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

Christian Marck*, Rym Kachouri-Lafond1, Ingrid Lafontaine2, Eric Westhof1, Bernard Dujon2 and Henri Grosjean3

Service de Biochimie et de Génétique Moléculaire, Bât 144. CEA/Saclay, 91191 Gif-sur-Yvette, France, 1Institut 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, 2Unité de Génétique Moléculaire des Levures, Institut Pasteur, 25 rue du Dr Roux, 75724 Paris, France and 3Laboratoire 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 50 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 A420 (7,8) which was previously considered to be a Pol II product; and ZOD1, whose function remains unknown (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...
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 promoter 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 GTTTCGA n n C, 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 G18 and G19 [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 Chordata 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 archiascomycte 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 archiascomycte 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 anticondons 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 G19C36 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
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 CCG 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 three 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 and in and around the mature product sequences using the consensus tryGnnAnmmG 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 methionine (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; ‘1’ 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 ‘2’ 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 G10; open boxes indicate other exceptions (see Figure 5C). The indications ‘∆A’, ‘∆G’, ‘∆U’ and ‘∆C’ emphasize the lack of tRNA bearing anticodon starting with the indicated nucleotide (first base of anticodon). The ‘A or G’ sparing rule (‘AG’ or ‘AA’) as well as ‘AU’ and ‘AC’ 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 different 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.
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 multicyclic 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...
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 correlation.
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 phylotype-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 ‘A34 or G34-sparing’ in which either an A34- or G34-containing tRNA decodes the two pyrimidine ending codons. In these cases, the A34 is always posttranscriptionally modified into inosine (I34) [reviewed in (73)] while the G34, is often modified into Gm34 or Q34 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 14 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 ‘ΔA34’ or ‘ΔG34’ 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 ‘A34 or G34-sparing’ (15)).

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 G34-sparing’ strategy, as expected (tRNA-Leu (A34AG). The five hemiascomycetes S.cerevisiae, C.glabrata, K.wallii, C.albicans and S.castellii, lack a few more tRNAs with anticodon starting with C34 (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 ‘reertoire’ (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).
a UAG anticodon (as in *S.cerevisiae*, *C.glabrata*, *K.lactis*, *E.gossypii*, *S.castellii* and *S.pombe*) in which U₃₄ is not posttranscriptionally modified (78), or by two tRNAs (in *K.waltii* and *Y.lipolytica*), one with U₃₄AG and the other with C₃₄AG 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 U₃₄ of this tRNA-Leu (UAG) is modified in other yeasts. The peculiar decoding of

![Diagram](image_url)
the Leu CUN codons in \textit{D. hansenii} and \textit{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 \textit{Y. lipolytica} (as the archiascomycete \textit{S. pombe}), uses the typical ‘Eukaryotic-type G34-sparing’ strategy, all other hemiascomycetes use a third type of sparing, known as ‘U34-sparing’ strategy, which is specific to arginine CGN codons and only known in Bacteria (21). To read Arg CGN codons, \textit{Y. lipolytica} and \textit{S. pombe} use a tRNA-Arg (A34CG) reading CGU and CGC codons and a tRNA-Arg (U34CG) reading CGA codons [and also CGG codons if the tRNA-Arg (U34CG) is absent, as in \textit{Y. lipolytica}]. In this case, U34 is possibly modified into a yet unknown derivative of the type mcmU like in tRNA-Arg [mcmU34 (C34)] (33)]. In all other eight hemiascomycetes, the tRNA-Arg (U34CG) is and a single tRNA-Arg (A34CG) reads the three Arg codons CGU, CGC and also CGA. In this case, A34 is probably modified into I34 and the decoding of CGA codon involves a wobble I34A34 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 (C34CG) and, in these eight organisms obeying the U34-sparing strategy, this tRNA becomes essential, while it may be absent when the tRNA-Arg (U34CG) is present as in \textit{Y. lipolytica}, \textit{D. melanogaster} and \textit{Encephalitozoon cuniculi} (usual C34-sparing strategy) (Figure 3C).

