or ΔG34
ΔU34 or ΔC34
ΔC34
ΔG34 in E
(ΔA34 in A B)
ΔA34
in E A B
ΔU34 or ΔC34
ΔC34
T C A G
Phe Leu Ser
Leu Pro
Ile Met/iMet Thr
Val Ala
codons
A G
Tyr * * Cys * Trp
His Gln Arg
Asn Lys Ser Arg
Asp Glu Gly
codons
total tRNA species 42 42 (41) 42 43 43 42 (41) 44 46 total = 426
total variant tRNAs 55 47 55 50 69 58 54 72 83 60 total = 603
total tDNAs 274 207 217 162 189 200 131 274 510 171 total = 2335
Genome size (Mb) 12.1 12.3 10.6 10.1 8.7 12.2 14.9 11.4 20.5 12.6
SACE CAGL KLWA KLLA ERGO DEHA CAAL SACA YALI SCPO

in three short contigs (G194contig_278, G194contig_341 and G194contig_362); (ii) these three contigs display continuous low GC content (about 20% compared to 44% for the total of all contigs); (iii) each of these 13 single copy tDNA was markedly different from other bona fide nuclear and multiple copy tDNA bearing the same anticodon; (iv) Blast search of these three contigs revealed high scores with the mitochondrial genome of the close species *K.lactis*. These three low GC content contigs were therefore considered as actual fragments of the mitochondrial genome of *K.waltii* and not as ancient permanent inclusions of its mitochondrial genome into the nuclear genome.

In *K.waltii* (KLWA), the genes encoding tRNA-Leu (CAA, decoding the UUG codon) and tRNA-Arg (CCG, decoding the CGG codon) were not identified (in Figure 1, these missing genes are indicated by a '/' sign). In *S.castellii* (SACA), tRNA-Pro (AGG, decoding the CCU and CCC codons) is also missing. For these two genomes, (as well as for *C.albicans* (CAAL), the genomic sequence is not complete.

p-Distance analysis of tRNA genes sequences

In order to align perfectly the sequences, introns (if any, located between nt 37 and 38), the base 47 (not always present) and the V-arm extension (from positions 47 to 48, present only in Leu and Ser isoacceptors) were removed. All sequence variations due to the polymorphism of some genes (e.g. a GC to AT base pair change in a stem) and not located in the eliminated regions listed above were selected for the p-distance analysis (some examples are given in Figure 4A). Only one tDNA copy was retained per family of strictly identical sequences and this lead to a total of 603 different sequences out of a total of 2335 tDNA sequences examined in this work. This number represents an intermediate between the total number of different types of tRNA for all ten genomes (426 tRNA/anticodon types) and the total number of tRNA genes (2335 genes). The comparative analysis of the 603 sequences required the computation of 181 503 pairwise p-distance values. The p-distance is defined as the number of nucleotide sites, which are different between any pair of sequences compared, divided by the total number of common nucleotides. The largest p-distance we observed was that between tDNA-Leu (AAG) from *Y.lipolytica* and tDNA-Glu (CTC) from *D.hansenii* (54 positions different out of 75). A histogram of the p-distance is shown in Figure 4B. The p-distance tree presented in Figure 4C was built by the Neighbor-Joining method (35) implemented within the MEGA2 software (36).

Search for A- and B-box promoter sequences in ncRNA Pol III genes other than tDNAs

Four ncRNA Pol III genes were also considered: *SNR6* (U6), *SNR52*, *RPR1* and *SCR1*. For *S.cerevisiae*, the boundaries of the mature products of these four genes were taken from the gene definition in SGD (URL's given in Supplementary Data). These four Pol III genes were identified in the four recently sequenced genomes (*C.glabrata*, *K.lactis*, *D.hansenii* and *Y.lipolytica*) as follows: *SNR6*: this gene was previously identified by a BlastN search (24); *SNR52*: following a genomic BlastN search run on each genome with the *S.cerevisiae* gene as entry. In other genomes, *SNR6* and *SNR52* genes were identified with BlastN when not annotated. *RPR1*: this gene was recently identified in the ten genomes explored and more (4). *SCR1*: this gene, previously identified in *C.glabrata*, *K.lactis*, *D.hansenii* and *Y.lipolytica* on the basis of a structural identity with the *S.cerevisiae* genes (see Supplementary Table S6 in (24)), was identified thanks to the conservation of the P6 and P8 helices (for nomenclature, see (37)). All RPR1 or SCR1 RNAs from the ten genomes were structurally aligned (Figure 7) and the boundaries of the mature products deduced from those of the *S.cerevisiae* mature RNAs.

The A- and B-boxes of all four genes from *S.cerevisiae* had already been identified and verified experimentally: *SNR6* (3); *SNR52* (6); *RPR1*, (38); *SCR1*, (5). A- and B-boxes of these four genes were searched in other genomes in and around the mature product sequences using the consensus TRGYnnAnnnG and GWTCRAnnC as determined from tDNA sequences analysis (this work, Figure 5E). The putative A-box of *SCR1* genes was located eight bases downstream that previously proposed (5). Detailed data about A- and B-boxes of these four ncRNA genes are presented in Figure 6.

RESULTS

Identification of tRNA genes: unusual sequence features and polymorphism

Using a combination of cloverleaf structure detection algorithm (21) and tRNAscan-SE (34), 2335 genes encoding nuclear tRNA molecules (tDNA) were identified from genomic sequences of ten yeast species (listed in Supplementary Data). Among them, 47 different anticodons were identified. The numbers of genes encoding each isoacceptor are given in Figure 1. Note the significant variations between species.

With a few exceptions, all tRNAs obey the canonical eukaryotic cloverleaf model (Figure 2) initially established

Figure 1. tRNA/anticodon and tDNA usages in the ten genomes. The upper and lower panels correspond to the left and right part of the conventional genetic code tabulation (upper and lower insets at right, respectively). The ten genomes are designated by their acronyms; the '#' signs following some acronyms denote genomes of low coverage or without full assembled chromosomes (uncomplete genomes). The 64 regular codons and anticodons, plus the initiator mehionine (iMet), are listed vertically. Anticodons which are never used in tRNAs (in any of the three domains of life) are replaced by '—'. Numbers in the main array indicate the number of tRNA genes present per genome for the given anticodon; '-' signs stand for no gene; '/' signs indicate that one copy (at least) of the tDNA is probably present in the actual genome but could not be identified in the available sequences. The '+' signs report tRNA genes that harbour an intron and '±' those in which the intron is absent from some copies. Grey background indicates tDNAs that exhibit deviation from the consensus sequences defined for the A-box at G_{10}; open boxes indicate other exceptions (see Figure 5C). The indications 'ΔA', 'ΔG', 'ΔU' and 'ΔC' emphasize the lack of tRNA bearing anticodon starting with the indicated nucleotide (first base of anticodon). The 'A or G' sparing rule ('ΔG' or 'ΔA') as well as 'ΔU' and 'ΔC' rules are summarized at right with 'E', 'A' and 'B' to indicate whether the rule applies to each of the three domains of life (Eukaryotes, Archaea and Bacteria, respectively). Black boxes emphasize the decoding of leucine and arginine (see Figure 3 for details) in which double arrows denote anticodons of variable usage among the ten genomes. In the grey rectangle at bottom are given the number of tRNA species per genome (number of differents anticodons with initiator and elongator tRNA-Met considered as different), the number of variant tRNA genes, the total number of tRNA genes per genome and the genome size expressed in Mb. For *K.waltii* (KLWA) and *S.castellii* (SACA), values of 41 different tRNA (in brackets) are underestimated due to the uncompleteness of sequence data. These species probably harbour 43 and 42 tRNA, respectively.

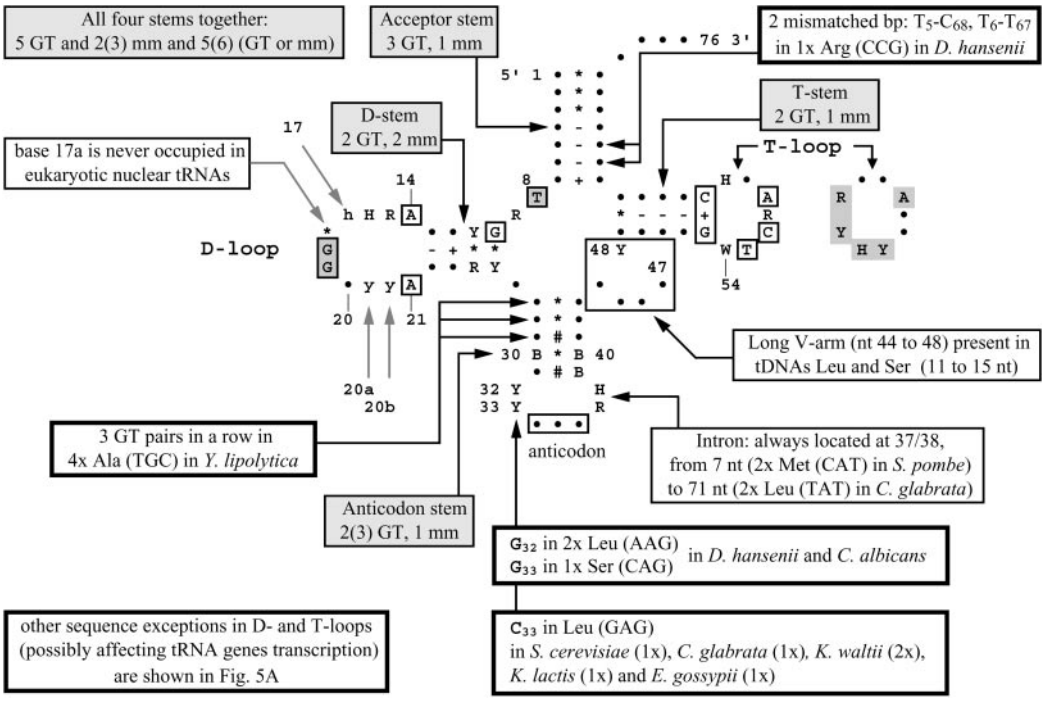


Figure 2. Features and exceptions in the tRNA cloverleaf model from the ten genomes. The sequence presented under the cloverleaf (2D) model is the consensus sequence of the 274 tRNA genes from *S.cerevisiae*. The conventional IUB/IUPAC degenerate DNA (or RNA with T changed for U) alphabet (32) is used in this and following figures: **R** (pu**r**ine), A or G; **Y** (p**y**rimidine), C or T; **S** (**s**trong), G or C; **W** (**w**eak), A or T; **M** (a**m**ino), A or C; **K** (**k**eto), G or T, **B** (not A), C, G or T; **D** (not C), A, G or T ; **H** (not G), A, C or T; **V** (not T), A, C or G; **N** (a**n**y), A, C, G or T; small dots also indicate n (any nucleotide). The key to base pairing symbols is: '+',Watson–Crick base pairing only; '*',Watson–Crick pairing or GT/TG pairing; '#', Watson–Crick pairing or mismatch; '-', Watson–Crick pairing or GT/TG pairing or mismatch. The arrows indicate the four variable positions of the D-loop which are (or not) occupied (indicated with lower case letters). Boxed nucleotides are those conserved in nearly all tDNAs; grey background indicates nucleotides requested by the eukaryotic cloverleaf model used to search tRNA genes in the genomes (21) (for clarity, the T-loop is drawn twice). The details of this model are indicated in the surounding grey-backgrounded boxes (one for each stem and one at top left applying to all four stems together). In these boxes, the maximum number of GT or TG base pairs (GU or UG in mature tRNA) is indicated as 'GT' and the maximum number of non Watson–Crick base pairs (mismatched) are indicated as 'mm'. Values in parenthesis were used only for tDNA search in *Y.lipolytica* (YALI). Remarkable features are indicated in the other boxes, sequence exceptions are the heavy lined boxes. More exceptions to conserved bases in the D- and T-loops, that may affect tRNA genes transcription, are shown below in Figure 5A.

by comparing tDNA sequences from *S.cerevisiae* and *S.pombe* and five other eukaryotes (21). Features specific to unique tDNAs are listed in Supplementary Table 2 and illustrated in Figure 2. For example, in *S.pombe*, the three copies of the tDNA-Ser (anticodon GCT) harbour an extra '20c' base in the D-loop. In *D.hansenii* and *C.albicans*, a special tDNA-Ser (CAG) that reads the CUG codon as Ser instead of Leu contains an unusual G at position 33 (39–41). In the same organism, the tDNA-Leu (AAG) that reads the CUU, CUC and CUA codons, contains an unusual G at position 32 (see also below and in Figure 3). A few tDNAs contains unusual accumulations of non-Watson–Crick base in the cloverleaf stems. For example, in *Y.lipolytica*, tDNA-Ala (TGC) has two mismatched base pairs and four GT base pairs, three of which are present in a row in the anticodon stem. Earlier analysis (21) showed that a maximum of two GT pairs in the anticodon stem and five GT or mismatched base pairs in all stems are present in the canonical eukaryotic cloverleaf. A near perfect conservation of the features characteristic of tDNA-iMet (21) is observed: AT pair in 1–72 (TA in *S.pombe*), positions 17, 17a, 20a and 20b unoccupied, GGGCT in 29–33 (AGGCT in *C.albicans*), and A in 54 and 60. Unusual sequence features occurring in the D- and T-loops, that may affect the transcription of tRNA genes, are discussed below.

All members of a multicopy tDNA gene family in a given yeast species (tRNAs harbouring the same anticodon) display neighbour (sometimes strictly identical) sequences (from nt 1 to 73), with the exception of two cases. This is why the total number of distinct tRNA molecules in each yeast species exceeds the number of isoacceptor tRNA ('tRNA species' and 'variant tRNAs', respectively, indicated at bottom of Figure 1). Slight sequence variations are frequent (e.g. a GC pair changed into an AT in some copies). Except in two cases, no markedly different tDNAs coding for the same amino acid were identified. The tDNA-Arg (CCG) departs from other tDNA-Arg in five of the ten genomes investigated. Also, the two copies of tDNA-Thr (CGT) of *Y.lipolytica* differ at 20 of the 75 positions of the tRNA molecule and, moreover, one has an intron (13 nt) whereas the other does not. These two cases are examined in details below.

Size and presence of introns in pre-tRNAs are highly variable

In eukaryotic pre-tRNA molecules encoded by nuclear genes, the introns, when present, are always located between 37 and 38 nt (1 nt downstream of the anticodon). One exception is the nucleomorph (remnant eukaryotic nucleus) of the cryptophyte

Guillardia theta (42) where introns (as short as 3 nt) have recently been found at non-canonical positions. The conservation of the intron position is dictated by the eukaryotic pre-tRNA-specific splicing machinery that has probably evolved from the unique ancestral machinery of Archaea (43,44) [reviewed in (45–47)]. In Bacteria, the only introns found in the tRNA genes are autocatalytic group I introns located within the anticodon loop [reviewed in (48,49)].

