D396–D398 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki130

MDS_IES_DB: a database of macronuclear and micronuclear genes in spirotrichous ciliates

Andre R. O. Cavalcanti, Thomas H. Clarke and Laura F. Landweber*

Department of Ecology and Evolutionary Biology, Princeton University, NJ 08544, USA

Received August 11, 2004; Revised and Accepted October 25, 2004

ABSTRACT

Ciliated protozoa have two kinds of nuclei: Macronuclei (MAC) and Micronuclei (MIC). In some ciliate classes, such as spirotrichs, most genes undergo several layers of DNA rearrangement during macronuclear development. Because of such processes, these organisms provide ideal systems for studying mechanisms of recombination and gene rearrangement. Here, we describe a database that contains all spirotrich genes for which both MAC and MIC versions are sequenced, with consistent annotation and easy access to all the features. An interface to query the database is available at http://oxytricha.princeton.edu/dimorphism/database.htm.

INTRODUCTION

Ciliates are microbial eukaryotes characterized by the presence of nuclear dimorphism—each ciliate cell contains two kinds of nuclei: a somatic macronucleus (MAC)—which provides templates for the transcription of all genes required for vegetative growth, and a germline nucleus—the micronucleus (MIC)—used for the exchange of meiotic products during conjugation (sexual reproduction).

In spirotrichous (formerly hypotrichous) ciliates, the MAC genome consists of thousands of gene-sized chromosomes (also referred to as 'nanochromosomes'), which exist in high copy number (~1000 copies). These molecules are assembled from sequences in the MIC genome called macronuclear destined sequences (MDSs). Within the MIC sequences, short AT-rich non-coding sequences called internal eliminated sequences (IESs) bound by short repeats (Pointers) interrupt the MDSs. IESs are precisely excised during macronuclear development (1).

In several genes, the order of the MDSs in the MIC sequence does not parallel their order in the MAC sequence. These are called scrambled genes. Scrambling has been characterized in three different genes: α-telomere binding protein (2,3), actin I (4,5) and DNA polymerase α (6,7). MDSs, IESs and pointers can also be designated as scrambled or non-scrambled based on their location within a gene (7,8).

Because most genes in spirotrichous ciliates undergo several layers of rearrangement during macronuclear development, these organisms provide ideal systems for studying the mechanisms of gene recombination and rearrangement. Currently, several sequences are available for micronuclear and macronuclear versions of spirotrich genes. However, within available public databases, like GenBank, these sequences are difficult to access as many of them have either incomplete or inconsistent annotation. Furthermore, many of the spirotrich macronuclear sequences are annotated under different unpublished guidelines, and for some micronuclear sequences, the MDS, IES and pointer annotations are not available publicly. A further difficulty arises from the fact that the fields supported by these databases are insufficient to describe the complexity of the information in these genes.

To solve these problems we built a database, the MDS_IES_DB, designed to collect all spirotrich genes for which complete or near complete sequences of both micronuclear and macronuclear versions exist. This database should serve as a single location from which the entire set of micronuclear and macronuclear sequence data can be accessed and cross-analyzed. We provide the most consistent and up-to-date annotation of the micronuclear sequences, so that analysis is not biased by differences in annotation.

ANNOTATION OF SPIROTRICHOUS GENES

We collected all ciliate sequences from spirotrichs with either full or nearly completed micronuclear and macronuclear sequences from GenBank, along with unpublished sequences from our own laboratory.

Sequence annotation was extracted from the GenBank files or from the original papers. When the annotation was not available we used the program Gene Unscrambler (9), to automate the annotation process. Pointer sequences are usually defined as the overlap between two consecutive MDS. Following (7), we allowed the presence of one mismatch in the pointers if such a mismatch is followed by a string of three

*To whom correspondence should be addressed. Tel: +1 609 258 1947; Fax: +1 609 258 7892; Email: lfl@princeton.edu

