*Nucleic Acids Research, 2006, Vol. 34, Database issue* ***D145–D149***
*doi:10.1093/nar/gkj084*

# MODOMICS: a database of RNA modification pathways

**Stanislaw Dunin-Horkawicz[1], Anna Czerwoniec[2], Michal J. Gajda[1], Marcin Feder[1], Henri Grosjean[3] and Janusz M. Bujnicki[1,2,*]**

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology ul. Ks., Trojdena 4, PL-02-190 Warsaw, Poland, [2]Bioinformatics Laboratory, Institute for Molecular Biology and Biotechnology, Adam Mickiewicz University, Umultowska 89, PL-61-614 Poznan, Poland and [3]Laboratoire d'Enzymologie et Biochimie Structurales, CNRS, Bld 34, 1 Avenue de la Terrasse, F-91198 Gif-sur-Yvette, France

Received August 15, 2005; Accepted October 12, 2005

## ABSTRACT

**MODOMICS is the first comprehensive database resource for systems biology of RNA modification. It integrates information about the chemical structure of modified nucleosides, their localization in RNA sequences, pathways of their biosynthesis and enzymes that carry out the respective reactions. MODOMICS also provides literature information, and links to other databases, including the available protein sequence and structure data. The current list of modifications and pathways is comprehensive, while the dataset of enzymes is limited to *Escherichia coli* and *Saccharomyces cerevisiae* and sequence alignments are presented only for tRNAs from these organisms. RNAs and enzymes from other organisms will be included in the near future. MODOMICS can be queried by the type of nucleoside (e.g. A, G, C, U, I, $m^1A$, $nm^5s^2U$, etc.), type of RNA, position of a particular nucleoside, type of reaction (e.g. methylation, thiolation, deamination, etc.) and name or sequence of an enzyme of interest. Options for data presentation include graphs of pathways involving the query nucleoside, multiple sequence alignments of RNA sequences and tabular forms with enzyme and literature data. The contents of MODOMICS can be accessed through the World Wide Web at http://genesilico.pl/modomics/.**

## INTRODUCTION

Naturally occurring RNAs contain numerous chemically altered nucleosides. They are formed by enzymatic modification of the primary transcripts during the complex RNA maturation process. To date, over 100 structurally distinguishable modified nucleosides originating from different types of RNAs from many diverse organisms of the three major phylogenetic domains of life have been reported and collected in the RNAMods database (1,2). The location and distribution of various types of modification vary greatly between different RNA molecules, organisms and organelles. The largest number of modified nucleosides with the greatest structural diversity is found in transfer RNAs (tRNAs) (3). Other types of RNA (rRNA, mRNA, snRNA, snoRNA and even the recently discovered miRNA) also contain modified nucleosides; however, they cannot match tRNAs with respect to abundance and diversity of modifications [see ref. (4) and the collection of reviews in books (5,6)].

## DATABASE CONTENT

We have collected all the currently known RNA modifications from the RNAMods database (2) and the tRNA sequences from the Bayreuth database (3). We have extended the set of RNA modifications to include the most recent additions reported in the literature and we have carefully refined the type and location of modifications in *Escherichia coli* and *Saccharomyces cerevisiae* tRNA sequences based on an extensive survey of the published data. Furthermore, based on a very extensive literature search we have compiled the set of known and predicted modification reactions (and the responsible enzymes, if available) that lead from the unmodified precursors to all known modified nucleosides. For all RNA modification enzymes from *E.coli* and *S.cerevisiae*, we have collected the available information concerning alternative gene and protein names, amino acid sequence, 3D structure, cofactor requirements and literature (PubMed) references to the key analyses. To date, such a comprehensive set of

*To whom correspondence should be addressed. Tel: +48 22 597 0750; Fax: +48 22 597 0715; Email: iamb@genesilico.pl
Correspondence may also be addressed to Henri Grosjean. Tel: +33 1 6982 3468, Fax: +33 1 6982 3129; Email: henri.grosjean@lebs.cnrs-gif.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

metabolic reactions involving nucleosides in RNA has never been reported. Similarly, our collection of enzymes is much more complete, accurate and up-to-date than the subsets of proteins involved in metabolism of RNA modifications stored in general-purpose enzyme and genome databases, such as BRENDA (7), EcoCyc (8), SGD (9), etc.

