ELSEVIER

Molecular and Biochemical Parasitology 97 (1998) 221–224

MOLECULAR AND BIOCHEMICAL PARASITOLOGY

Short communication

Genetic nomenclature for *Trypanosoma* and *Leishmania*

Christine Clayton [a,*], Mark Adams [b], Renata Almeida [c], Théo Baltz [d], Mike Barrett [e], Patrick Bastien [f], Sabina Belli [g], Stephen Beverley [h], Nicolas Biteau [d], Jenefer Blackwell [c], Christine Blaineau [f], Michael Boshart [i], Frederic Bringaud [d], George Cross [j], Angela Cruz [k], Wim Degrave [l], John Donelson [m], Najib El-Sayed [b], Gioliang Fu [c], Klaus Ersfeld [n], Wendy Gibson [o], Keith Gull [n], Alasdair Ivens [p], John Kelly [q], Daniel Lawson [r], John Lebowitz [s], Phelix Majiwa [t], Keith Matthews [n], Sara Melville [u], Gilles Merlin [d], Paul Michels [v], Peter Myler [w], Alan Norrish [c], Fred Opperdoes [v], Barbara Papadopoulou [x], Marilyn Parsons [w], Thomas Seebeck [y], Deborah Smith [p], Kenneth Stuart [w], Michael Turner [e], Elisabetta Ullu [z], Luc Vanhamme [aa]

[a] *Zentrum für Molekulare Biologie, Im Neuenheimer Feld 282, D-69120 Heidelberg, Germany*
[b] *TIGR, Rockville, Maryland, USA*
[c] *Addenbrookes Hospital, Cambridge, UK*
[d] *Université Bordeaux II, Bordeaux, France*
[e] *University of Glasgow, Glasgow, UK*
[f] *CNRS EP 0613, Montpelier, France*
[g] *Université de Lausanne, Epalinges, Switzerland*
[h] *Washington University, St. Louis, Missouri, USA*
[i] *Max-Planck-Institut für Biochemie, Munchen, Germany*
[j] *The Rockefeller University, New York, New York, USA*
[k] *University of Sao Paolo, Ribeirao Preto, Brazil*
[l] *FIOCRUZ, Rio de Janiero, Brazil*
[m] *University of Iowa, Iowa City, USA*
[n] *University of Manchester, Manchester, UK*
[o] *University of Bristol, Bristol, UK*
[p] *Imperial College, London, UK*
[q] *London School of Hygiene and Tropical Medicine, London, UK*
[r] *Sanger Centre, Cambridge, UK*
[s] *Purdue University, West Lafayette, Indiana, USA*
[t] *ILRI, Nairobi, Kenya*
[u] *University of Cambridge, Cambridge, UK*
[v] *Christian de Duve Institute of Cellular Pathology, Brussels, Belgium*
[w] *Seattle Biomedical Research Institute, Seattle, Washington, USA*
[x] *Centre de Recherche de CHUL, Québec, Canada*
[y] *University of Bern, Bern, Switzerland*
[z] *Yale University, New Haven, Connecticut, USA*
[aa] *Université Libre de Bruxelles, Brussels, Belgium*

Received 11 May 1998; received in revised form 22 July 1998; accepted 31 July 1998

* Corresponding author. Tel.: +49 6221 546876; fax: +49 6221 545894; e-mail: cclayton@sun0.urz.uni-heidelberg.de

PII S0166-6851(98)00115-7

Keywords: Nomenclature; Trypanosoma; Leishmania

1. Introduction

The increasing availability of kinetoplastid gene sequences and mutants, combined with the wide use of genetic manipulation to create progressively more complex strains, has made the development of a unified genetic nomenclature imperative. We suggest here the use of nomenclature adapted from that accepted for the yeast *Saccharomyces cerevisiae* [1]. Yeast has been chosen as a basis for three main reasons: First, appropriate nomenclature for nearly all operations performed in kinetoplastids is already in place. Second, it is the only eucaryote for which the complete genomic sequence is available, and the function of these genes is being determined by a large number of laboratories. Yeast is therefore likely to be a major source of sequence information indicating the possible function of new kinetoplastid genes. Third, the complementation of *S. cerevisiae* mutants is being increasingly used as a way to confirm kinetoplastid gene function.