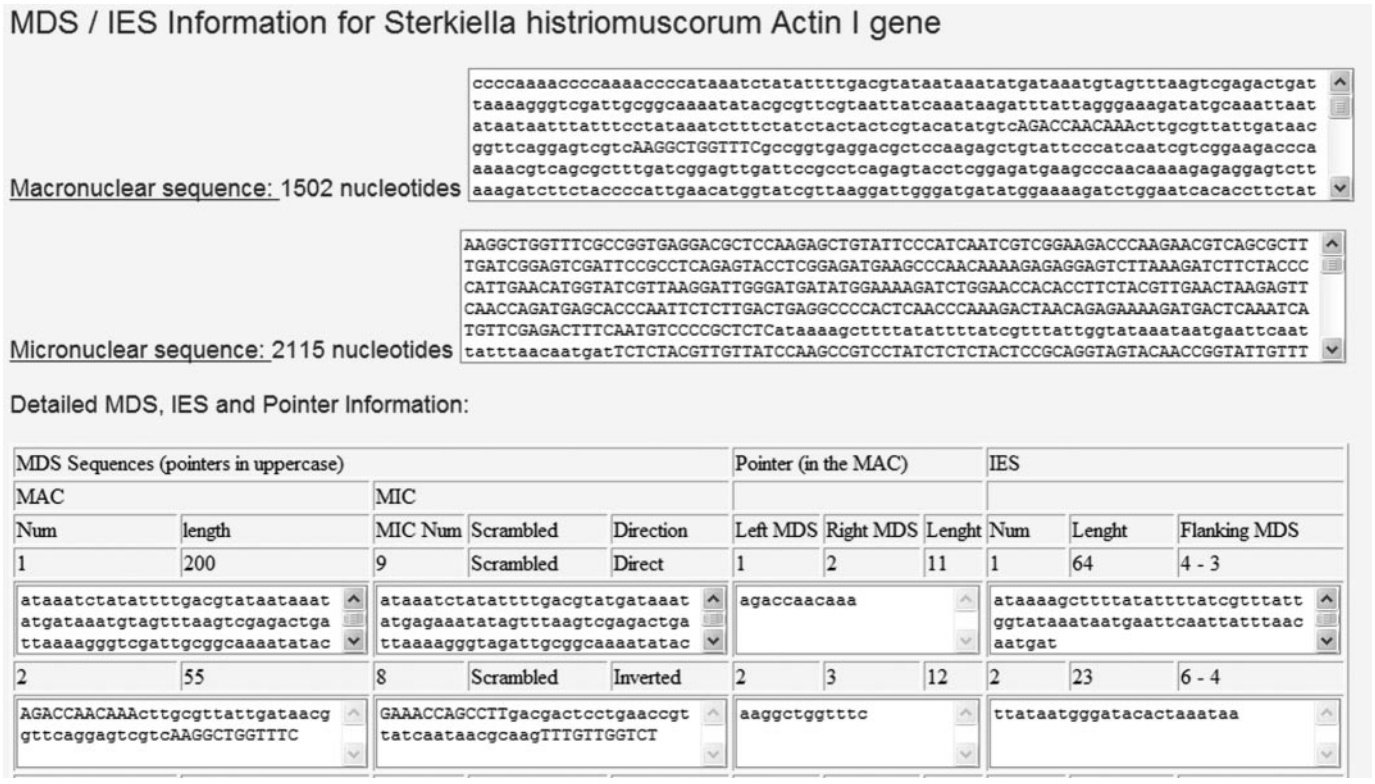

Figure 1. Screen dump of the web interface showing the organization of the *Oxytricha trifallax* (*Sterkiella histriomuscorum*) actin I gene. The first textbox gives the annotated macronuclear sequence of the gene, with MDS and pointers in uppercase and IES and intergenic sequences in lowercase. The second textbox gives the annotated micronuclear sequence of the gene, with MDS and telomeres in lowercase and pointers in uppercase. The MDS, IES and Pointer sequences are given in the lower textboxes. Only the first three MDS, pointers and IES are shown. Note that the sequence of an MDS can be slightly different between the macronuclear and micronuclear versions due to allelic variation. Also in some cases only the macronuclear version of an MDS was sequenced, and in these cases the micronuclear box says 'missing'.

or more consecutive matches. All annotation was manually verified and pointer sequences were extended when possible.

DATABASE DESCRIPTION

The database was built using MySQL (version 4.0.16). The web interface was built with Perl using the CGI, DBI and GD Perl modules. The database consists of four tables: the head table, and tables for MDS, IES and Pointer data.

Within the database, there are 29 different pairs of micronuclear and macronuclear genes. These sequences represent 11 different gene families and 13 different organisms. Twelve of these genes pairs are scrambled, coming from three gene families and eight different organisms. There are three incomplete micronuclear sequences.