## DATABASE ORGANIZATION AND ACCESS

MODOMICS is a relational database that links together the datasets of modification reactions/pathways, RNA-modifying enzymes and the sequences of target RNAs, which can be queried via three convenient menus, 'PATHWAYS', 'ENZYMES' and 'RNAs'. The modification pathways are represented as a set of graphs visualized with GRAPHVIZ (http://www.graphviz.org/), in which the nodes represent the modified nucleosides, and the edges represent the transformations (real or putative) between them, e.g. enzymatic reactions. Currently, the pathway dataset comprises five graphs, corresponding to all known modifications of adenosine (A), cytidine (C), guanosine (G), uridine (U) and the queuosine pathway (Q), of which the graph of U modifications is clearly most diversified. The 'PATHWAYS' menu offers a variety of filtering options to display a whole graph or its fragment corresponding to reactions that occur in a particular kingdom of life (Eukaryota, Archaea, Bacteria or viruses) or an organellum (currently only mitochondrial modifications are supported), in a particular subset or RNAs (tRNAs, rRNAs, mRNAs, small RNAs or chromosomal RNAs), and hypermodifications that result from processing of an already modified nucleoside (e.g. I and its derivatives, Ψ and its derivatives, etc.). All nodes of the (sub)graph, e.g. the images of all nucleosides, are hyperlinked to dynamically generated windows comprising two panels (example shown in Figure 1). The upper panel displays basic information about the selected nucleoside, including its common name, one letter code in the tRNA sequence database, symbol used in MODOMICS and the chemical structure in the SMILES code hyperlinked to the corresponding image generated using the SMI2GIF script kindly provided by the Daylight Chemical Information Systems Inc (http://www.daylight.com/). The lower panel includes a subgraph of reactions leading to the selected nucleoside from its precursor(s) (unless the selected nucleoside is the precursor itself), and all known hypermodifications formed from that nucleoside.

All edges of the (sub)graph, e.g. arrows that connect the images of nucleosides, are hyperlinked to static 'reaction' windows comprising one or more panels (example shown in Figure 2). The upper panel displays basic information about the selected reaction, namely the type of RNA in which it is conducted and its occurrence in the phylogenetic context (inferred from the knowledge of the phylogenetic distribution of the substrate and the product nucleoside). Other panels display information about enzymes known to catalyze the selected reaction in different substrates and in different organisms. Currently, the database of enzymes includes only proteins from *E.coli* and *S.cerevisiae*, but will be expanded in the future and may eventually comprise all orthologs of the functionally characterized enzymes identifiable in fully sequenced genomes. At this moment, however, MODOMICS includes only enzymes and corresponding protein cofactors that have been experimentally validated, of which most (but not all) have known amino acid sequences. The enzyme panels are also directly accessible from the 'ENZYMES' menu, which lists all entries of the current database in the tabular form, which can be sorted by the name(s) (e.g. Trm1, Dus2, MnmC, etc.), enzyme type (e.g. methyltransferase, deaminase, etc.) or the organism of origin. The enzyme database can be also searched by identifying entries, whose fields match the query formulated as a regular expression (e.g. 'Trm*', '*pus*', 'm1*' or '*transferase', etc.). The enzyme names include both the traditional one(s), which have been quite erratically inferred from different characteristics of the gene, the protein, the organism, the type of the reaction, or even the whole pathway (e.g. Tgs1p, Abd1p, Gar1p, Nop1p, MnmA, GidA, etc.), as well as a novel name given according to the newly developed, uniform nomenclature (H.G. and J.M.B., manuscript in preparation). The enzyme panel lists all relevant types of reactions between nucleosides in different substrates and provides the EC numbers (if available) and links to the corresponding entries in the BRENDA database (7). Other information about the enzyme concerns the sequence (if available) and includes the name of the open reading frame in the genome, the NCBI GenPept (10) and SwissProt (11) accession numbers, the experimentally solved structure in the PDB (12) (in the future it will also include experimentally validated computational models) and the literature information concerning experimental characterization of the protein, identification of the gene, and possibly a structural analysis, together with links to the relevant database entries.

The 'RNAs' menu allows displaying multiple sequence alignments of homologous RNA families, in which the modifications have been observed. Currently, this dataset is complementary to the enzyme dataset and includes only tRNA sequences from *S.cerevisiae* (both cytoplasmic and mitochondrial) and *E.coli*, and the corresponding unmodified tDNA sequences. In the future it will be extended to manually validated tRNA sequences from other model organisms, for which the tRNA data are available (3), as well as to tDNA sequences predicted from the fully sequenced genomes [e.g. the tRNomics dataset from ref. (13)]. It will also include other RNA sequences with identified modifications, such as rRNAs and various small non-coding RNAs. The content of the currently available tRNA sequence alignment can be filtered according to several options, including the organism and strain or taxon of origin, anticodon and amino acid specificity.