Two unusual tRNA decode Leu CUU, CUc, CUA and Ser CUG codons in \textit{D. hansenii} and \textit{C. albicans}

In \textit{D. hansenii} and \textit{C. albicans}, the ‘Leu’ codon, CUG, is read as Ser (40) (Figure 3A and B). The tRNA-Leu (U34AG), 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 (A34AG), 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 I34A34 wobble pairing exists in the codon-anticodon pairing during translation on the ribosome. This decoding strategy is analogous to the ‘Bacterial-type U34-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 (U34AG) in both \textit{D. hansenii} and \textit{C. albicans} (U34-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 (\textit{D. hansenii} and \textit{C. albicans}). For this tRNA, we do not know what is its decoding capability compared to a more ‘normal’ tRNA and whether A34 is posttranscriptionally modified into I34 (83,84).

The presence of G32 or G33 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 (A34AG) [possibly (I34AG)] and CUG as serine in the case of tRNA-Ser (C34AG). The coexistence of these two types of unusual tRNA among the tRNA population within the same organism (\textit{D. hansenii} and \textit{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

\textbf{Figure 4. Distance tree analysis of 603 tDNA isoacceptor sequences from nine hemiascomycetous genomes and \textit{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 anticonds 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 \textit{S. pombe} (SCP0) 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 incomplete 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 including two neighbour isoacceptors. (3) The gene of the tDNA-Arg (CCG) of \textit{S. cerevisiae} which is related to the tRNA-Asp (TTC) 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 \textit{D. hansenii} (DEHA) and \textit{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 \textit{D. hansenii} (DEHA) and \textit{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 \textit{D. hansenii} clusters here and not in the extra Arg (CCG) cluster].
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 A14 is always occupied or not, thus generating a total of six possible patterns. (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 Figure 1. (indicated by +)
is that of tDNA missing in the elongator tDNA-Met cluster (light box) [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 AUG 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 (G38 and U73 into C38 and G73, 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 TFIIC

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 TFIIC. 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 the 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 (T3, G10, A14, G19 and A21 in the A-box; G53, T55, C56, A58 and C61 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 G10 are mostly found in the tDNA-Leu and tDNA-Ser of D.hansenii,
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 A18 (instead of G18), which probably allows a A18U53 tertiary base pair instead of the bifurcated GU pair in the tRNA transcript (88,89).

In S. pombe, the single copy tRNA-Arg (CCG) harbours an unusual A18U53 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 G18 or G19, as shown in Figure 5E) is always found, in the genomic sequences, four bases downstream of the universally conserved A14. 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 monocopy 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).