Among the 47 tRNA species identified in the ten yeast genomes studied, 40 display an intron in at least one species (Figure 1). tRNA-iMet (CAU) is the only isoacceptor that never bears an intron. Introns are universal to all pre-tRNA-Ile (UAU), pre-tRNA-Ser (CGA), pre-tRNA-Ser GCU) and pre-tRNA-Tyr (GUA). With 27 of the 44 pre-tRNAs containing intron, *Y.lipolytica* is the richest intron-containing yeast and eukaryotic species known to date. Next comes *C.albicans* with 24 intron-containing pre-tRNAs (over a total of 42). The remaining eight genomes display 10 to 16 pre-tRNAs harbouring introns.

Intron sizes range from 7 nt (pre-tRNA-Met (CAU) in *S.pombe*) to 71 nt (pre-tRNA-Ile (UAU) from *C.glabrata*). Sequence variation is observed among the different copies of a same tRNA species. For example, among the 27 copies of *Y.lipolytica* pre-tRNA-Glu (CUC), two copies have a 20 nt intron and the 25 other ones harbour 9 or 10 nt introns. Moreover, in some other types of pre-tRNA of *Y.lipolytica*, the intron is simply missing in some tDNA copies. For example, among the 12 copies of pre-tRNA-His (GUG) of *Y.lipolytica*, only four copies have introns of 14, 15, 16 and 22 nt, respectively and the remaining seven copies lack the intron. Such polymorphism is not reported in other Eukarya, except a single case in *Caenorhabditis elegans* [see in (21)]. Preliminary analysis of recently available *Cryptococcus neoformans* genome shows that the presence or absence of introns in the multiple copies of the same isoacceptor is more encountered than initially thought (C. Marck, unpublished data).

No obvious correlations emerge from the presence or absence of introns in the various isoacceptors. However, the universal presence of intron in pre-tRNA-Ile (UAU) and pre-tRNA-Tyr (GUA) in yeasts and in all other sequenced eukaryotes is probably related to the need of specific post-transcriptional isomerization of uridine into pseudouridine (Ψ) (50,51). This modification is catalysed by tRNA pseudouridine synthases during pre-tRNA maturation, respectively PUS1 [catalyzing the intron-dependent formation of Ψ_{34} and Ψ_{36} in pre-tRNA-Ile (52,53)] and PUS7 [catalyzing the intron-dependent formation of Ψ_{35} in pre-tRNA-Tyr (54)]. Likewise, in all yeast genomes analysed (except *D.hansenii*), tRNA-Leu (CAA) harbours an intron that, in *S.cerevisiae*, is known to be essential for the intron-dependent formation of m^5C_{34} catalysed by the TRM4 methyltransferase (55,56). The identity of C_{34} modification in intron-less tRNA-Leu ($C_{34}AA$) of *D.hansenii* is not known but we may anticipate it is 2′-*O*-methyl-C_{34} (Cm_{34}) as in the intron-less tRNA-Leu ($Cm_{34}AA$) of *Candida cylindracea* and *Drosophila melanogaster* [see in (33)]. In *S.cerevisiae*, the methylation of residue 34 in tRNA-Leu ($Cm_{34}AA$), as well as tRNA-Phe ($Gm_{34}AA$) and tRNA-Trp ($Cm_{34}CA$) is catalysed by the TRM7 methyltransferase only in intron-less pre-tRNA, thus subsequently after the removal of intron (57,58). No such correlation can be made between the apparently universal presence of introns in pre-tRNA-Ser (CGA) and (GCU) and any modified nucleotides that are present in the corresponding mature tRNA, or possible alternate base pairing configurations [see (59)].

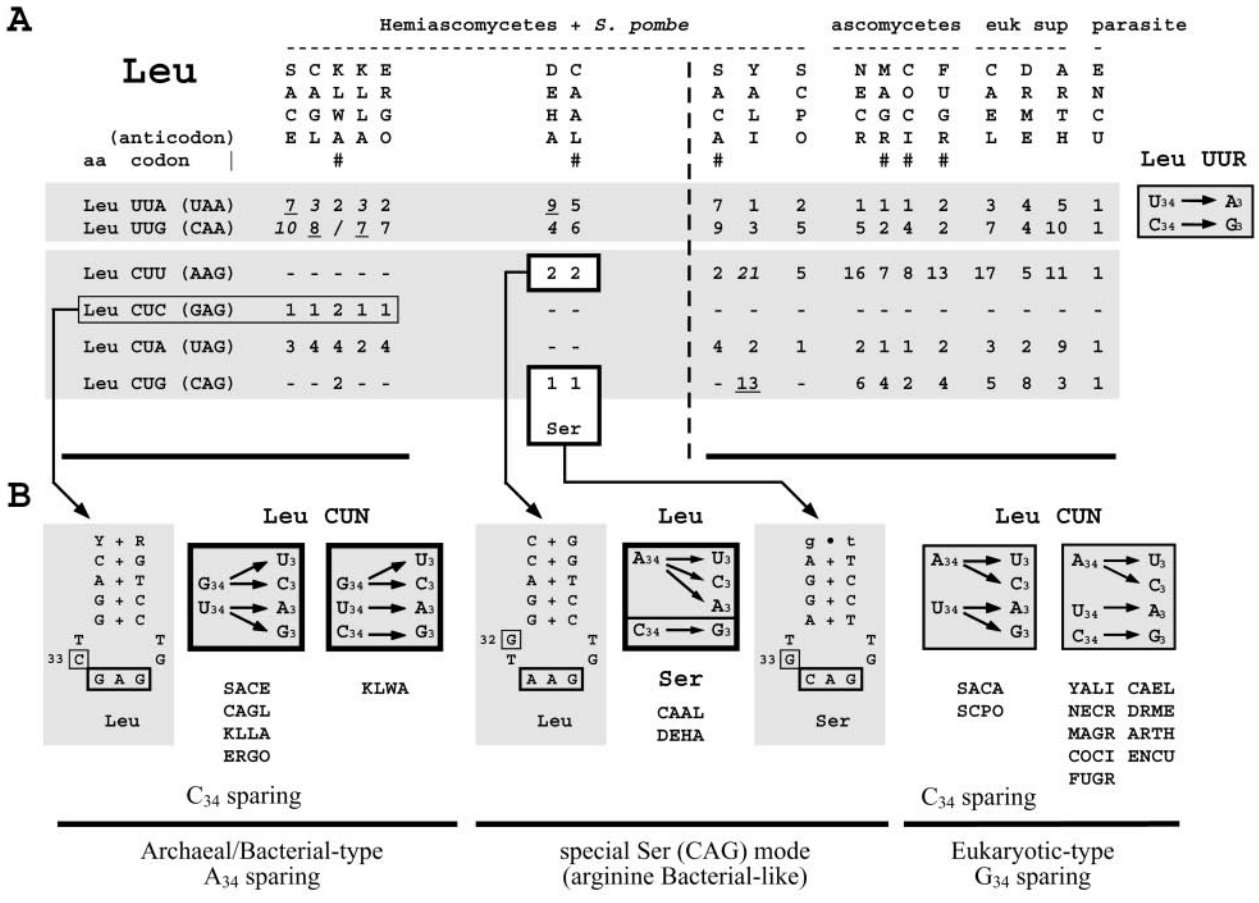
Distribution of tRNA genes, multicopy families and cotranscribed paired genes

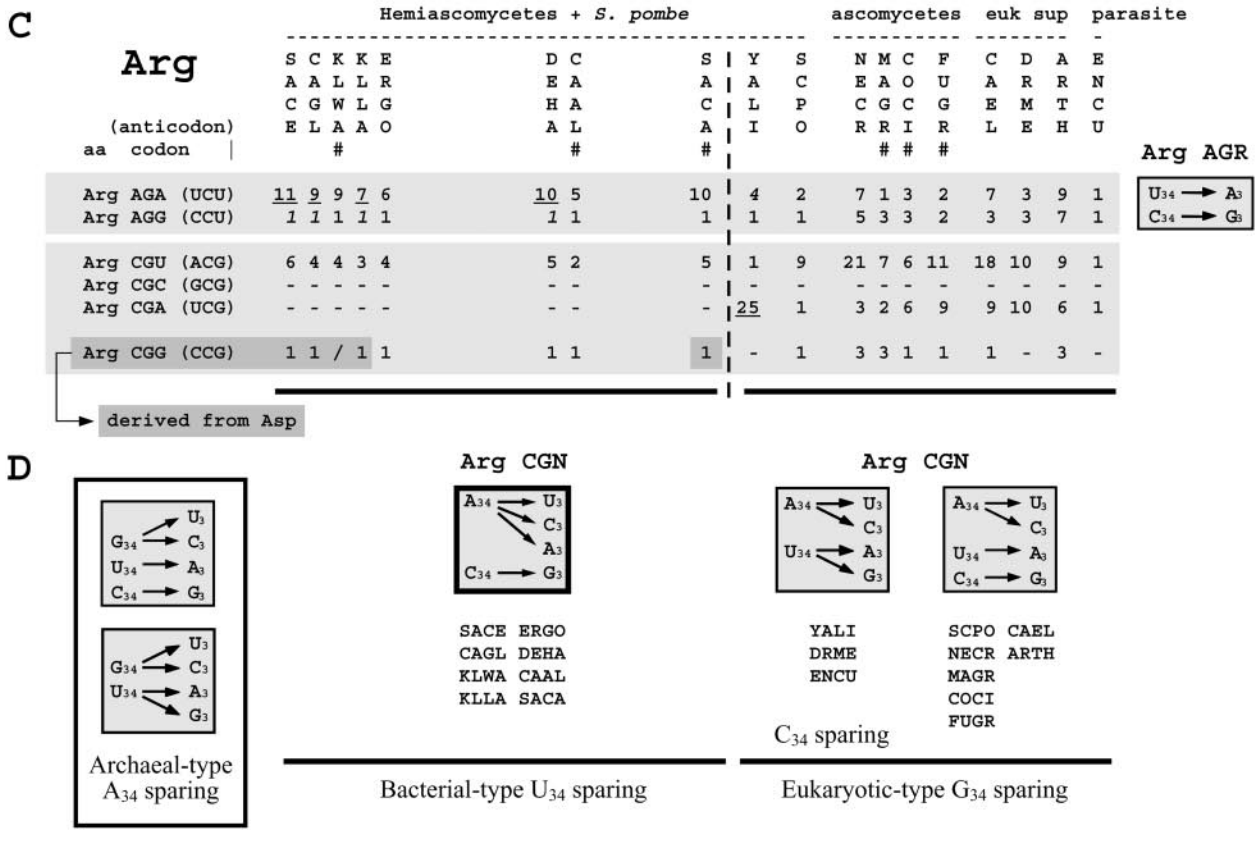
The number of copies encoding the same tRNA varies widely between different isoacceptors within a single yeast species and for the same isoacceptor between different yeast species. For example, several tDNAs exist as single copies while 34 copies of the (DNA-Lys (CTT) are present in *Y.lipolytica*. The same tDNA-Glu (CTC) is encoded by 1 up to 27 copies according to the species (Figure 1). Highly redundant tRNA genes usually correspond to abundant cellular tRNA molecules and poorly redundant tDNA to minor cellular tRNA (60–62) [reviewed in (63), see also (64)]. The total number of tRNA genes varies from 131 (in *C.albicans*) to 510 (in *Y.lipolytica*—see bottom of Figure 1), while 274 tDNA were reported for *S.cerevisiae* (62,65). After correction for genome length, it appears that the tRNA gene density is three times higher in *Y.lipolytica* than in *C.albicans*.

For all hemiascomycetous species, tRNA genes appear scattered throughout the genome. In *S.cerevisiae*, 39 pairs of tRNA genes result from the ancestral duplication of the whole genome (66). Consistent with their variation in total number, the average distances between two successive tRNA genes on the chromosome maps range from 40 kb in *Y.lipolytica* to 110 kb in *C.albicans*. No gene cluster was found in the hemiascomycetous genomes examined except in *D.hansenii* where eight identical tandem co-oriented copies of a tDNA-Lys (CTT) are present on chromosome B. The distances separating these genes (188 to 1855 bp) indicate independent transcription. This is consistent with the frequent formation of tandem genes in this particular species (22).

Clusters of tDNAs have been detected in a variety of eukaryotic genomes including *D.melanogaster* (67) and the archiascomycete *S.pombe*. In the latter case 27 tDNAs are found in the 50 kb region surrounding chromosome B centromere (68), and 20 other tDNA in the 75 kb region around the chromosome C centromere (31). It is remarkable that no such cluster exists in any of the hemiascomycetes studied. Instead, most tDNA are scattered throughout the genome in a random orientation relative to flanking genes.

In studying the localization of tDNA in hemiascomycetes, we were surprised to observe numerous cases of head-to-tail pairs of tDNA. In such pairs, the distance between the two genes ranges from 5 to 26 nt (Supplementary Table 3). This distance is shorter than the minimal sequence required for the independent transcription of the second gene [about 100 nt, (2)]. A few of the pairs had already been noted, as in *S.cerevisiae* and *S.pombe* (69–71) and two more recently discovered in the *S.cerevisiae* (62,65), but their almost universal presence in hemiascomycetes was not suspected. In yeast and *Xenopus* oocyte nuclear extract, the co-transcription of the paired tDNAs into a single precursor followed by processing to mature-size tRNA molecules was experimentally demonstrated (70,72). It is likely that the same mechanism operates for all pairs now identified. In agreement with this hypothesis, all genes are always co-oriented in a given pair and the short intergenic sequences show no obvious


A
Leu
Hemiascomycetes + S. pombe
ascomycetes
euk sup
parasite
(anticodon)
aa codon
Leu UUR
Leu UUA (UAA)
Leu UUG (CAA)
Leu CUU (AAG)
Leu CUC (GAG)
Leu CUA (UAG)
Leu CUG (CAG)
Ser
B
Leu CUN
Leu
Ser
Leu CUN
SACE CAGL KLLA ERGO
KLWA
CAAL DEHA
SACA SCPO
YALI NECR DRME MAGR ARTH COCI ENCU FUGR
C34 sparing
C34 sparing
Archaeal/Bacterial-type A34 sparing
special Ser (CAG) mode (arginine Bacterial-like)
Eukaryotic-type G34 sparing


C
Arg
Hemiascomycetes + S. pombe
ascomycetes
euk sup
parasite
(anticodon)
aa codon
Arg AGR
Arg AGA (UCU)
Arg AGG (CCU)
Arg CGU (ACG)
Arg CGC (GCG)
Arg CGA (UCG)
Arg CGG (CCG)
derived from Asp
D
Arg CGN
Arg CGN
SACE ERGO CAGL DEHA KLWA CAAL KLLA SACA
YALI DRME ENCU
SCPO CAEL NECR ARTH MAGR COCI FUGR
C34 sparing
Archaeal-type A34 sparing
Bacterial-type U34 sparing
Eukaryotic-type G34 sparing

Pol III terminator (tracks of T's), although the strength of terminators is difficult to predict (15).