The most important aspects of the proposed nomenclature system are listed below.

2. Basic priciples

1. Wild-type genes are in italicised upper case, 3–6 letters e.g. *DHFRTS* (see below for more details) [2]. A gene is assumed wild-type unless there is evidence to the contrary. The reference wild-type sequences will ultimately be those of the genome-project strains.
2. If several functional alleles must be distinguished, these are designated by hyphenated numbers, e.g. *DHFRTS-1*, *DHFRTS-2*, etc. It follows that the basic gene name should preferably not include hyphens.
3. Multiple related but non-identical genes can be named with added letters or numbers, without hyphenation, e.g. *PGKA*, *PGKB*, *PGKC*, *THT1*, *THT2* [3,4].
4. The nature of the source organism can, if necessary within a discussion comparing species or sub-species, be indicated by a prefix. For example, *LmjDHFRTS* and *LmxDHFRTS* could distinguish the *L. major* and *L. mexicana* *DHFRTS* genes.
5. RNA products have the same (italicised) designations as the corresponding genes e.g. *DHFRTS* RNA.
6. Protein products have the same designations as the corresponding genes but are not italicised, e.g. PGKA, *Lmx*DHFRTS. (The species prefix, when it has to be used, would be italicised as usual.)
7. Promoters are denoted by subscript, e.g. P_{RRNA} for ribosomal RNA promoter.

3. Designation of mutations

1. Mutant versions of genes, RNAs and proteins are shown in lower case, e.g. the *dhfrts-3* mutant gene would encode the dhfrts-3 protein. (Note that this is true even when the mutation is dominant for a particular phenotype.)
2. Deletions can be designated by Δ, e.g.Δ*dhfrts*. This symbol implies that most or all of the gene, and certainly all functions of the encoded RNA or protein, are gone. Small deletions can be designated as mutants (point 2) and described separately.
3. Diploidy is indicated using /. For example, a strain with a deletion of one *DHFRTS* gene would be *Δdhfrts/DHFRTS*; a strain with one mutant copy and one wild-type copy could be *dhfrts-3/DHFRTS*. A triploid might be *Δdhfrts/Δdhfrts/DHFRTS*.

4. Genetic manipulation

1. Gene replacements are indicated by a double colon (the standard indication of a double-

strand break and reunion). For example, a strain in which one copy of the entire *DHFRTS* coding region had been replaced by the neomycin phosphotransferase gene *NEO*, and the other by the hygromycin resistance gene *HYG*, would be *Δdhfrts::NEO/Δdhfrts::HYG*.

2. Insertional inactivations can be denoted by ^. For example, placing *NEO* into the middle of the *DHFR* gene would give *^dhfrts::NEO*.
3. Fusion proteins-in $5' \rightarrow 3'$ order, for example *DHFRTS::GFP* is a fusion of GFP (at the C-terminus) with DHFR (at the N-terminus).
4. Extrachromosomal elements and episomes (i.e. anything with non-mendelian inheritance) are in square parentheses, e.g. a strain with *[pX HYG GFP]* would carry a pX-based plasmid [5] expressing hygromycin resistance and the green fluorescent protein. The presence of a high copy number could, if important, be indicated by a subscript e.g. $[pX\ HYG\ GFP]_{85}$.
5. Other characteristics can be added as superscripts, e.g. GFP^{Ti} for a tetracycline-inducible GFP gene.

As an example, we could take a wholly fictitious *L. major* strain with the following genotype: $TETR\ BLE\ TbPEX11^{Ti}\ HYG\Delta pex11::NEO/\Delta pex11::NEO\ [pX\ GFP\ PAC]_{20}$

This would carry

1. the procaryotic tetracycline repressor *TETR* [6];
2. *BLE* (phleomycin/bleomycin resistance marker);
3. a tetracycline-inducible copy of the *T. brucei* homologue of the yeast *PEX11* gene, *TbPEX11*;
4. *HYG*;
5. a complete homozygous knockout of the endogenous *PEX11* locus made using the *NEO* marker;
6. copies of a pX-based episome encoding green fluorescent protein and the *PAC* (puromycin resistance) marker.