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 posttranscriptionally (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 (‘ΔA/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.
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. cerevisiae 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 (G34 to A34-sparing, node #3). Remarkably, S.castellii has reverted to the usual eukaryotic G34-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 T15G48 is present in all five tDNA-Phe (TGG) in C.albicans replaces the usual R15Y48 pair; G21 is found instead of the universal A21 in all three tDNA-Met (CAT) in S.pombe; the A33T61 pair, which makes the outer bases of the B-box, is substituted to G33C61 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) ancestor, 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 RPRI 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 Généoleurs 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, Généoleurs sequencing consortium), the Ministère de la Jeunesse, de l’Éducation et de la Recherche (ACI IMPBio n°IMPB114 ‘Généoleurs 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 RNA P RNA in Candida glabrata. RNA, 11, 1064–1072.
5. Dieci, G., Biard, S., Jeong, S., Stribinskis, V., Zhu, Y., Ramos, K.S., Westhof, E. and Li, Y. (2005) Intragenic promoter adaptation and facilitated RNA polymerase III recycling in the transcription of S. cerevisiae, the 7SL RNA gene for Saccharomyces cerevisiae. J. Biol. Chem., 277, 6903–6914.
6. Harisbeyn, O., Gendrel, C., Soulard, 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. Lowc, T.M. and Eddy, S.R. (1999) A computational screen for methylation switch snoRNAs in yeast. Science, 283, 1168–1171.
8. Bonneret, C., Frintz, L. and Luftz, G. (2003) Functional redundancy of Spb1p and a snR52-dependent mechanism for the 2-0 ribose methylation of a conserved RNA position in yeast. Mol. Cell, 12, 1309–1315.
9. Moqtaderi, Z. and Sh ruth, 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., Dechamps, F. and Sentenac, A. (1995) The essential function of TFIIF 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 initiation. Trends Biochem. Sci., 28, 202–209.
14. Ferrari, R., Rivetti, A., X., Acker, J. and Dieci, G. (2004) Distinct roles of transcription factors TFIIIB and TFIIIC in RNA polymerase III transcription initiation. 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) Mutualational analysis of the transcription start site of the yeast rRNA (Leu3) gene. Nucleic Acids Res., 23, 2914–2918.
17. Guiodiori, S., Percudani, R., Braglia, P., Ferrari, R., Guarnieri, E., Onotello, S. and Dieci, G. (2003) A composite upstream sequence motif potentiates rRNA gene transcription in yeast. J. Mol. Biol., 333, 1–20.
18. Dieci, G., Percudani, R., Guiodiori, S. and Bottarelli, L. and Onotello, S. (2000) TFIIIC-dependent in vitro transcription of yeast rRNA genes. J. Mol. Biol., 299, 601–613.
19. Galli, G., Hoffstetter, H. and Birnstiel, M.L. (1981) Two conserved sequence blocks within eukaryotic rRNA 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. Marc, C. and Grosjean, H. (2002) rNontomics: analysis of RNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticond-sparing strategies and domain-specific features. RNA, 8, 1189–1232.
22. Dujon, B. (2005) Hemiascomycetes yeasts are the forefront of comparative genomics. Curr. Opin. Genet. Dev., 15, 614–620.
23. Dujon, B. (2005) Eukaryotic genome evolution: yeasts zoom in molecular mechanisms. Trends Genet., in press.
24. Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., Montigny, J., Marc, C., Neveglickc, C., Talla, E. and et al. (2004) Genome evolution in yeasts. Nature, 430, 35–44.
25. Golfeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Gilbert, F., Hoheisel, J.D., Jacq, C., Johnston, M. and et al. (1996) Life with 6000 genes. Science, 274, 546, 563–567.
26. Clift, P., Sadrasanam, P., Desikan, A., Fulton, L., Fulton, B., Major, 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 Saccharomycetaeae, and the proposal of the new genera Lachancea, Nakasosmes, Naumovia, Vanderwalltozyma and Zygotortalaspora. 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., Lerach, A., Gates, K., Steiner, S., Mohr, C., Pohmann, R., Lupe, P., Choi, S. and et al. (2004) The Ashtya 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., Thorstensen, Y.R., Agabian, N., Magee, P.T. and 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., Sgourou, P., Peat, N., Hayles, J., Baker, S. and 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 Vassileenko, K.S. (2005) Compilation of RNA sequences and sequences of rRNA genes. Nucleic Acids Res., 33, D139–D140.