Interestingly, the tDNA pairs differ from one yeast species to the next, with only limited conservation (e.g. tDNA-Arg/tDNA-Asp pairs found in *S.cerevisiae*, *S.castellii* and *K.lactis*) and the pairs are often found in multiple copies within a genome, e.g. six occurrences of the tDNA-Ile/tDNA-Ala pair in *D.hansenii*. In some cases, the pair is composed of two identical tDNA but in most cases, two distinct tDNA are involved. The expansion of identical pairs in a genome suggests successive duplications of the pairs within each phylogenetic branch through a yet unknown mechanism. Single copies of the tRNA genes identical to those involved in pairs are also present in the same genome. No correlation can be made yet between the decoding capacity of each tRNA of the pairs and the level of their expression. It is possible that some enzymatic modification of nucleotides (like in the case of intron-containing tRNA, see above) or correct folding is however dependent on the expression of such paired pre-tRNAs.

A combination of universal and phylum-specific strategies are used to decode the genetic information in hemiascomycetes

Three major sparing strategies allow an organism to read the genetic code information (insets in Figure 1) with a limited repertoire of anticodons in the tRNAs (21). The first sparing strategy is the universal 'A_{34} or G_{34}-sparing' in which either an A_{34}- or G_{34}-containing tRNA decodes the two pyrimidine ending codons. In these cases, the A_{34} is always posttrancriptionally modified into inosine (I_{34}) (73) [reviewed in (74)] while the G_{34}, is often modified into Gm_{34} or Q_{34} derivatives [reviewed in (75,76)]. The mutually exclusive existence of an A or G at the first position of the anticodon is true in the 16 four-codon boxes of the genetic code, despite the fact that a given tRNA species (given anticodon) is usually encoded by multiple genes. Such cases of tRNA sparing are indicated by arrows and symbols 'ΔA_{34}' or 'ΔG_{34} in Figure 1. As a consequence of this first sparing rule, the maximum number of tRNA species in any organism cannot exceed 46 (64 codons, minus 3 stop codons, minus 16 cases of 'A_{34} or G34-sparing' and plus 1 for the iMet codon leads to 46). This rule holds for all three biological domains: Archaea, Bacteria or Eukarya. This number of 46 tRNA/anticodon species is reached in *S.pombe* but hemiascomycetous yeasts lower the number of tRNAs (Figure 1 bottom). The use of anticodon starting with either A_{34} or G_{34} is the same in all three kingdoms for Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, Cys, Trp, Arg, Ser (AGY codons only) and Gly (Figure 1, Phe in upper panel, others in lower panel). In contrast, this choice differs between Eukarya and Archaea/Bacteria for the other amino acids (Figure 1, upper panel). Eukarya use the anticodons starting with A_{34} (G_{34}-sparing, noted 'ΔG_{34}' in Figure 1) while Archaea and Bacteria use anticodons starting with G_{34} (A_{34}-sparing, noted 'ΔA_{34}') (21).

To achieve additional reduction (down to 44 or 42 tRNA types), hemiascomycetous yeasts use a second sparing strategy, known as 'C_{34}-sparing strategy', which is also used in all three domains of life (indicated as 'ΔC_{34}' in Figure 1). When the tRNA with anticodon starting with C_{34} is absent, the cognate G_3-ending codon is read by the U_{34}-containing isoacceptor tRNA. In this case, U_{34} is always modified [reviewed in (75–77)]. In the genome of *Y.lipolytica*, a set of 44 tRNA is enough to read all 62 codons, the tRNA-Arg ($C_{34}CG$) and tRNA-Gly ($C_{34}CC$) being both absent. Other genomes, like *S.cerevisiae, C.glabrata, K.waltii, C.albicans* and *S.castellii*, lack a few more tRNAs with anticodon starting with C_{34} (Figure 1). Note that *C.glabrata* and *K.lactis* display exactly the same set of 42 tRNA (this work) as *S.cerevisiae* (62,65). As a matter of fact, this moderate variation in the tRNA 'repertoire' (between 42 and 46 tRNA) hides drastic changes in the way each individual yeast decodes Leu and Arg with respect to other eukaryotes (discussed below).

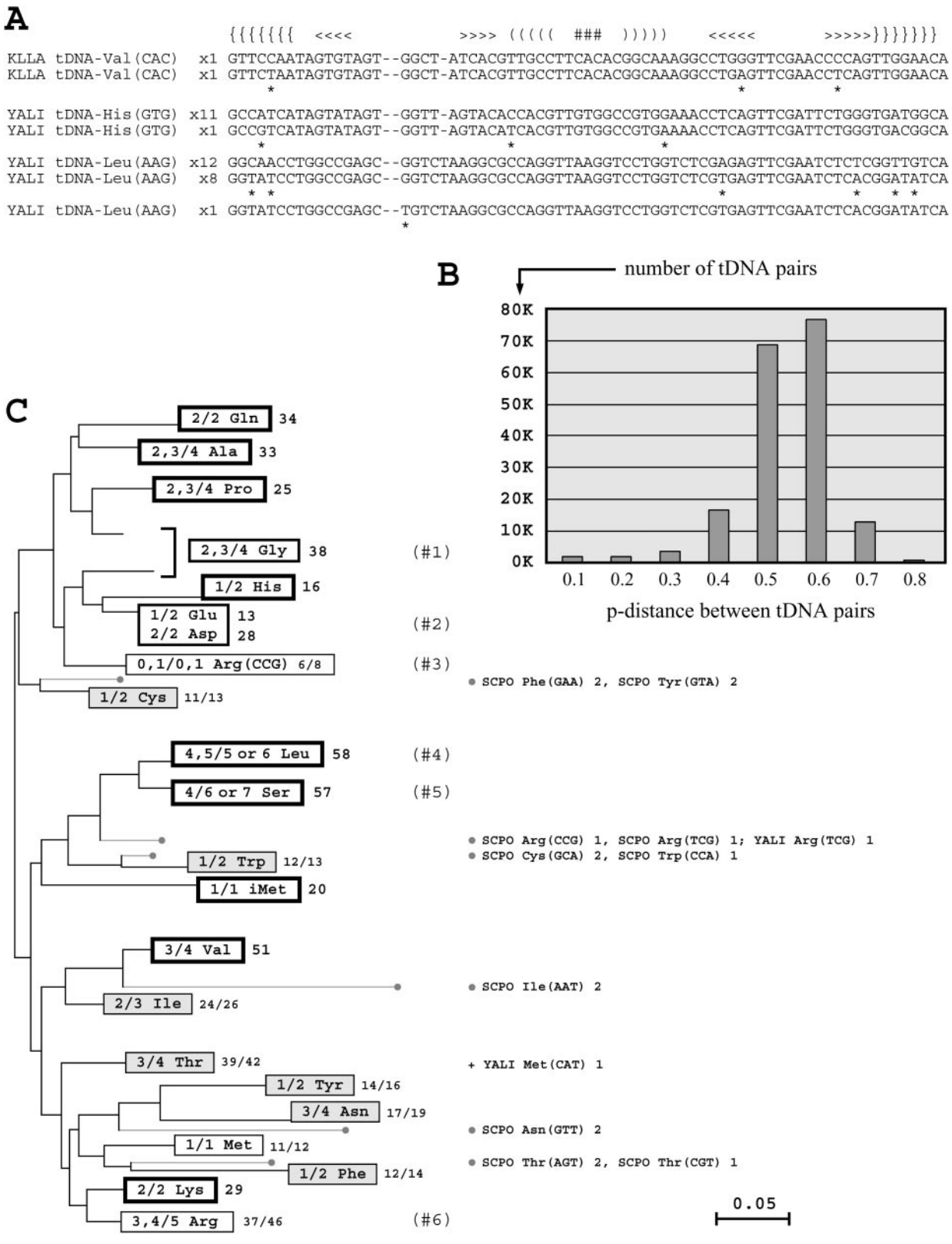
Hemiascomycetes imitate bacteria to read the Leu CTN and Arg CGN codons

The largest variability of tRNA repertoire between yeasts occurs in the decoding of the Leu CTN and Arg CGN codons (boxed in Figure 1). This situation is illustrated more in details in Figure 3A–D together with additional eukaryotes. Genes coding for each of the tRNA reading one of the two purine-ending Leu codons UUA and UUG are universally present. In contrast, two distinct strategies are used to read the four Leu CUN codons. Four of the nine hemiascomycetes and *S.pombe* use the 'Eukaryotic-type G_{34}-sparing' strategy, as expected (tRNA-Leu ($A_{34}AG$). The five hemiascomycetes *S.cerevisiae, C.glabrata, K.waltii*, *K.lactis* and *E.gossypii,* which belong to the same evolutionary branch, use the 'Bacterial/Archaeal A_{34}-sparing' (tRNA-Leu ($G_{34}AG$). This case is unique among eukaryotes.

The other two Leu CUR codons (CUA and CUG in Figure 3A) are read by either a unique tRNA-Leu harbouring

Figure 3. Various strategies used to decode the Leu UUR/CUN and Arg AGR/CGN codons. In addition to the ten genomes explored in this work, data of other eukaryotic genomes are reported for comparison: NECR, *Neurospra crassa* (99); MAGR, *Magnaporthe grisea* [Data Version 10/31/2003 (Release 2.3)]; COCI, *Coprinus cinereus* (Data Version 6/25/2003); FUGR; *Fusarium graminearum* (Data Version 3/11/2003); CAEL, *C.elegans* (100); DRME, *D.melanogaster* (101); ARTH, *Arabidopsis thaliana* (102) and ENCU, *E.cuniculi* (103). '#' signs indicate genomes of low coverage or without full assembled chromosomes. (**A**) Numbers indicate how many genes encode tRNA reading the UUR and CUN leucine codons (with '-' standing for no gene). Underline and italic numbers indicate the most and second most used codons [taken from (24), Supplementary Table S3]. All genomes harbours the two tRNA-Leu (UAA) and (CAA). The two heavy boxes emphasize the particular situation in *D.hansenii* (DEHA) and *C.albicans* (CAAL): i.e. the tRNA-Leu G_{32} (AAG) reads the three codons CUU, CUC and CUA (top box) and the single copy tRNA-Ser G_{33} (CAG) reads the CUG codon (bottom box). The dashed vertical bar separates the species using an eukaryotic-type of sparing (at right) from those using bacteial types (at left). (**B**) Schematic representation of the decoding of the Leu CUN codon in the different genomes indicated. The sequences of the anticodon stems and loops of the three remarkable tRNA-Leu (GAG) of *S.cerevisiae* and four related genomes, tRNA-Leu (AAG) and tRNA-Ser (CAG) of *D.hansenii* and *C.albicans* are shown to illustrate that all three harbour an unusual nucleotide (boxed) close to the anticodon (boxed): C_{33}, G_{32} or G_{33}, respectively. (**C** and **D**) A similar presentation is used for the decoding of the six arginine codons. Darker grey background denote, among the tDNA-Arg (CCG), which ones are presumably derived from the tDNA-Asp (GTC) (see Figure 4B). In the uncomplete genome of *K.waltii* (KLWA), this tDNA was not found, but it is probably present in this organism.

a UAG anticodon (as in *S.cerevisiae, C.glabrata, K.lactis, E.gossypii*, *S.castellii* and *S.pombe*) in which U_{34} is not posttranscriptionally modified (78), or by two tRNAs (in *K.waltii* and *Y.lipolytica*), one with U_{34}AG and the other with C_{34}AG anticodon. Since no RNA sequence of the fully mature tRNA-Leu (UAG) other than that of *S.cerevisiae* is available, we do not know whether U_{34} of this tRNA-Leu (UAG) is modified in other yeasts. The peculiar decoding of



the Leu CUN codons in *D.hansenii* and *C.albicans* (because of a change in the amino acid assignment for one of these codons—shown in Figure 3A and B) is commented below.

Another example of an imitation of bacterial sparing strategy is found in the decoding of the Arg CGN codons (Figures 3C and D). Whereas *Y.lipolytica* (as the archiascomycete *S.pombe*), uses the typical 'Eukaryotic-type G_{34}-sparing' strategy, all other hemiascomycetes use a third type of sparing, known as 'U_{34}-sparing' strategy, which is specific to arginine CGN codons and only known in Bacteria (21). To read Arg CGN codons, *Y.lipolytica* and *S.pombe* use a tRNA-Arg ($A_{34}CG$) reading CGU and CGC codons and a tRNA-Arg ($U_{34}CG$) reading CGA codons [and also CGG codons if the tRNA-Arg ($U_{34}CG$) is absent, as in *Y.lipolytica*]. In this case, U_{34} is possibly modified into a yet unknown derivative of the type mcm^5U like in tRNA-Arg [$mcm^5U_{34}CU$) (33)]. In all other eight hemiascomycetes, the tRNA-Arg ($U_{34}CG$) is and a single tRNA-Arg ($A_{34}CG$) reads the three Arg codons CGU, CGC and also CGA. In this case, A_{34} is probably modified into I_{34} and the decoding of CGA codon involves a wobble $I_{34}A_3$ pairing mode which was initially anticipated (79,80) but only recently demonstrated to occur during mRNA decoding on the ribosome (81). The CGG codons are read by a second tRNA-Arg ($C_{34}CG$) and, in these eight organisms obeying the U_{34}-sparing strategy, this tRNA becomes essential, while it may be absent when the tRNA-Arg ($U_{34}CG$) is present as in *Y.lipolytica*, *D.melanogaster* and *Encephalitozoon cuniculi* (usual C_{34}-sparing strategy) (Figure 3C).