## AVAILABILITY

The data are accessible freely for research purposes at the URL http://genesilico.pl/modomics/. Images of pathways are available by Email request to the first author (at sdh@genesilico.pl). Scientists interested in adding data (modifications, RNA sequences, enzymes, etc.) or in using the enzyme information in MODOMICS to predict the modification status of RNA molecules from the DNA sequence are encouraged to contact the senior authors (at iamb@genesilico.pl or henri.grosjean@lebs.cnrs-gif.fr). This article should be cited in research projects assisted by the use of MODOMICS.

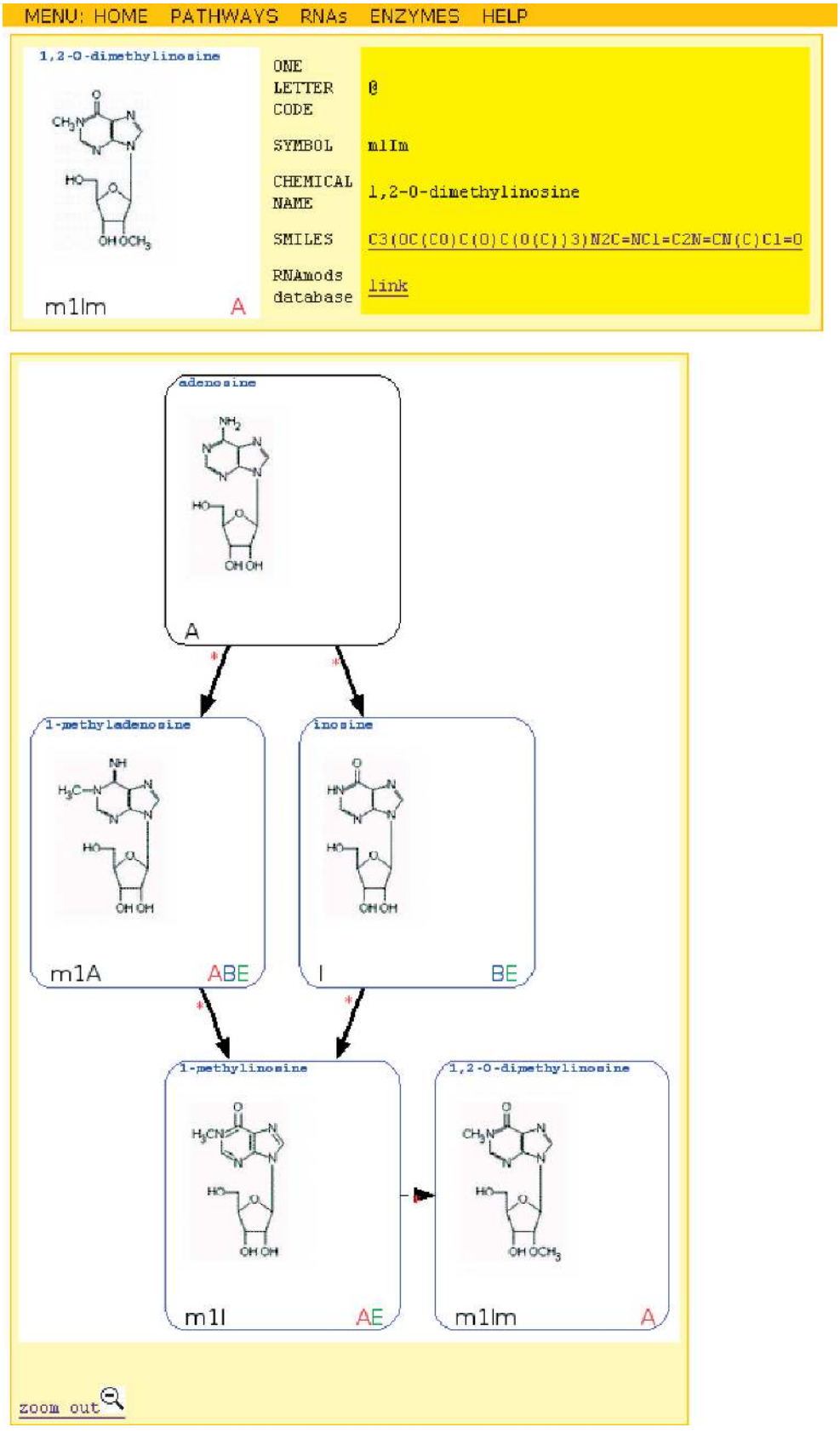


**Figure 1.** A subgraph of adenosine modifications, showing the pathways of $m^1Im$ synthesis (14). Experimentally verified reactions catalyzed by known enzymes from the MODOMICS dataset are indicated by solid arrows, putative reactions or those catalyzed by unknown enzymes are indicated by broken arrows (in this example: $m^1I \rightarrow m^1Im$). The phylogenetic occurrence of nucleosides is shown by colored letters: Archaea (red A), Bacteria (blue B) and Eukaryota (green E).