34. Lowe, T.M. and Eddy, S.R. (1997) rNascan-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., Przyborska,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. (2003) Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: an example of subfunctionalization. *Proc. Natl Acad. Sci. USA*, 102, 8933–8938.
43. Trotta,C.R., Miao,F., Arn,E.A., Stevens,S.W., Ho,C.K., Rauhut,R. and Soll,D. (2000) Transfer RNA genes and retroelements in the yeast genome. *Gene*, 238, 253–261.
44. Branlant,C. (2003) The tRNA splicing in archaea and eukaryotes. *Archaea: an example of subfunctionalization.* 2nd edn, ASM press, pp. 561–584.
45. Trotta,C.R., Miao,F., Arn,E.A., Stevens,S.W., Ho,C.K., Rauhut,R. and Soll,D. (2000) Transfer RNA genes and retroelements in the yeast genome. *Gene*, 238, 253–261.
46. DeLotto,R. and Schedl,P. (1984) A Drosophila melanogaster transfer RNA gene cluster at the cytogenetic locus 90BC. *J. Mol. Biol.*, 179, 587–605.
47. Kuhn,R.M., Clarke,L. and Carbon,J. (1991) Clustered tRNA genes in *Saccharosaccharomyces pombe* centromeric DNA sequence repeats. *Proc. Natl Acad. Sci. USA*, 88, 1306–1310.
48. Edgell,D.R., Belfort,M. and Shub,D.A. (2000) Barriers to intron function in yeast tRNA gene as promoter to transcribe promoter-deficient downstream RNAs. *Biochim. Biophys. Acta*, 1516–1531.
49. Schmidt,O., Mao,J., Ogden,R., Beckmann,J., Sakano,H., Abelson,J. and Soll,D. (1980) Dimeric transfer RNA precursors in S. pombe. *Cell*, 21, 509–516.
50. Schmidt,O., Mao,J., Ogden,R., Beckmann,J., Sakano,H., Abelson,J. and Soll,D. (1980) Dimeric transfer RNA precursors in yeast. *Nature*, 287, 750–752.
51. Johnson,P.F. and Abelson,J. (1998) Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of isoaccepting 34 in the anticodon of transfer RNA. *J. Mol. Biol.*, 262, 437–458.
52. Kuhn,R.M., Clarke,L. and Carbon,J. (1991) Clustered tRNA genes in *Saccharosaccharomyces pombe* centromeric DNA sequence repeats. *Proc. Natl Acad. Sci. USA*, 88, 1306–1310.
53. Hampton,J.R. and Soll,D. (1987) Transfer RNA gene as promoter to transcribe promoter-deficient downstream RNAs. *Biochim. Biophys. Acta*, 1131, 62–68.
54. Johnson,P.F. and Abelson,J. (1983) The yeast tRNA*ye* gene is essential for correct modification of its tRNA product. *Nature*, 302, 681–687.
55. Lim,V.I. and Curran,J.F. (2001) Analysis of codon:anticodon interaction in *Saccharomyces cerevisiae*: a review. *Biochim. Biophys. Acta*, 1516–1531.
56. Lim,V.I. and Curran,J.F. (2001) Analysis of codon:anticodon interaction in *Saccharomyces cerevisiae*: a review. *Biochim. Biophys. Acta*, 1516–1531.
57. Weissenbach,J., Dirheimer,G., Falloff,R., Sanseau,J. and Falloff,E. (1977) Yeast tRNA*leu* (anticodon U–A–G) translates all six leucine codons in extracts from interferon treated cells. *FEBS Lett.*, 82, 71–76.
58. 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.
59. Murphy,F.V.IV. and Ramakrishnan,V. (2004) Structure of a tRNA:pseudouridine synthase Pus1p. *Nucleic Acids Res.*, 32, 1235–1238.
60. Murphy,F.V.IV. and Ramakrishnan,V. (2004) Structure of a tRNA:pseudouridine synthase Pus1p. *Nucleic Acids Res.*, 32, 1235–1238.
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., Erioni, G., Giewe, 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.

88. Dirheimer, G., Keith, G., Dumas, P. and Westhof, E. (1995) Primary, secondary, and tertiary structures of tRNAs. In Soll, 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 Maraiya, 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 Brown, 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 rRNAtrp 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 T to two intragenic regions. Proc. Natl. Acad. Sci. USA, 84, 8768–8772.

98. Kruszka, K., Barneche, F., Guyot, A., Alphas, 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., Cornu, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaille, E., Brottier, P. and 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.