Two unusual tRNA decode Leu CUU, CUC, CUA and Ser CUG codons in *D.hansenii* and *C.albicans*

In *D.hansenii* and *C.albicans*, the 'Leu' codon, CUG, is read as Ser (40) (Figure 3A and B). The tRNA-Leu ($U_{34}AG$), which reads the CUA codon in all other eight yeasts investigated, is missing in these two genomes. Consequently, the Leu codon CUA must be read by the tRNA-Leu ($A_{34}AG$), which cannot be the major tRNA-Leu according to the codon usage and the existence of only two gene copies in each genome. This hypothesis implies that an $I_{34}A_3$ wobble pairing exists in the codon-anticodon pairing during translation on the ribosome. This decoding strategy is analogous to the 'Bacterial-type U_{34}-sparing' mode of reading the three Arg codons CGU, CGC and CGA as discussed above [see also left part of Figure 3C and D and Figure 6 in (21)].

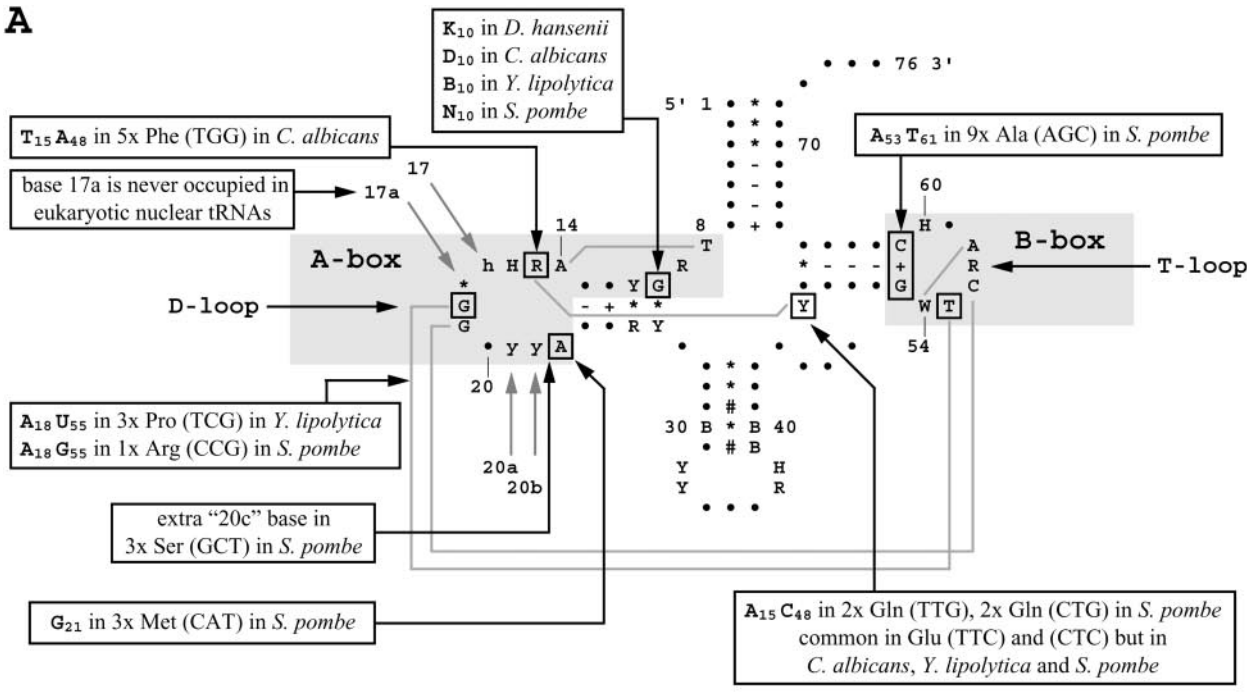
In summary, the absence of a tRNA-Leu ($U_{34}AG$) in both *D.hansenii* and *C.albicans* (U_{34}-sparing strategy) appears consistent with the need to avoid any misreading of the Ser codon CUG of the CUN decoding box. The four codons of this box are read by only two types of tRNA isoacceptors. The first type charges Ser for the CUG codon and possesses an uncommon G at position 33 of the anticodon loop (39–41,82). The second one charges Leu for the three codons CUU, CUC and CUA and possesses also an unusual G, but located at position 32, instead of the universal pyrimidine found in 4000 tDNAs analysed (21). The G at position 32 cannot result from sequencing errors because it is found in the two gene copies in each genome (*D.hansenii* and *C.albicans*). For this tRNA, we do not know what is its decoding capability compared to a more 'normal' tRNA and whether A_{34} is posttrancriptionnally modified into I_{34} (83,84).

The presence of G_{32} or G_{33} instead of the universal pyrimidines [C or U, see in (47)] probably alters the anticodon stem–loop structure and allows accurate readings of CUU, CUC and also CUA as leucine in the case of tRNA-Leu ($A_{34}AG$) [possibly ($I_{34}AG$)] and CUG as serine in the case of tRNA-Ser ($C_{34}AG$). The coexistence of these two types of unusual tRNA among the tRNA population within the same organism (*D.hansenii* and *C.albicans*) is therefore not a coincidence but rather an important novel feature of the decoding strategy in these microorganisms [see also (85)].

Global distance analysis of tDNA reveals sequence conservation and functional recruitment

Evolutionary relationship between the tRNA gene species of the different yeasts were investigated using a pairwise p-distance matrix analysis carried over the 603 variant tDNA sequences (see bottom of Figure 1) identified in the ten yeasts. Some examples of tDNA sequences prepared for p-distance computation are shown in Figure 4A. All pairwise distances were computed after removal of the intronic sequences (if any), and of base 47 and V-arm sequences in

Figure 4. Distance tree analysis of 603 tDNA isoacceptor sequences from nine hemiascomycetous genomes and *S.pombe*. (**A**) Examples of tDNA sequences prepared for the p-distance analysis. The intron and sequences between nt 46 and 48 were removed to obtain perfectly aligned sequences, all 75 nt long. The stems are symbolized with '{}', acceptor stem; '<>', D- and T-stems; '()', anticodon stem; the anticodon is indicated with '###'. Stars indicate sequence variations. (**B**) The 603 different isoacceptor tDNA sequences analysed (see Materials and Methods) generate 181 503 pairwise p-distances. This histogram displays the number of p-distance values between two tDNA versus the value of the p-distance. (**C**) A p-distance unrooted tree was computed from the p-distance matrix. For the sake of clarity, this tree is presented vertically and the sub-trees in which neighbour tDNA encoding the same amino acid cluster together are symbolized by boxes. The actual branches inside the sub-trees extends rightwards far beyond the right edge of the boxes. The number of anticodons and codons specific to the amino acid are given inside each box (e.g. '2,3/4 Ala' means 2 or 3 anticodons and 4 codons for alanine). The three types of boxes are as follows: (i) Heavy lined boxes: all isoacceptors for a given amino acid (whatever are the anticodons) from the ten genomes cluster together; the total number of corresponding sequence types are given outside the boxes at right. (ii) Light boxes: not all the tDNA isoacceptors from the ten genomes cluster together; numbers at right indicate the fraction of tDNA that cluster over the total number of sequences considered for the amino acid. (iii) Light boxes filled with grey: all isoacceptors from the nine hemiascomycetes cluster together, but not those from *S.pombe* (SCPO) which are indicated at the right side (long grey horizontal branches ending with a dot). The signs '+' at right indicate extra tDNA sequences that cluster inside uncomplete sub-trees (light boxes). For clarity, some vertical spacing was introduced in the drawing of the tree but the length of the horizontal branches was not modified. Notes: (#1) tDNAs-Gly split into two clusters; the upper one contains most of the tDNAs-Gly (TTC). (#2) tDNAs-Asp and tDNAs-Glu do not form two separate clusters but a single one mixing these two neighbour isoacceptors. (#3) The gene of the tDNA-Arg (CCG) of *S.cerevisiae* [which is related to the tRNA-Asp (GTC) of the same organism (see text)] as well as those from five other genomes (over eight harbouring such as tDNA) define a special cluster close to the Asp/Glu one (see also note #6). (#4) in *D.hansenii* (DEHA) and *C.albicans* (CAAL), the CUG codon is used for serine instead of leucine and only five codons remain for leucine. (#5) This cluster includes the two special single copy tDNA-Ser (CAG) of *D.hansenii* (DEHA) and *C.albicans* (CAAL) (CUG is a 7th serine codon in these two genomes). (#6) This cluster gets together tDNA-Arg other than tDNA-Arg (CCG) (however that of *D.hansenii* clusters here and not in the extra Arg (CCG) cluster).



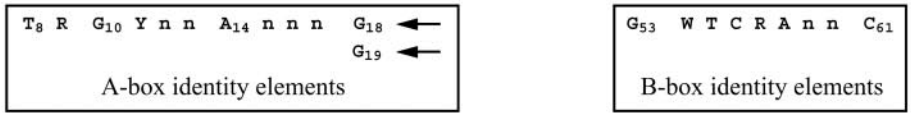
C

	A-box: D-stem	A-box: D-loop	B-box: T-loop
S. cerevisiae	T_8 R	G_{10} Y n n	A_{14} R H h * G_{18} G_{19} n y y A_{21}	G_{53} W T C R A n H C_{61}
S. castellii	T_8 R	G_{10} Y n n	A_{14} R Y y * G_{18} G_{19} n y y A_{21}	G_{53} W T C R A n H C_{61}
C. glabrata	T_8 R	G_{10} Y n n	A_{14} R H y * G_{18} G_{19} n y t A_{21}	G_{53} W T C R A n H C_{61}
K. waltii	T_8 R	G_{10} Y n n	A_{14} R n y * G_{18} G_{19} n y t A_{21}	G_{53} W T C R A n H C_{61}
K. lactis	T_8 R	G_{10} Y n n	A_{14} R n y * G_{18} G_{19} n y y A_{21}	G_{53} W T C R A n H C_{61}
E. gossypii	T_8 R	G_{10} Y n n	A_{14} R H y * G_{18} G_{19} n y y A_{21}	G_{53} W T C R A n H C_{61}
D. hansenii	T_8 R	K_{10} Y n n	A_{14} R H y * G_{18} G_{19} n y t A_{21}	G_{53} W T C R A n H C_{61}
C. albicans	T_8 R	D_{10} Y n n	A_{14} D n y * G_{18} G_{19} n b t A_{21}	G_{53} W T C R A n H C_{61}
Y. lipolytica	T_8 R	B_{10} Y n n	A_{14} R H h * R_{18} G_{19} n b y A_{21}	G_{53} W T C R A n H C_{61}
S. pombe	T_8 R	N_{10} B n n	A_{14} R n y * R_{18} G_{19} n b t R_{21}	R_{53} W K C R A n H Y_{61}

D

S. cerevisiae	T_8 R G_{10} Y n n A_{14} R H H G_{18} G_{19} n y y A_{21}
	T_8 R G_{10} Y n n A_{14} R H G_{18} G_{19} n y y A_{21}
	T_8 R G_{10} Y n n A_{14} R H H G_{18} G_{19} n y A_{21}
	T_8 R G_{10} Y n n A_{14} R H G_{18} G_{19} n y A_{21}
	T_8 R G_{10} Y n n A_{14} R H H G_{18} G_{19} n A_{21}
	T_8 R G_{10} Y n n A_{14} R H G_{18} G_{19} n A_{21}

E



tDNAs-Leu and tDNAs-Ser. The repartition of the pairwise distances obained (Figure 4B) shows a majority of p-distance values in the range 0.5–0.6 (50–60% difference). The p-distance tree derived from this matrix (according to Materials and Methods) is shown in Figure 4C. Remarkably, for tDNA specific for Gln, Ala, Pro, His, Leu, Ser, iMet, Val and Lys, orthologous tRNA genes (coding for the same amino acid) belonging to the nine hemiascomycetes or even to all ten yeast species cluster together. In such instances, the evolutionary divergence of sequences between orthologous tDNA of different yeast species is less than the divergence between paralogous tDNA species (charging different amino acids) within a single yeast species.

The sequences of three other isoacceptor families (specific for Gly, Asp and Glu) cluster less perfectly: indeed, tDNA-Gly split into two clusters while the tDNA specific for Asp and Glu are fused in the same cluster. The tDNA for amino acids Cys, Trp, Ile, Thr, Tyr, Asn and Phe cluster together for all hemiascomycetous yeasts but the clusters do not contain orthologs from *S.pombe*. The fact that initiator tDNA-Met and elongator tDNA-Met do not cluster together was expected due to clear singularities in the sequences for initiator tRNA [discussed in detail in (21)].

A novel case of 'tDNA mimicry' was identified: the only tDNA missing in the elongator tDNA-Met cluster (light box) is that of *Y.lipolytica* that clusters inside the Thr cluster (indicated by + YALI Met (CAT) on the right side of the Figure 4C). This tDNA-Met (present in nine identical copies) is very close in sequence to one of the two copies of tDNA-Thr (TGT) of *Y.lipolytica* (57 positions identical) while these two copies diverge at 20 positions. These data are indicative of a possible tDNA capture (tDNA-Met derived of tDNA-Thr in *Y.lipolytica*) similar to the case of tDNA-Arg (CCG) commented below.

It is worth mentioning that the different tDNA-Leu and tDNA-Ser form a unique cluster, despite the fact that these amino acids correspond to two distinct decoding boxes (4 + 2 codons for each). This observation is consistent with the fact that these tDNAs are phylogenetically related (86). Interestingly, the tDNA-Ser harbouring a CAG anticodon, hence reading CUG as Ser instead of Leu in *C.albicans* and *D.hansenii* (detailed in Figure 3A and B and discussed above), clusters with the Ser-tDNAs and not with the Leu-tDNAs, thus attesting to its clear affiliation to the tDNA-Ser family. In contrast, the sequences of the five tDNA-Arg isoacceptors, which also belong to two different decoding boxes (CGN and AGR), are split into two separated clusters. The first cluster (noted '3,4/5 Arg', at bottom of Figure 4B) contains all tDNA-Arg except six of the eight tDNA-Arg (CCG) that form a separate cluster (noted '0,1 0,1 Arg (CCG)') close to the tDNA-Asp/tDNA-Glu cluster. Fender and coworkers proposed that the arginine specific tRNA (CCG) gene from *S.cerevisiae* (as well as those of *Saccharomyces uvarum*, *Zygosaccharomyces rouxii*, *C.glabrata* and *K.lactis*) is a remnant of a former aspartate acceptor (87). The conversion of only two bases (G_{38} and U_{73} into C_{38} and G_{73}, respectively) in an *in vitro* transcript of tDNA-Arg (CCG) is sufficient to allow mutant tRNA-Arg (CCG) to become an aspartate acceptor (87). We now show that this tDNA-Arg (CCG), which does not exist in *Y.lipolytica*, is also presumably derived from the tDNA-Asp (GTC) in *S.castellii* and *E.gossypii* but not in *D.hansenii* and *C.albicans* (tDNA-Arg (CCG) allowing us to date the recruiting event on the hemiascomycete tree (see Discussion).