The database contains information on 440 MDS pairs (each pair composed of the MIC and the MAC version of a given MDS), 392 IES and 361 pointer triples (each pointer has two active copies in the MIC and one copy in the MAC) (7). Out of the 440 MDSs, 235 are scrambled, and 65 are in the opposite strand in the MIC. A total of 320 IESs and 202 pointers are scrambled.

For each pair of genes in the database the user can see the micronuclear and macronuclear organization and has the option to see all the MDS, IES and pointer sequences (Figure 1). Another option is to download the MIC sequence with the MDSs and pointers in uppercase and the IESs in lowercase. It is also possible to graphically compare the organization of several genes.

We expect this database to grow substantially with the completion of the genome sequencing of *Oxytricha trifallax* (*Sterkiella histriomuscorum*) (10–13).

AVAILABILITY

The MDS_IES_DB is available online at http://oxytricha.princeton.edu/dimorphism/database.htm, together with a program—Gene Unscrambler (9)—to automatically annotate MDSs, Pointers and IESs in macronuclear and micronuclear genes. This web page also contains links to the results of a pilot genome project of the macronucleus of *O.trifallax* (11–13).

A manual with more detailed description of the analyses available is also accessible in the above address. Corrections, new entries, errors and/or omissions and other material for inclusion in the database are welcome and should be sent to acavalca@princeton.edu.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Wei-Jen Chang and Li Chin Wong for contributing sequences to the database before publication and National Science Foundation award 0121422 and National Institute of General Medical Sciences award GM59708 to L.F.L.

REFERENCES

1. Prescott,D.M. (1994) The DNA of ciliated protozoa. *Microbiol. Rev.*, **58**, 233–267.
2. Mitcham,J.L., Lynn,A.J. and Prescott,D.M. (1992) Analysis of a scrambled gene: the gene encoding alpha-telomere-binding protein in *Oxytricha nova*. *Genes Dev.*, **6**, 788–800.
3. Prescott,J.D., DuBois,M.L. and Prescott,D.M. (1998) Evolution of the scrambled germline gene encoding α-telomere binding protein in three hypotrichous ciliates. *Chromosoma*, **107**, 293–303.
4. DuBois,M. and Prescott,D.M. (1995) Scrambling of the actin I gene in two *Oxytricha* species. *Proc. Natl Acad. Sci. USA*, **92**, 3888–3892.
5. Hogan,D.J., Hewitt,E.A., Orr,K.E., Prescott,D.M. and Müller,K.M. (2001) Evolution of IESs and scrambling in the actin I gene in hypotrichous ciliates. *Proc. Natl Acad. Sci. USA*, **98**, 15101–15106.
6. Hoffman,D.C. and Prescott,D.M. (1996) The germline gene encoding DNA polymerase alpha in the hypotrichous ciliate *Oxytricha nova* is extremely scrambled. *Nucleic Acids Res.*, **24**, 3337–3340.
7. Landweber,L.F., Kuo,T.C. and Curtis,E.A. (2000) Evolution and assembly of an extremely scrambled gene. *Proc. Natl Acad. Sci. USA*, **97**, 3298–3303.
8. Prescott,D.M. (2000) Genome gymnastics: unique modes of DNA evolution and processing in ciliates. *Nature Rev. Genet.*, **1**, 191–198.
9. Cavalcanti,A.R.O. and Landweber,L.F. (2004) Gene Unscrambler for detangling scrambled genes in ciliates. *Bioinformatics*, **20**, 800–802.
10. Powell,K. (2002) Second round of gene sequencing goes down to the farm. *Nature*, **419**, 237.
11. Doak,T.G., Cavalcanti,A.R.O., Stover,N., Dunn,D.M., Weiss,R., Herrick,G. and Landweber,L.F. (2003) Sequencing the *Oxytricha trifallax* macronuclear genome: a pilot project. *Trends Genet.*, **19**, 603–607.
12. Cavalcanti,A.R.O., Dunn,D.M., Weiss,R., Herrick,G., Landweber,L.F. and Doak,T.G. (2004) Sequence features of *Oxytricha trifallax* (class Spirotrichia) macronuclear telomeric and subtelomeric sequences. *Protist*, **155**, 311–322.
13. Cavalcanti,A.R.O., Stover,N.A., Orecchia,L., Doak,T.G. and Landweber,L.F. (2004) Coding properties of *Oxytricha trifallax* (*Sterkiella histriomuscorum*) macronuclear chromosomes: analysis of a pilot genome project. *Chromosoma*, **113**, 69–76.