5. Unresolved issues

Insertions between existing genes that do not cause a deletion. If it is imperative to show the site of integration of a gene, it might be indicated by leaving the targeted site in upper case e.g. for an integration of *NEO* in the tubulin locus, *TUB::NEO*. This is probably best avoided if possible as it could cause confusion!

Linkage of two integrated genes (for example, insertion of a plasmid containing *GFP* and *HYG* into the tubulin locus). Indistiguishable copies within tandem arrays pose a problem, especially as copy numbers may vary between strains and even between chromosome homologues. If numbering becomes essential a solution may be to use the cross-hatch symbol #, e.g. *SLRNA # 3* for the third gene in the array.

Nomenclature to distinguish the multiple VSG expression sites and the genes they contain is outside the scope of this communication.

6. Choosing names for genes

Gene names are groups of three to six letters without any interruptions, if possible an abbreviation of the name of encoded protein or RNA. Ultimately it will be necessary to generate and maintain standardised lists of known genes and their names in salivarian trypanosomes and *Leishmania* as part of the central parasite genome database. In the meantime, if a homologous gene has already been characterised in another trypanosomatid, the same abbreviation should be adopted if possible. Also, if the function of the gene product has been demonstrated, and if there is a functional homologue in *Saccharomyces*, the use of the *Saccharomyces* abbreviation is preferred unless this has already been used for a different trypanosomatid gene. In many cases judgements on whether or not a gene can be accepted as a homologue will be made independently by the referees of corresponding manuscripts. The majority of published kinetoplastid gene and protein names already either conform with this nomenclature system or can adapted with minimal alteration. Sometimes re-naming will be necessary, because homologues or functions have been identified, or to conform with new standardised nomenclature systems. It is hoped that any conflicts which arise between laboratories can be re-

solved privately by the researchers involved. An example of a gene registration system, as well as lists of all currently-known *S. cerevisae genes*, can be accessed at the yeast genome database: http://genome-www.stanford.edu/Saccharomyces/. Another possible scource of help is [7]. The parasite genome web site is accessible via the Worldwide Web at http://www.ebi.ac.uk/parasites/parasite-genome.html.

7. Conclusion and acknowledgements

This system was discussed at a workshop at the Woods Hole Molecular Parasitology meeting (Massachusetts, USA) in September 1996 and again at a WHO-sponsored workshop for the *T. brucei* and *Leishmania* genome projects (Arcachon, France) in April 1998. We hope it will be useful and enjoy broad acceptance.

References

[1] Sherman F. Genetic Nomenclature. In: Strathern JN, Jones EW, Broach JR, editors. The molecular Biology of the Yeast Saccharomyces: Life cycle and Inheritance. Cold Spring Harbor: Cold Spring Harbor Press, 1981.

[2] Beverley SM, Ellenberger TE, Cordingley JS. Primary structure of the gene encoding the bifunctional dihydrofolate reductase-thymidylate synthase of *Leishmania major*. 1986;83:2584–2588.

[3] Gibson WC, Swinkels BW, Borst P. Post-transcriptional control of the differential expression of phosphoglycerate kinase genes in *Trypanosoma brucei*. J Mol Biol 1988;201:315–25.

[4] Bringaud F, Baltz T. Differential regulation of two distinct families of glucose transporter genes in *Trypanosoma brucei*. Mol Cell Biol 1993;13:1146–54.

[5] Lebowitz JH, Coburn CM, McMahon-Pratt D, Beverley S. Development of a stable *Leishmania* expression vector and application to the study of parasite surface antigen genes. Proc Natl Acad Sci USA 1990;87:9736–40.

[6] Wirtz LE, Clayton CE. Inducible gene expression in trypanosomes mediated by a procaryotic repressor. Science 1995;268:1179–83.

[7] Scientific Style and Format, Chapter 20, Council of Biology (Eds.) Cambridge, Cambridge University Press, UK, 1994.