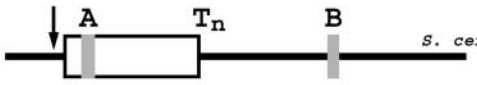
Only few conserved nucleotides internal to tDNA are major identity elements for their recognition by TFIIIC

The large collection of tDNA sequences extracted from ten yeasts allows for a better definition of the A- and B-consensus sequences which are recognized by the transcription factor TFIIIC. Only one G remains in the final genomic consensus of the A-box if the variable occupancy of the optional bases 17 and 17a of the D-loop is considered. The consensus sequence (in the form of a cloverleaf) of the 274 tDNAs from *S.cerevisiae* is shown in Figure 5A, while Figure 5B illustrates the variable distance between the A- and B-boxes and Figure 5C lists the conserved and semi-conserved nucleotides found in the tDNAs of each of the yeasts examined in this work. The position numbered 17a (indicated by arrow and asterisk in Figure 5A and B) is never occupied in any of eukaryotic tDNAs sequenced so far. This is a major difference with the situation in tDNAs of archaeal and bacterial genomes where position 17a (and 17) is occupied in 64 and 7% of the tDNAs, respectively [see Supplementary Table 1 in (21)]. Also few nucleotides are strictly conserved (T_8, G_{10}, A_{14}, G_{19} and A_{21} in the A-box; G_{53}, T_{55}, C_{56}, A_{58} and C_{61} in the B-box). Most of the sequence variability occurs in the evolutionarily distant yeast *S.pombe* (sequence exceptions are indicated in the boxes surrounding the cloverleaf in Figure 5A).

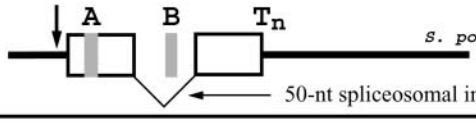
In the A-box, exceptions to the conserved G_{10} are mostly found in the tDNA-Leu and tDNA-Ser of *D.hansenii*,

Figure 5. tRNA genes promoter sequences and polymorphism of the A-box/D-loop and B-box/T-loop sequences. (**A**) The consensus sequence of the 274 tRNA genes from *S.cerevisiae* is used to emphasize the conserved elements recognized by the RNA polymerase III machinery at the DNA level. These elements concentrate in the two areas referred to as the A- and B-boxes (grey boxes) which include, at the RNA level, the forward strand of the D-stem plus the D-loop (A-box) and the T-loop plus the terminal base pair of the T-stem (B-box), respectively. Typographical symbols and nucleotide abbreviations are given in the legend to Figure 2. The arrows indicate the four variable positions of the D-loop (occupied, or not, indicated with lower case letters). Exceptions in the sequences of the A- and B- boxes of some tDNA are indicated in surounding boxes. Grey lines connecting nucleotides between the D- and T-loops represent tertiary base pairs. (**B**) Schematic representation of a tRNA gene without (left) or with (right) an intron. The solid black bar represents upstream and downstream DNA, the open rectangle(s) the mature product and the two grey rectangles the A- and B-boxes. Transcription starts about 20–25 bases upstream the A-box (vertical arrows) and terminates inside the poly-T track (T_n). (**C**) Consensus sequences observed in the A- and B-boxes for the ten genomes; the position 17a (second vertical arrow) is never occupied in the nuclear eukaryotic tDNA. Sequence deviations [detailed in (A)] at positions 10, 19, 53, 55 and 61 are boxed and the tDNA in which they occur are highlighted with grey background in Figure 1. (**D**) The six possible cases of A-box sequences (according to various occupancies at positions 17, 20a and 20b) written without gap. Various occupancies at positions 20a and 20b generate three possible patterns (no base, only 20a occupied, both 20a and 20b occupied). For each of these patterns, position 17a can be occupied or not, thus generating a total of six possible patterns. (**E**) Final consensus sequences (not taking into account the exceptions shown in (A and C)) of the A- and B-boxes used to search these elements in the other Pol III genes from the ten genomes (see Figure 6). Note that the fourth nucleotide downstream A_{14} is always a G (either G_{18} or G_{19}, twin horizontal arrows).

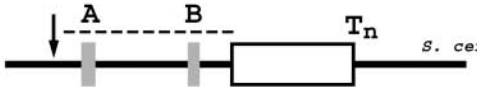
SNR6 (U6)



genes		mature product				A-box			ΔA-B		B-box			Δter-B
species	direction	boundaries		length	pos.	TRGYnnAnnnG				pos.	GWTCRAnnC			
chromosome/contig														
S. cerevisiae	L >	366244	366347	104	+21	T TGGTcaAtttG A			195	+227	C GTTCGAacC T			122
S. castellii	713 >	86932	87034	103	+21	T TGGTcaAtttG A			185	+217	C GTTCGAtcC T			113
C. glabrata	M >	1158035	1158137	103	+21	T TGGTcaAtttG A			188	+220	A GTTCGAttC T			116
K. waltii	175 <	289	392	104	+21	T TGGTcaAtttG A			208	+240	C GTTCGAaaC T			135
K. lactis	F >	1858716	1858817	102	+21	T TGGTcaAtttG A			248	+280	C GTTCGAatC A			177
E. gossypii	G <	811786	811889	104	+21	T TGGTcaAtttG A			182	+214	C GTTCGAggC G			109
D. hansenii	D <	1319741	1319842	102	+21	T TtGTcaAtttt A			204	+236	G GTTCAAttC C			133
C. albicans	2500 >	66590	66690	101	+21	C TtGTcaActtt A			205	+237	A GATtcAatC T			113
Y. lipolytica	F <	1089094	1089197	104	+21	A TGGTcaAtttG A			212	+244	A GTaCAAtaC T			139
S. pombe (#1)	A <	2562271	2562423	153	+21	T TGGTcaAattG A			41	+73	G GTTCGAgtC A			na

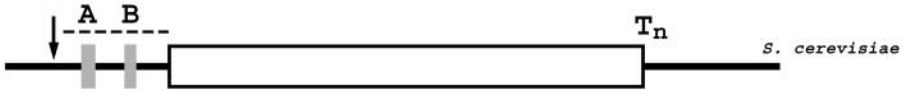


SNR52

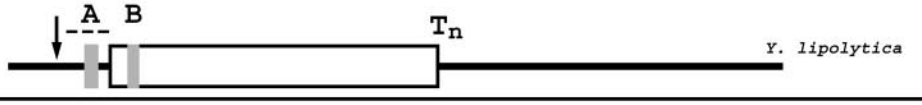


S. cerevisiae	E >	431115	431216	102	-139	T TGGgctAgcgG T			93	-35	C GTTCGAaaC T			
S. castellii	627 <	23031	23130	100	-65	T TGGTcaAgtgG T			25	-29	G GTTCGAaaC T			
C. glabrata	K >	180575	180680	106	-208	G TGGCgcAacgG T			101	-96	T GATCGAgtC C			
K. waltii	139 >	16141	16244	104	-138	G TcGCccAgcgG G			53	-74	C GTTCGAatC C			
K. lactis	F <	1157744	1157873	130	-94	G TAGCttAgtgG G			74	-9	A GTTCGAatC T			
E. gossypii	F >	737999	738086	88	-73	G TGGCccAgcgG G			31	-31	G GTTCGAggC C			
D. hansenii	F >	265190	265298	109	-88	T TGGCacAgtcG T			48	-29	G GTTCGActC G			
C. albicans	10254 >	52700	52823	124	-109	T TGGCatAgctG A			67	-31	A GTTCGAgaC T			
Y. lipolytica	D >	677176	677284	109	-118	A TGGCacAgcgG T			33	-74	A GTTCGAatC T			
S. pombe (#2)	A >	339556	339638	83	na	na			na	na	na			

RPR1

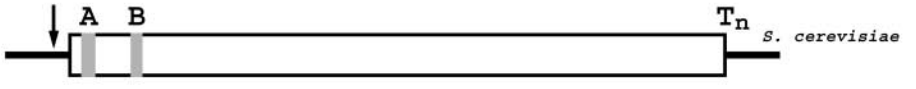


S. cerevisiae	E <	117667	118035	369	-64	G TGGCgcAcatG G			24	-29	G GAaCGAaaC T			
S. castellii	583 >	8832	9184	353	-61	T TGGTtcAccaG G			24	-26	C GATCAAaaC T			
C. glabrata	L >	877630	878778	1149	-121	G TGGTtcAgtgG T			101	-9	A GTTCGAaaC C			
K. waltii	194 <	3421	3751	331	-45	C TGGaccAgttG G			25	-9	C GTTCGAatC C			
K. lactis	E <	2036599	2037139	541	-44	G TGcCgtAgctG G			25	-8	G GTTCGAatC T			
E. gossypii	E <	530133	530512	380	-43	C TGGCcgggcgG C			29	-3	G GTTCGAgaC A			
D. hansenii	C <	876076	876392	317	-71	A TGGTtaAccat A			56	-4	G GTTCcAacC T			
C. albicans	10241 <	31955	32291	337	-14	T TGGCgtAatcG C			102	+100	C GcTCcAtcC T			
Y. lipolytica	E <	2070712	2070999	288	-28	C TGGCctAacgG T			68	+52	A GTTCGAatC C			



S. pombe (#2)	C <	989942	990834	893	na	na			na	na	na			

SCR1



S. cerevisiae	E >	441983	442504	522	+18	C TGGTgggatgG G			24	+53	G GAaCAAatC C			
S. castellii (#3)	627 >	30509	31057	549	+17	T TAGTggAacca A			24	+52	G GAaCGAaaC T			
S. castellii	710 >	40849	41397											
C. glabrata	K >	171939	172438	500	+18	T TAGTggAattG T			24	+53	G GTTCGAacC C			
K. waltii	230 >	21426	21840	415	+17	A cAGTggAactG T			24	+52	G GTTCAAacC C			
K. lactis	F >	1162042	1162526	485	- 7	C TGtCgaAggaG T			25	+30	G GTTCGAagC C			
E. gossypii	F <	769197	769578	382	+17	T TGGCcggaagG G			36	+64	C GTTCGAacC T			
D. hansenii	F >	1783871	1784130	260	+17	A TGGCagAagcG C			45	+73	G GTTCGAtct C			
C. albicans	10053 >	92936	93200	265	+17	T TAGCggAagcG T			50	+78	G GcTCGAtcC C			
Y. lipolytica (#4)	A <	176311	176581	271	+18	T TGtCggAgtgG T			21	+50	C GTTCGAgtC T			
Y. lipolytica	D >	819163	819438	276	+18	T TGtCggAgtgG T			21	+50	C GTTCGAgtt C			
S. pombe	A <	4262510	4262776	267	+15	T TGGTcgAagtG T			37	+63	G GTTCGAgtC T			

C.albicans and *Y.lipolytica* and *S.pombe* (these tDNAs are shown as grey background in Figure 1). At position 18, an A (instead of G) is found twice. The tDNA-Pro (TGG) (3 copies) of *Y.lipolytica* harbours an unusual A_{18} (instead of G_{18}), which probably allows a $A_{18}U_{55}$ tertiary base pair instead of the bifurcated GU pair in the tRNA trancript (88,89). In *S.pombe*, the single copy tRNA-Arg (CCG) harbours an unusual $A_{18}G_{55}$ tertiary base pair, and no other tDNA bearing G at position 55 has been identified.

Taking into account the various combinations of bases present or absent at the four positions 17, 17a, 20a and 20b of the cloverleaf, six different DNA patterns are possible (Figure 5D). Similar results and conclusions are obtained with the tDNAs of the nine other yeasts (data not shown). Remarkably, if the six patterns are combined, a G (either G_{18} or G_{19}, as shown in Figure 5E) is always found, in the genomic sequences, four bases downstream of the universally conserved A_{14}. We therefore hypothesized that the minimal identity elements of the A-box (A-box signature) recognized by the *S.cerevisiae* transcription factor TFIIIC is the 11 nt sequence TRGYnnAnnnG, ending with only one G (n being any base, R a purine and Y a pyrimidine).

Similarly, we assume that the minimal identity elements (sequence signature) of the B-box are the 9 nt sequence GWTCRAnnC (W meaning A or T, Figure 5E). A consensus sequence comprising 11 bases (i.e. including 52–62 bp) was previously reported for the B-box (19), but it clearly appears that 52–62 bp is not conserved for tRNA, even within *S.cerevisiae* [see also data in Supplementary Table 2 in (21)]. A remarkable exception concerns the nine copies of tDNA-Ala (AGC) from *S.pombe* that harbour an A at position 53 and a T at position 61. This GC to AT base pair change at both edges of the B-box is probably counterbalanced by the greater role of upstream TATA sequence that helps the binding of the second factor TFIIIB in *S.pombe* (90). Given these minimal consensus for the A- and B-boxes, extracted from the tRNA genes (Figure 5E), we checked whether they could be retrieved in other Pol III genes from the ten genomes.

Other Pol III genes harbour the same minimal promoter sequences as the tDNAs

We then investigated whether the minimal A and B sequences obtained from the tDNA analysis (shown in Figure 5E) are also retrieved in the four other Pol III genes *SNR6*, *SNR52*, *RPR1* and *SCR1* common to the nine hemiascomycetes and *S.pombe*. The multicopy 5S gene, which is also transcribed by Pol III was not investigated here because it is recognized by its specific transcription factor, TFIIIA. In the case of *S.cerevisiae*, the A- and B-promoter sequences of all four genes have been experimentally investigated [*SNR6* (3); *SNR52* (6); *RPR1* (38); *SCR1* (5)]. In these genes, the promoters are always internal to the primary transcript but, in contrast with tRNA genes, they are, in some cases, external to the mature product (see schemes in Figure 6).

Among the four genes considered, *SNR6* (U6) appears the most conserved in sequence. In *S.cerevisiae* the B-box is exceptionally located about 120 bases beyond the gene and an upstream TATA promoter element is also present (3). The extragenic B block of *SNR6* was located in the orthologous genes at comparable distance (109–177 nt, Figure 6). A specialized chromatin structure appears to dictate this peculiar organization of the two promoter sequences (91). Remarkably, the U6 gene of *S.pombe* is the only Pol III gene known to be interrupted by a spliceosomal intron (92), and the B-box (which perfectly fits the consensus) is located inside this intron rather than downstream of the gene. This peculiar type of organization of the U6 gene is specific to the *Schizosaccharomyces* genus (93).