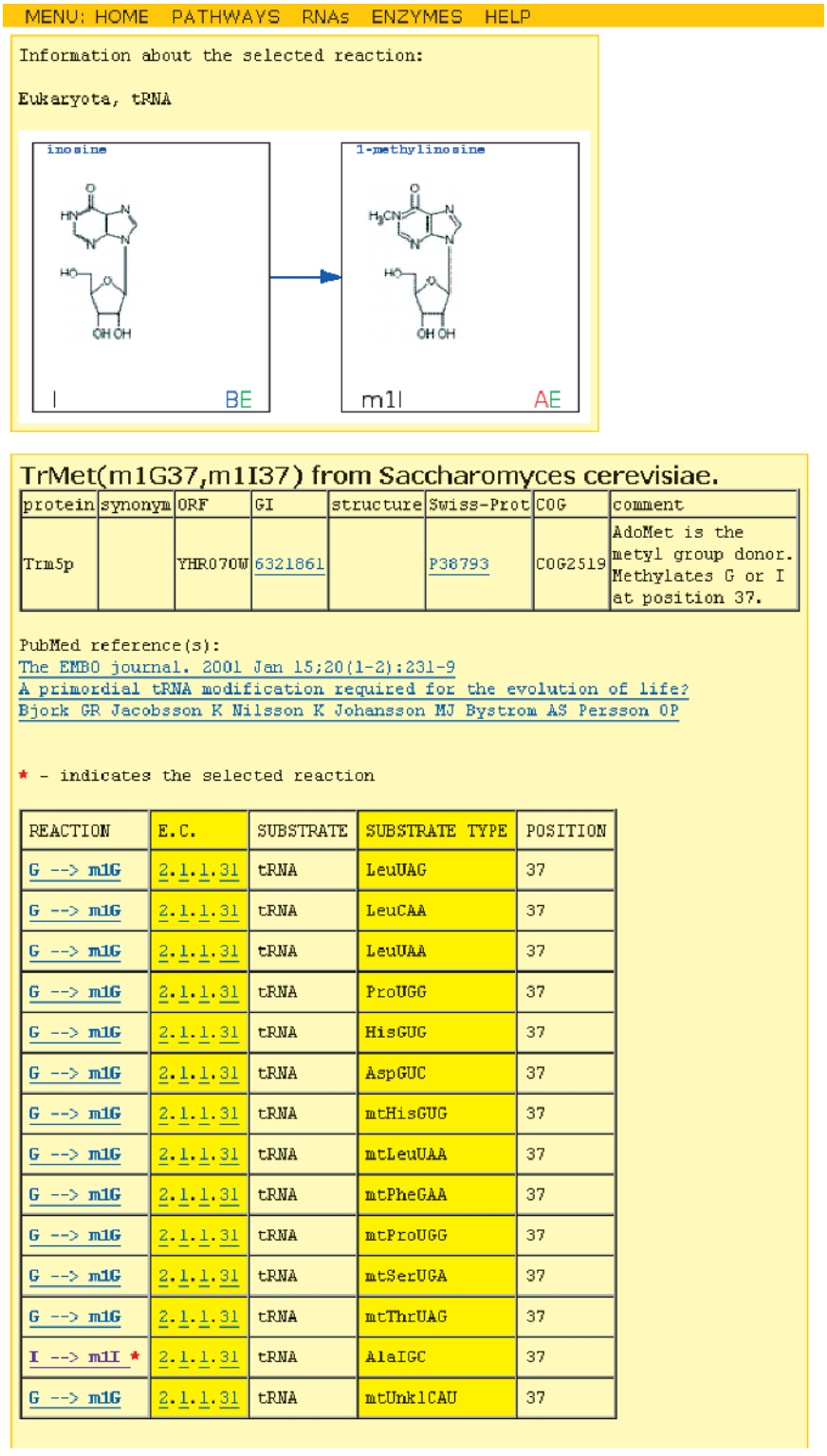


**Figure 2.** Reaction panel illustrates the methylation of inosine (I) to $N^1$-methylinosine ($m^1I$) catalyzed by the Trm5p enzyme at the position 37 of the anticodon loop of $tRNA^{Ala}$ (anticodon IGC), which also introduces the $m^1G$ modification in a larger subset of other yeast tRNAs (15). The hyperlink to the G→$m^1G$ reaction reveals a similar panel (data not shown) with Trm5p, as well as three other enzymes, namely tRNA MTases Trm10p in *S.cerevisiae* (16) and TrmD (17), and rRNA MTase RlmA(I) (18) (both in *E.coli*). Note that in Archaeal tRNAs, the enzymatic formation of $m^1I$, which exclusively occurs at the position 57 of the Ψ-loop, follows a completely different metabolic route [data not shown, for details see (14,19)].

## ACKNOWLEDGEMENTS

The authors are grateful to Daylight Chemical Information Systems Inc. for providing their SMI2GIF service. MODOMICS would not be possible without the data from publicly available databases, in particular the RNAMods database and the Bayreuth tRNA dataset. J.M.B. was supported by the EMBO/HHMI Young Investigator Programme. S.D.-H. was supported by the NIH (grant 5R01AI056034-02). Travel expenses between Poland and France were covered by EGIDE (Polonium grant #07558WK) and by the CEMBM Center of Excellence grant from the 5th Framework Programme of the European Union (contract #QLK6-CT-2002-90363). The Open Access publication charges for this article were waived by Oxford University Press.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Limbach,P.A., Crain,P.F. and McCloskey,J.A. (1994) Summary: the modified nucleosides of RNA. *Nucleic Acids Res*., **22**, 2183–2196.
2. Rozenski,J., Crain,P.F. and McCloskey,J.A. (1999) The RNA Modification Database: 1999 update. *Nucleic Acids Res*., **27**, 196–197.
3. Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*., **33**, D139–D140.
4. McCloskey,J.A. and Rozenski,J. (2005) The Small subunit rRNA Modification Database. *Nucleic Acids Res*., **33**, D135–D138.