The next two genes, *SNR52* (6) and *RPR1* (94) share a common organization (at least in *S.cerevisiae* and related genomes): the A- and B-boxes are internal to the transcript but external to the matured product since a leader sequence (dashed lines in Figure 6) is cleaved posttranscriptionally. No structural constraint applies, at the RNA level in this leader sequence and larger variations in the locations of A- and B-boxes occur in *SNR52* (−208 to −65 for the A-box, −96 to −9 for the B-box). In the *S.cerevisiae* gene, a TTTTTT sequence is present 3′ of the A-box; this sequence was shown to be a weak Pol III transcriptional terminator (30% efficient) in the *SNR52* context (15). In the gene of *S.pombe*, we did not identify any A- or B-box nor a Pol III terminator poly-T, arguing for a Pol II transcription for this gene as is the case of most snRNAs in yeasts.

The *RPR1* genes previously identified (4) were structurally aligned. Two types of promoter organization (internal or external to the mature product) can be distinguished (Figures 6 and 7A) in accordance with the phylogenetic distances. From *S.cerevisiae* to *K.lactis*, both the A- and B-boxes are located upstream of the matured product (as in *SNR52* genes). In *E.gossypii* and *D.hansenii*, the B-box terminates in the mature product (inside the 5′ strand of the P1 helix) while, in *C.albicans* and *Y.lipolytica*, the B-box is fully internal to the mature product (boxed in Figure 6). In

Figure 6. Conservation of the A- and B-boxes promoter sequences in four RNA Pol III genes. The schematic representation of the genes uses the same convention as in Figure 5B. Coordinates of the mature products are given with a letter indicating the chromosome (or the contig number) and the direction ('>',Watson strand; '<', Crick strand). Positions of the first base of A- and B-boxes (conserved T for the A-box, conserved G for the B-box) are given with respect to the first nucleotide of the mature product (numbered +1). A positive coordinate indicates that the promoter sequence is located inside the mature product (as in *SNR6* and *SCR1*); while a negative coordinate indicates that the promoter sequence is located in a leader sequence cleaved posttrancriptionally (shown as dashed lines in *SNR52* and *RPR1*). 'ΔA-B' indicates the distance (nt) separating the A- and B- boxes. Additionally, for the *SNR6* genes, the distance between the 3′ end of the gene and the external 3′ B-box is given ('Δter/B'). Nucleotides corresponding to 'n' (any nucleotide) in the genomic sequences are written in lower case. Exceptions in the positions conserved or semi-conserved in the consensus are also written in lower case. The nucleotides preceding and following the A- and B-boxes in the genomic sequences are also reported (separated with a blank) to better enhance the actual boundaries of the sequences putatively recognized by TFIIIC. The 'na' indication stands for 'not applicable'. Boxes hilite the peculiar organization of the *S.pombe SNR6* gene (B-box inside a spliceosomal intron) and that of *RPR1* gene of *E.gossypii* and *D.hansenii* (B-box overlapping the 5′ boundary of the mature product) and *C.albicans* and *Y.lipolytica* (B-box internal). Notes: (#1) In the *SNR6* gene of *S.pombe* (black box), the B-box is not located beyond the gene but inside a 50 nt spliceosomal intron located at positions 51–100; boundaries and length reported here include the intron. (#2) The *SNR52* and *RPR1* genes of *S.pombe* are probably Pol II genes as no A/B-boxes nor poly-T terminators are present. (#3) Two 100% identical *SCR1* genes are present in *S.castellii*. (#4) Two *SCR1* genes (94% identity) are present in *Y.lipolytica*.

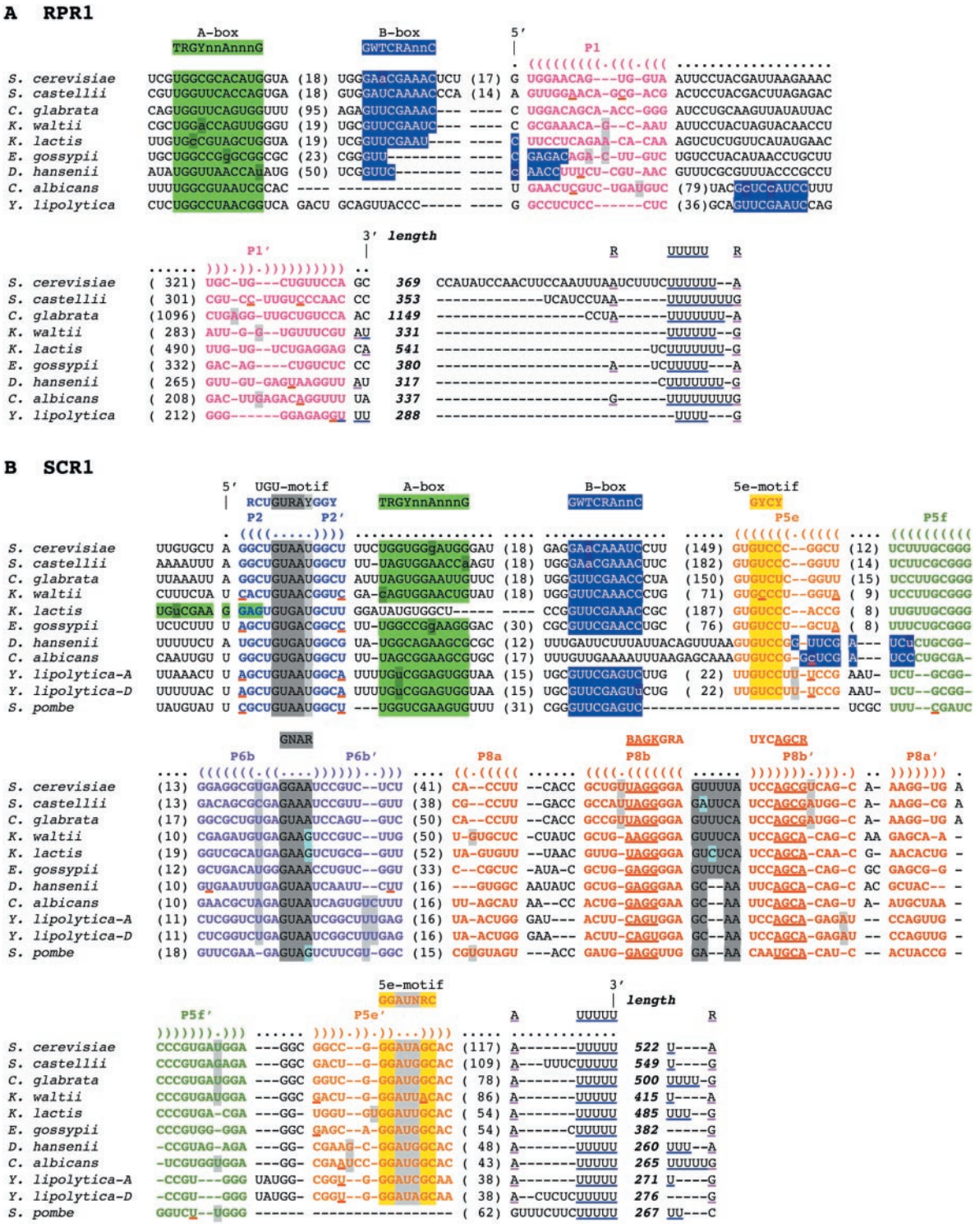


Figure 7. Structural alignment of the RPR1 and SCR1 RNAs from nine hemiascomycetes and *S. pombe*. Header lines indicates the A- and B-promoter elements (green and blue background, respectively; white letters for B) and some consensus sequence elements. On the next two lines are displayed the names of helices with the bracket notation: a dot indicates a single stranded nucleotide and brackets, open for the 5′ end and closed for the 3′ end, indicate helices. Regions for which the structure is not specified are represented as single strands (with dots). Sequences of each species are aligned in a phylogenetic order favouring closest homology between neighbour genomes. First column indicates species names. In each sequence, a dash sign (-) indicates a 1 nt gap, whereas the number of nucleotides (in brackets) indicates a longer gap. Underlined pairs of nucleotides in red color indicate that they do not form nor a Watson–Crick or a GU wobble base pair. Bulges are highlighted in light grey and terminal loops in dark grey. Some nucleotides are highlighted in light blue to emphasize the variations occurring only in one sequence. The lowercase letters, highlighted in a darker color, indicate the nucleotides which are different from the consensus sequence of each promoter. The boundaries of the mature products are indicated with '5′' and '3′'. With the exception of *S.pombe*, the Pol III termination signals (poly-T, underlined with blue) are followed by A or G (underlined in pink color) indicating an efficient terminator. (**A**): RPR1 RNAs, the product of the *S.pombe* gene (transcribed by Pol II) is not shown. (**B**): SCR1 RNAs, the genes of the two RNAs from *Y.lipolytica* are located on chromosomes A and D. The nucleotide abbreviations are given in the legend to Figure 2.

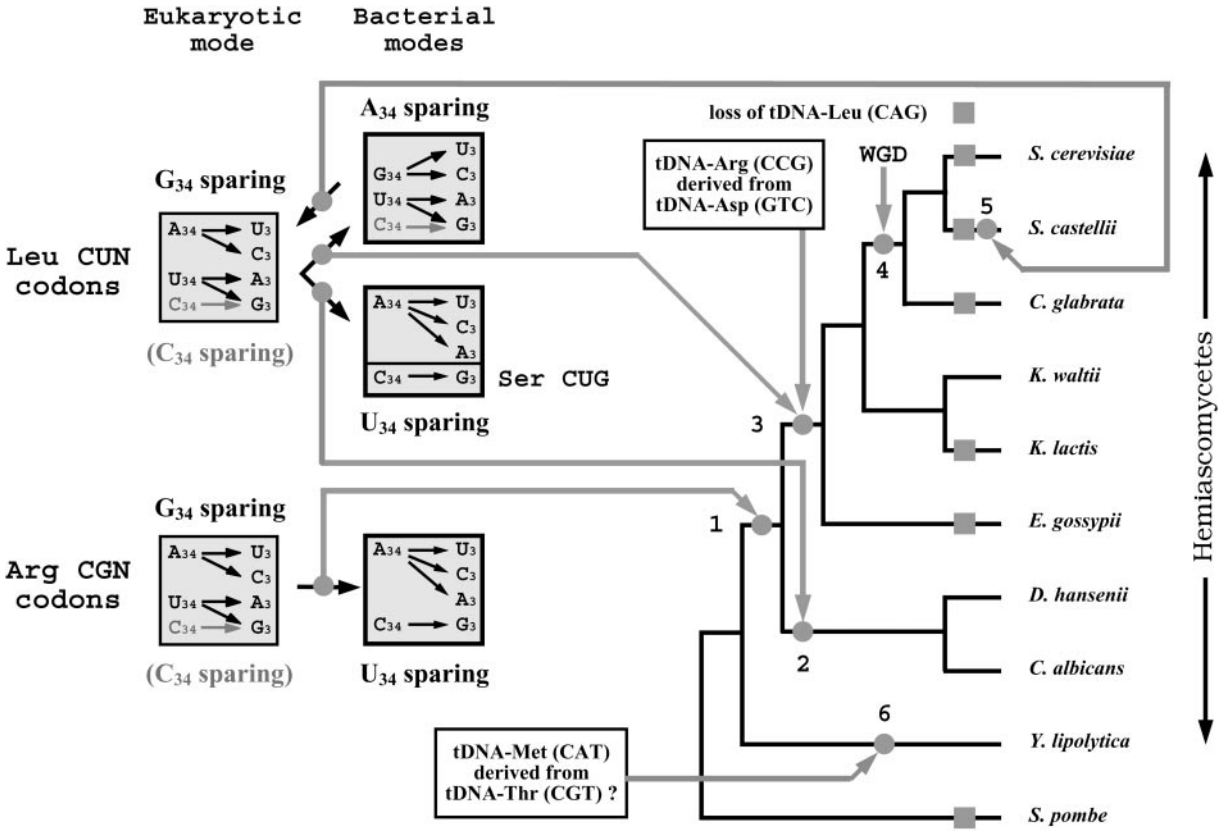


Figure 8. Localization of decoding changes for Leu CUN and Arg CGN in the phylogeny of hemiascomycetes. The schematic phylogeny of hemiascomycetes versus *S.pombe* is taken from (22,23). The two four-codon boxes at left show the regular decoding rules of Leu CUN and Arg CGN codons in eukaryotes. In both boxes, the tRNA with anticodon starting with C_{34} (shown in grey) is dispensable. Grey arrows indicate which is the switch in decoding occurring at the targeted node. Node 1: switch in the decoding of arginine CGN codons from G34-sparing to U_{34}-sparing; node 2: change in the genetic code (CUG codon reassigned to serine) in the *Candida* genus and switch in the decoding of leucine CUN codons from G34-sparing to U_{34}-sparing; node 3: switch in the decoding of leucine CUN codons from G_{34} to A_{34}-sparing; node 4: location of the ancestral whole genome duplication (WGD) (104); node 5: *S.castellii* reverts to the standard eukaryotic rule for the decoding of Leu CUN codons. Grey squares at left indicate which species have lost the tDNA-Leu (CAG) (C_{34}-sparing); Note the similarity of the decoding of Leu CUN in the *Candida* genus with that of Arg CGN in yeasts from *S.cerevisiae* to *C.albicans*.

C.albicans, the B-box is located in the 5′ strand of the P7 helix and, for *Y.lipolytica,* in the P3 helix. Similarly to the *SNR52* gene, the *RPR1* gene of *S.pombe* is probably a Pol II gene (no A- or B-box and no poly-T terminator could be identified). Exceptions to the B-box consensus were found for *S.cerevisiae*, *D.hansenii* and *C.albicans*. In *S.cerevisiae*, an A nucleotide at the third position (instead of a T, also seen in *SCR1* of *S.cerevisiae* and *S.castellii*) does not prevent TFIIIC recognition (38). We noticed a common variation (C at the fifth position) in *D.hansenii* and *C.albicans* genes and C at the second position (instead of W) is also observed in the *C.albicans SCR1* gene.

The *SCR1* genes from the ten genomes were structurally aligned, based on the conservation of P6 and P8 helices (Figure 7B) and the location of the A- and B-boxes carefully examined with respect to the RNA secondary structure. The A-box was previously located at position 10 of the *S.cerevisiae* gene by Dieci and coworkers (5) (starting nucleotide is U with light grey background in the UGU motif, Figure 7B). Alternatively, the A-box might be located 8 nt downstream (at position 18 of *SCR1*, green background), where the A-box consensus (TRGYnnAnnnG) is nearly satisfied for nine out of the ten genomes. Mutation of GG at position 19–20 of *SCR1* (positions 18 and 19 in tRNA) affects TFIIIC binding, thus suggesting that these 2 nt do belong to the A-box (5). This experimental result fits with the two possible locations for the A-box (starting at 10 or 18 in *S.cerevisae SCR1*). Clearly, in the case of *K.lactis*, none of the two A-box positions reasonably fits the consensus while an A-box, with a single variation (at 3rd position), can be found slightly upstream at position −7. In *SCR1*, the B-box is located 24 to 50 nt downstream the A-box in a region of weak sequence conservation, except in *D.hansenii* and *C.albicans* where the B-box overlaps the 5′ strands of P5e and P5f helices.

DISCUSSION

We present the first comprehensive genome wide analysis of Pol III-dependent genes in ten eukaryotes (nine hemiascomycetes and the archiascomycete *S.pombe*). This exhaustive analysis unearthed several original observations. Unexpected features for decoding were first revealed. Yeasts close to *S.cerevisiae* follow the bacterial sparing rules to decode Leu CUN and Arg CGN codons. Such changes, which are unique among eukaryotes, can be precisely dated on the

phylogeny of hemiascomycetes. As shown in Figure 8, the most ancient switch appears to be the change of decoding Arg CGN codons from the regular eukaryotic to a bacterial-type (node #1). The change in the genetic code that reassigned the CUG codon to Ser occurred later, in the branch leading to the *Candida* genus (*D.hansenii* and *C.albicans*, node #2). Independently, in another branch leading to other hemiascomycetes, including *S.cerevisiae*, the decoding of Leu CUN codons switches from the eukaryotic to bacterial mode (G_{34}- to A_{34}-sparing, node #3). Remarkably, *S.castellii* has reverted to the usual eukaryotic G_{34}-sparing (node #5). The capture of tDNA-Asp leading to a novel tDNA-Arg (CCG) appears to be also concomitant with the events occurring at node #3. Finally, the loss of tDNA-Leu (CAG) seems to have occurred several times independently (in these cases, the CUG codon is read by tRNA-Leu (UAG)).

The large size of the collection of tDNA sequences originating from a single eukaryotic phylum allows extensive comparisons between both orthologous genes (i.e. between yeast species) and paralogous genes within each species. For a given tDNA species (given anticodon), the large variation in the number of gene copies is particularly remarkable [e.g. 1–27 copies for tDNA-Glu (CTC)]. This variation in number is at least partly correlated to variation in codon usage between yeast species. It is also remarkable that within a yeast species, the various gene copies are always (or nearly) identical. Remarkably, specific deviations with respect to the eukaryotic cloverleaf model apply to all gene copies within a genome. For example, the tertiary base pair $T_{15}A_{48}$ present in all five tDNA-Phe (TGG) in *C.albicans* replaces the usual $R_{15}Y_{48}$ pair; G_{21} is found instead of the universal A_{21} in all three tDNA-Met (CAT) in *S.pombe*; the $A_{53}T_{61}$ pair, which makes the outer bases of the B-box, is substituted to $G_{53}C_{61}$ in all nine copies of tDNA-Ala (AGC) in *S.pombe*). This suggests a specific role for such deviations and also the existence of a survey mechanism permanently unifying the different tDNA copies of the same tDNA (same anticodon) within each species.

The sequence homogeneity between orthologous tDNA (tDNA coding for the same amino acid in different genomes) contrasts with the sequence divergence between paralogous tDNAs (tDNAs bearing different amino acid within a same genome) as shown by our p-distance analysis. Note that a similar histogram of distance (Figure 4B) was already reported several years ago with a much more limited tDNA set, insufficient for phylogenetic analysis (86,95). With our new dataset that includes ~600 different tDNA sequences, single clustering of orthologous tDNA was observed for most amino acids, with the sole exception of tDNA from *S.pombe*, offering the opportunity to examine the significance of the exceptions to this rule. A first exception is the close relation between the tDNA-Arg (CCG) and the tDNA-Asp (GTC) in yeasts close to *S.cerevisiae* (87). Actually, the origin of the tDNA-Arg (CCG) in the two related genomes *D.hansenii* and *C.albicans* still appears unclear. While the tDNA-Arg (CCG) of *D.hansenii* sides together with other tDNA-Arg within the main tDNA-Arg cluster, that of *C.albicans* sides into the extra cluster defined by five other tDNA-Arg (CCG) (Figure 4C). For the time being, it seems reasonable to conclude that tDNA-Arg (CCG) from *D.hansenii* is a regular tDNA-Arg, not derived from a tDNA-Asp (GUC) anscestor, and that this is also the case for *C.albicans*. It remains that the emergence of the tDNA-Arg (CCG) (Figure 8, node #3) is complex and that detailed analyses of more genomes are necessary to clarify its origin in the different organisms, including hemiascomycetes. The second exception is the intriguing clustering of the tDNA-Met (CAT) from *Y.lipolytica* into the Thr cluster that suggests a possible case of capture (Figure 8, node #6). Here again, more genomes (close to *Y.lipolytica*) will be needed to conclude unambiguously.

Prior to this work, the definition of the promoter elements in the A-box recognized by TFIIIC was uncertain. We used the most representative class of Pol III genes, the tRNA genes, which always amount to more than 41 different types of genes and more than 100 gene copies per genome (up to 500), to extract the A- and B-boxes genomic signatures. These short sequence elements were searched and retrieved in four other Pol III ncRNAs from the ten genomes (except two cases of probable Pol II transcribed genes in *S.pombe*). Examination of the 39 A- and B-box sequences (Figure 6) shows that the consensus signatures are indeed found always at appropriate locations, with a few sequence exceptions.

Directed mutagenesis experiments have established that the B-box is the most critical region for TFIIIC binding and that the interaction between A-box and TFIIIC is less important to the stability of the DNA–TFIIIC complex (96). Among the 2nd, 4th and 5th positions of the B-box (equivalent to positions 54, 56 and 57 of the tRNA), the 4th position, always occupied by a C, is the most critical and its replacement by G lowers the *in vitro* binding affinity of TFIIIC by 370-fold (96). Only in one case, divergence at the 4th position of the B-box over 39 exists (a T is present instead of C in the *SNR6* gene of *C.albicans*). In accordance with the less prominent role of the A-box, more numerous cases of sequence deviations were observed. Nevertheless, A-boxes were always localized no more than 21 nt away from the 5′ end of the mature products, which fits with a distance of about 25 nt between the A-box and the start of transcription. The shortest A-B distance observed (24 nt) is greater than the minimal distance experimentally determined for the correct binding of TFIIIC (21 nt) (97). In 35 cases over 39, the terminator (poly-T) is followed by A or G, which is indicative of an efficient Pol III termination (15).

In contrast to the high conservation of the A- and B-promoter elements throughout the ten genomes, their locations are highly variable, depending on the gene and on the genome. For example, the *RPR1* B-box, which is external to the mature product in the yeasts from *S.cerevisiae* to *K.waltii*, becomes internal in *C.albicans* and *Y.lipolytica*. This illustrates the adaptability of the Pol III transcription machinery to overcome the additional constraints exerted on an internal B-box at the RNA level. In these ten genomes, cases of dicistronic Pol III genes (98) were searched, but none except the tDNA pairs were found. Preliminary investigations for tDNA pairs in higher eukaryotes also remained unsuccessful, suggesting that this type of organization and the mechanism that maintain species-specific pairs are restricted to yeasts.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Jean-Luc Souciet (Strasbourg) and all the members of the Génolevures Consortium for stimulating discussions. The authors thank Yves Boulard (Saclay) for help in running tRNAscan-SE. The authors acknowledge Valérie de Crécy-Lagard (University of Florida) for her help in improving the manuscript. The sequencing projects of *C.glabrata*, *K.lactis*, *D.hansenii* and *Y.lipolytica* were supported by the Consortium National de Recherche en Génomique (to Génoscope and to Institut Pasteur Génopole), the CNRS (GDR 2354, Génolevures sequencing consortium), the Ministère de la Jeunesse, de l'Éducation et de la Recherche (ACI IMPBio n°IMPB114 'Génolevures en ligne') and the 'Conseil Régional d'Aquitaine' ('Génotypage et Génomique Comparée'). The *Magnaporthe grisea* sequencing project is performed by Ralph Dean, Fungal Genomics Laboratory at North Carolina State University (www.fungalgenomics.ncsu.edu), and Center for Genome Research (www.broad.mit.edu). The *Coprinus cinereus* and *Fusarium graminearum* sequencing projects are performed at the Broad Institute and are supported by the National Research Initiative, which is within the U.S. Department of Agriculture's (USDA's) Cooperative State Research Education and Extension Service, and reviewed through the USDA/NSF Microbial Genome Sequencing Project. E.W. and B.D. are members of Institut Universitaire de France. Funding to pay the Open Access publication charges for this article was provided by Commissariat à l'Énergie Atomique.

Conflict of interest statement. None declared.

REFERENCES

1. Sentenac,A. (1985) Eukaryotic RNA polymerases. *CRC Crit. Rev. Biochem.*, **18**, 31–90.
2. White,R.J. (1998) *RNA Polymerase III Transcription, 2nd edn*. Springer-Verlag/Landes Bioscience, NY.
3. Brow,D.A. and Guthrie,C. (1990) Transcription of a yeast U6 snRNA gene requires a polymerase III promoter element in a novel position. *Genes Dev.*, **4**, 1345–1356.
4. Kachouri,R., Stribinskis,V., Zhu,Y., Ramos,K.S., Westhof,E. and Li,Y. (2005) A surprisingly large RNase P RNA in *Candida glabrata*. *RNA*, **11**, 1064–1072.
5. Dieci,G., Giuliodori,S., Catellani,M., Percudani,R. and Ottonello,S. (2002) Intragenic promoter adaptation and facilitated RNA polymerase III recycling in the transcription of SCR1, the 7SL RNA gene of *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **277**, 6903–6914.
6. Harismendy,O., Gendrel,C.G., Soularue,P., Gidrol,X., Sentenac,A., Werner,M. and Lefebvre,O. (2003) Genome-wide location of yeast RNA polymerase III transcription machinery. *EMBO J*, **22**, 4738–4747.
7. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
8. Bonnerot,C., Pintard,L. and Lutfalla,G. (2003) Functional redundancy of Spb1p and a snR52-dependent mechanism for the 2′-*O*-ribose methylation of a conserved rRNA position in yeast. *Mol. Cell*, **12**, 1309–1315.
9. Moqtaderi,Z. and Struhl,K. (2004) Genome-wide occupancy profile of the RNA polymerase III machinery in *Saccharomyces cerevisiae* reveals loci with incomplete transcription complexes. *Mol. Cell Biol.*, **24**, 4118–4127.
10. Schramm,L. and Hernandez,N. (2002) Recruitment of RNA polymerase III to its target promoters. *Genes Dev.*, **16**, 2593–2620.
11. Camier,S., Dechampesme,A.M. and Sentenac,A. (1995) The only essential function of TFIIIA in yeast is the transcription of 5S rRNA genes. *Proc. Natl Acad. Sci. USA*, **92**, 9338–9342.
12. Geiduschek,E.P. and Kassavetis,G.A. (2001) The RNA polymerase III transcription apparatus. *J. Mol. Biol.*, **310**, 1–26.
13. Dieci,G. and Sentenac,A. (2003) Detours and shortcuts to transcription reinitiation. *Trends Biochem. Sci.*, **28**, 202–209.
14. Ferrari,R., Rivetti,C., Acker,J. and Dieci,G. (2004) Distinct roles of transcription factors TFIIIB and TFIIIC in RNA polymerase III transcription reinitiation. *Proc. Natl Acad. Sci. USA*, **3**, 3.
15. Braglia,P., Percudani,R. and Dieci,G. (2005) Sequence context effects on oligo(dT) termination signal recognition by *Saccharomyces cerevisiae* RNA polymerase III. *J. Biol. Chem.*, **280**, 19551–19562.
16. Fruscoloni,P., Zamboni,M., Panetta,G., De Paolis,A. and Tocchini-Valentini,G.P. (1995) Mutational analysis of the transcription start site of the yeast tRNA(Leu3) gene. *Nucleic Acids Res.*, **23**, 2914–2918.
17. Giuliodori,S., Percudani,R., Braglia,P., Ferrari,R., Guffanti,E., Ottonello,S. and Dieci,G. (2003) A composite upstream sequence motif potentiates tRNA gene transcription in yeast. *J. Mol. Biol.*, **333**, 1–20.
18. Dieci,G., Percudani,R., Giuliodori,S., Bottarelli,L. and Ottonello,S. (2000) TFIIIC-independent *in vitro* transcription of yeast tRNA genes. *J. Mol. Biol.*, **299**, 601–613.
19. Galli,G., Hofstetter,H. and Birnstiel,M.L. (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*, **294**, 626–631.
20. Ciliberto,G., Raugei,G., Costanzo,F., Dente,L. and Cortese,R. (1983) Common and interchangeable elements in the promoters of genes transcribed by RNA polymerase III. *Cell*, **32**, 725–733.
21. Marck,C. and Grosjean,H. (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, **8**, 1189–1232.
22. Dujon,B. (2005) Hemiascomycetous yeasts are the forefront of comparative genomics. *Curr. Opin. Genet. Dev.*, **15**, 614–620.
23. Dujon,B. (2005) Eukaryoyic genome evolution: yeasts zoom in molecular mechanisms. *Trends Genet.*, in press.
24. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., De Montigny,J., Marck,C., Neuveglise,C., Talla,E. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.
25. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.
26. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
27. Kurtzman,C.P. (2003) Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Res.*, **4**, 233–245.
28. Kellis,M., Birren,B.W. and Lander,E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, **428**, 617–624.
29. Dietrich,F.S., Voegeli,S., Brachat,S., Lerch,A., Gates,K., Steiner,S., Mohr,C., Pohlmann,R., Luedi,P., Choi,S. *et al.* (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, **304**, 304–307.
30. Jones,T., Federspiel,N.A., Chibana,H., Dungan,J., Kalman,S., Magee,B.B., Newport,G., Thorstenson,Y.R., Agabian,N., Magee,P.T. *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc. Natl Acad. Sci. USA*, **101**, 7329–7334.
31. Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
32. Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, **13**, 3021–3030.
33. Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33**, D139–D140.
34. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
35. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
36. Kumar,S., Tamura,K., Jakobsen,I.B. and Nei,M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1245.