5. Grosjean,H. and Benne,R. (1998) *Modification and Editing of RNA*. ASM Press, Washington, DC.
6. Grosjean,H. (2005) *Fine-tuning of RNA Functions by Modification and Editing*. Springer-Verlag, Berlin-Heidelberg.
7. Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*., **32**, D431–D433.
8. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res*., **33**, D334–D337.
9. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al*. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res*., **32**, D311–D314.
10. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
11. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.
12. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al*. (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res*., **33**, D233–D237.
13. 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.
14. Grosjean,H., Constantinesco,F., Foiret,D. and Benachenhou,N. (1995) A novel enzymatic pathway leading to 1-methylinosine modification in *Haloferax volcanii* tRNA. *Nucleic Acids Res*., **23**, 4312–4319.
15. Bjork,G.R., Jacobsson,K., Nilsson,K., Johansson,M.J., Bystrom,A.S. and Persson,O.P. (2001) A primordial tRNA modification required for the evolution of life? *EMBO J*., **20**, 231–239.
16. Jackman,J.E., Montange,R.K., Malik,H.S. and Phizicky,E.M. (2003) Identification of the yeast gene encoding the tRNA $m^1G$ methyltransferase responsible for modification at position 9. *RNA*, **9**, 574–585.
17. Bystrom,A.S. and Bjork,G.R. (1982) Chromosomal location and cloning of the gene (*trmD*) responsible for the synthesis of tRNA ($m^1G$) methyltransferase in *Escherichia coli* K-12. *Mol. Gen. Genet*., **188**, 440–446.
18. Gustafsson,C. and Persson,B.C. (1998) Identification of the *rrmA* gene encoding the 23S rRNA $m^1G745$ methyltransferase in *Escherichia coli* and characterization of an m1G745-deficient mutant. *J. Bacteriol*., **180**, 359–365.
19. Roovers,M., Wouters,J., Bujnicki,J.M., Tricot,C., Stalon,V., Grosjean,H. and Droogmans,L. (2004) A primordial RNA modification enzyme: the case of tRNA ($m^1A$) methyltransferase. *Nucleic Acids Res*., **32**, 465–476.