37. Zwieb,C., van Nues,R.W., Rosenblad,M.A., Brown,J.D. and Samuelsson,T. (2005) A nomenclature for all signal recognition particle RNAs. *RNA*, **11**, 7–13.
38. Lee,J.Y., Evans,C.F. and Engelke,D.R. (1991) Expression of RNase P RNA in *Saccharomyces cerevisiae* is controlled by an unusual RNA polymerase III promoter. *Proc. Natl. Acad. Sci. USA*, **88**, 6986–6990.
39. Ohama,T., Suzuki,T., Mori,M., Osawa,S., Ueda,T., Watanabe,K. and Nakase,T. (1993) Non-universal decoding of the leucine codon CUG in several *Candida* species. *Nucleic Acids Res.*, **21**, 4039–4045.
40. Santos,M.A., Perreau,V.M. and Tuite,M.F. (1996) Transfer RNA structural change is a key element in the reassignment of the CUG codon in *Candida albicans*. *EMBO J*, **15**, 5060–5068.
41. Perreau,V.M., Keith,G., Holmes,W.M., Przykorska,A., Santos,M.A. and Tuite,M.F. (1999) The *Candida albicans* CUG-decoding ser-tRNA has an atypical anticodon stem–loop structure. *J. Mol. Biol.*, **293**, 1039–1053.
42. Kawach,O., Voss,C., Wolff,J., Hadfi,K., Maier,U.G. and Zauner,S. (2005) Unique tRNA introns of an enslaved algal cell. *Mol. Biol. Evol.*, **22**, 1694–1701.
43. Trotta,C.R., Miao,F., Arn,E.A., Stevens,S.W., Ho,C.K., Rauhut,R. and Abelson,J.N. (1997) The yeast tRNA splicing endonuclease: a tetrameric enzyme with two active site subunits homologous to the archaeal tRNA endonucleases. *Cell*, **89**, 849–858.
44. Tocchini-Valentini,G.D., Fruscoloni,P. and Tocchini-Valentini,G.P. (2005) Structure, function, and evolution of the tRNA endonucleases of Archaea: an example of subfunctionalization. *Proc. Natl Acad. Sci. USA*, **102**, 8933–8938.
45. Belfort,M. and Weiner,A. (1997) Another bridge between kingdoms: tRNA splicing in archaea and eukaryotes. *Cell*, **89**, 1003–1006.
46. Trotta,C.R. and Abelson,J. (1999) tRNA splicing: an RNA world add-on or an ancient reaction? In *The RNA World. 2nd edn*. Cold Spring Harbor Laboratory Press, pp. 561–584.
47. Marck,C. and Grosjean,H. (2003) Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. *RNA*, **9**, 1516–1531.
48. Edgell,D.R., Belfort,M. and Shub,D.A. (2000) Barriers to intron promiscuity in bacteria. *J. Bacteriol.*, **182**, 5281–5289.
49. Haugen,P., Simon,D.M. and Bhattacharya,D. (2005) The natural history of group I introns. *Trends Genet.*, **21**, 111–119.
50. Szweykowska-Kulinska,Z., Senger,B., Keith,G., Fasiolo,F. and Grosjean,H. (1994) Intron-dependent formation of pseudouridines in the anticodon of *Saccharomyces cerevisiae* minor tRNA(Ile). *EMBO J.*, **13**, 4636–4644.
51. Johnson,P.F. and Abelson,J. (1983) The yeast tRNATyr gene intron is essential for correct modification of its tRNA product. *Nature*, **302**, 681–687.
52. Motorin,Y., Keith,G., Simon,C., Foiret,D., Simos,G., Hurt,E. and Grosjean,H. (1998) The yeast tRNA:pseudouridine synthase Pus1p displays a multisite substrate specificity. *RNA*, **4**, 856–869.
53. Hellmuth,K., Grosjean,H., Motorin,Y., Deinert,K., Hurt,E. and Simos,G. (2000) Cloning and characterization of the *Schizosaccharomyces pombe* tRNA:pseudouridine synthase Pus1p. *Nucleic Acids Res.*, **28**, 4604–4610.
54. Behm-Ansmant,I., Urban,A., Ma,X., Yu,Y.T., Motorin,Y. and Branlant,C. (2003) The *Saccharomyces cerevisiae* U2 snRNA:pseudouridine-synthase Pus7p is a novel multisite-multisubstrate RNA:Psi-synthase also acting on tRNAs. *RNA*, **9**, 1371–1382.
55. Strobel,M.C. and Abelson,J. (1986) Intron mutations affect splicing of *Saccharomyces cerevisiae* SUP53 precursor tRNA. *Mol. Cell Biol.*, **6**, 2674–2683.
56. Motorin,Y. and Grosjean,H. (1999) Multisite-specific tRNA:m^5C-methyltransferase (Trm4) in yeast *Saccharomyces cerevisiae*: identification of the gene and substrate specificity of the enzyme. *RNA*, **5**, 1105–1118.
57. Jiang,H.Q., Motorin,Y., Jin,Y.X. and Grosjean,H. (1997) Pleiotropic effects of intron removal on base modification pattern of yeast tRNA-Phe: an *in vitro* study. *Nucleic Acids Res.*, **25**, 2694–2701.
58. Pintard,L., Lecointe,F., Bujnicki,J.M., Bonnerot,C., Grosjean,H. and Lapeyre,B. (2002) Trm7p catalyses the formation of two 2′-*O*-methylriboses in yeast tRNA anticodon loop. *EMBO J.*, **21**, 1811–1820.
59. Willis,I., Hottinger,H., Pearson,D., Chisholm,V., Leupold,U. and Söll,D. (1984) Mutations affecting excision of the intron from eukaryotic dimeric tRNA precursor. *EMBO J.*, **3**, 1573–1580.
60. Ikemura,T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.*, **158**, 573–597.
61. Lloyd,A.T. and Sharp,P.M. (1993) Synonymous codon usage in *Kluyveromyces lactis*. *Yeast*, **9**, 1219–1228.
62. Percudani,R., Pavesi,A. and Ottonello,S. (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **268**, 322–330.
63. Ikemura,T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13–34.
64. Raymond,K.C., Raymond,G.J. and Johnson,J.D. (1985) *In vivo* modulation of yeast tRNA gene expression by 5′-flanking sequences. *EMBO J*, **4**, 2649–2656.
65. Hani,J. and Feldmann,H. (1998) tRNA genes and retroelements in the yeast genome. *Nucleic Acids Res.*, **26**, 689–696.
66. Seoighe,C. and Wolfe,K.H. (1999) Updated map of duplicated regions in the yeast genome. *Gene*, **238**, 253–261.
67. DeLotto,R. and Schedl,P. (1984) A *Drosophila melanogaster* transfer RNA gene cluster at the cytogenetic locus 90BC. *J. Mol. Biol.*, **179**, 587–605.
68. Kuhn,R.M., Clarke,L. and Carbon,J. (1991) Clustered tRNA genes in *Schizosaccharomyces pombe* centromeric DNA sequence repeats. *Proc. Natl Acad. Sci. USA*, **88**, 1306–1310.
69. Mao,J., Schmidt,O. and Söll,D. (1980) Dimeric transfer RNA precursors in *S. pombe*. *Cell*, **21**, 509–516.
70. Schmidt,O., Mao,J., Ogden,R., Beckmann,J., Sakano,H., Abelson,J. and Söll,D. (1980) Dimeric tRNA precursors in yeast. *Nature*, **287**, 750–752.
71. Straby,K.B. (1988) A yeast tRNAArg gene can act as promoter for a 5′ flank deficient, non-transcribable tRNASUP6 gene to produce biologically active suppressor tRNA. *Nucleic Acids Res.*, **16**, 2841–2857.
72. Otter,C.A., Edqvist,J. and Straby,K.B. (1992) Characterization of transcription and processing from plasmids that use polIII and a yeast tRNA gene as promoter to transcribe promoter-deficient downstream DNA. *Biochim. Biophys. Acta*, **1131**, 62–68.
73. Auxilien,S., Crain,P.F., Trewyn,R.W. and Grosjean,H. (1996) Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of inosine-34 in the anticodon of transfer RNA. *J. Mol. Biol.*, **262**, 437–458.
74. Grosjean,H., Auxilien,S., Constantinesco,F., Simon,C., Corda,Y., Becker,H.F., Foiret,D., Morin,A., Jin,Y.X., Fournier,M. *et al.* (1996) Enzymatic conversion of adenosine to inosine and to N^1-methylinosine in transfer RNAs: a review. *Biochimie*, **78**, 488–501.
75. Curran,J.F. (1998) Modified nucleosides in translation. In Grosjean,H. and Benne,R. (eds), *Modification and Editing of RNA*. ASM press, Washington DC, pp. 493–516.
76. Agris,P.F. (2004) Decoding the genome: a modified view. *Nucleic Acids Res.*, **32**, 223–238.
77. Takai,K. and Yokoyama,S. (2003) Roles of 5-substituents of tRNA wobble uridines in the recognition of purine-ending codons. *Nucleic Acids Res.*, **31**, 6383–6391.
78. Weissenbach,J., Dirheimer,G., Falcoff,R., Sanceau,J. and Falcoff,E. (1977) Yeast tRNALeu (anticodon U–A–G) translates all six leucine codons in extracts from interferon treated cells. *FEBS Lett.*, **82**, 71–76.
79. Crick,F.H. (1966) Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**, 548–555.
80. Lim,V.I. and Curran,J.F. (2001) Analysis of codon:anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure. *RNA*, **7**, 942–957.
81. Murphy,F.V.IV. and Ramakrishnan,V. (2004) Structure of a purine-purine wobble base pair in the decoding center of the ribosome. *Nature Struct. Mol. Biol.*, **11**, 1251–1252.
82. Yokogawa,T., Suzuki,T., Ueda,T., Mori,M., Ohama,T., Kuchino,Y., Yoshinari,S., Motoki,I., Nishikawa,K., Osawa,S. *et al.* (1992) Serine tRNA complementary to the nonuniversal serine codon CUG in *Candida cylindracea*: evolutionary implications. *Proc. Natl Acad. Sci. USA*, **89**, 7408–7411.
83. Curran,J.F. (1995) Decoding with the A:I wobble pair is inefficient. *Nucleic Acids Res.*, **23**, 683–688.
84. Boren,T., Elias,P., Samuelsson,T., Claesson,C., Barciszewska,M., Gehrke,C.W., Kuo,K.C. and Lustig,F. (1993) Undiscriminating codon

reading with adenosine in the wobble position. *J. Mol. Biol.*, **230**, 739–749.
85. Santos,M.A., Moura,G., Massey,S.E. and Tuite,M.F. (2004) Driving change: the evolution of alternative genetic codes. *Trends Genet.*, **20**, 95–102.
86. Cedergren,R.J., Sankoff,D., LaRue,B. and Grosjean,H. (1981) The evolving tRNA molecule. *CRC Crit. Rev. Biochem.*, **11**, 35–104.
87. Fender,A., Geslain,R., Eriani,G., Giege,R., Sissler,M., Florentz,C., Eddy,S.R. and Durbin,R. (2004) A yeast arginine specific tRNA is a remnant aspartate acceptor. *Nucleic Acids Res.*, **32**, 5076–5086 Print 2004.
88. Dirheimer,G., Keith,G., Dumas,P. and Westhof,E. (1995) Primary, secondary, and tertiary structures of tRNAs. In Söll,D. and RajBhandary,U. (eds), *tRNA: Structure, Biosynthesis, and Function*. ASM press, Washington DC, pp. 93–126.
89. Doyon,F.R., Zagryadskaya,E.I., Chen,J. and Steinberg,S.V. (2004) Specific and non-specific purine trap in the T-loop of normal and suppressor tRNAs. *J. Mol. Biol.*, **343**, 55–69.
90. Hamada,M., Huang,Y., Lowe,T.M. and Maraia,R.J. (2001) Widespread use of TATA elements in the core promoters for RNA polymerases III, II, and I in fission yeast. *Mol. Cell Biol.*, **21**, 6870–6881.
91. Kaiser,M.W., Chi,J. and Brow,D.A. (2004) Position-dependent function of a B block promoter element implies a specialized chromatin structure on the *S.cerevisiae* U6 RNA gene, SNR6. *Nucleic Acids Res.*, **32**, 4297–4305.
92. Tani,T. and Ohshima,Y. (1989) The gene for the U6 small nuclear RNA in fission yeast has an intron. *Nature*, **337**, 87–90.
93. Reich,C. and Wise,J.A. (1990) Evolutionary origin of the U6 small nuclear RNA intron. *Mol. Cell Biol.*, **10**, 5548–5552.
94. Lee,J.Y., Rohlman,C.E., Molony,L.A. and Engelke,D.R. (1991) Characterization of RPR1, an essential gene encoding the RNA component of *Saccharomyces cerevisiae* nuclear RNase P. *Mol. Cell. Biol.*, **11**, 721–730.
95. Cedergren,R.J., LaRue,B., Sankoff,D., Lapalme,G. and Grosjean,H. (1980) Convergence and minimal mutation criteria for evaluating early events in tRNA evolution. *Proc. Natl Acad. Sci. USA*, **77**, 2791–2795.
96. Baker,R.E., Gabrielsen,O.S. and Hall,B.D. (1986) Effects of tRNAtyr point mutations on the binding of yeast RNA polymerase III transcription factor C. *J. Biol. Chem.*, **261**, 5275–5282.
97. Baker,R.E., Camier,S., Sentenac,A. and Hall,B.D. (1987) Gene size differentially affects the binding of yeast transcription factor τ to two intragenic regions. *Proc. Natl. Acad. Sci. USA*, **84**, 8768–8772.
98. Kruszka,K., Barneche,F., Guyot,R., Ailhas,J., Meneau,I., Schiffer,S., Marchfelder,A. and Echeverria,M. (2003) Plant dicistronic tRNA-snoRNA genes: a new mode of expression of the small nucleolar RNAs processed by RNase Z. *EMBO J.*, **22**, 621–632.
99. Galagan,J.E., Calvo,S.E., Borkovich,K.A., Selker,E.U., Read,N.D., Jaffe,D., FitzHugh,W., Ma,L.J., Smirnov,S., Purcell,S. *et al.* (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, **422**, 859–868.
100. The *C.elegans* sequencing consortium (1998) Genome sequence of the nematode *C. elegans:* a platform for investigating biology. *Science*, **282**, 2012–2018.
101. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
102. The *Arabidopsis* genome initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
103. Katinka,M.D., Duprat,S., Cornillot,E., Metenier,G., Thomarat,F., Prensier,G., Barbe,V., Peyretaillade,E., Brottier,P., Wincker,P. *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, **414**, 450–453.
104. Seoighe,C. and Wolfe,K.H. (1999) Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.*, **2**, 548–554.