Published in final edited form as:

Trends Parasitol. 2011 May ; 27(5): 183–184. doi:10.1016/j.pt.2011.02.007.

Identity crisis? The need for systematic gene IDs

Marilyn Parsons^{1,2}, Peter J. Myler^{1,2,3}, Matthew Berriman⁴, David S. Roos⁵, and Kenneth D. Stuart^{1,2}

¹Seattle Biomedical Research Institute, 307 Westlake Avenue North, Seattle, WA 98109, USA

²Department of Global Health and University of Washington, Seattle, WA 98195, USA

³Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195, USA

⁴The Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK

⁵Department of Biology, University of Pennsylvania, Philadelphia PA 19104, USA

Abstract

Recent years have seen an explosion in the availability of protozoan pathogen genome sequences. Although data regarding the underlying genome sequence remain relatively stable after the initial draft, understanding of specific gene function is increasing rapidly. This dichotomy is reflected in the relative stability of systematic gene identifiers (SysIDs^{*}) in genome sequence databases, as compared to evolving and/or conflicting gene and gene product names. GenBank/EMBL/DDBJ accession numbers are important, but most protozoan parasite researchers use organism-based databases such as EuPathDB or GeneDB as their immediate resource for gene-based information because they not only provide sequence information but also functional information and links to references. Reference to SysIDs therefore provides a valuable bridge to this repository of information.

SysIDs^{*} are found in differently named fields in each database: EuPathDB (GeneID), GeneDB (Systematic Name or Gene), TDRTargets (Name or Identifier/Accession) and UniProt (Accession Number). They are typically provided in GenBank Entrez Gene in the 'locus tag' field or occasionally in the 'gene symbol' field. Abbreviations and databases: DDBJ, DNA Data Bank of Japan (<http://www.ddbj.nig.ac.jp/>); EMBL, European Molecular Biology Laboratory Nucleotide Sequence Database (<http://www.ebi.ac.uk/embl/>); EuPathDB, Eukaryotic Pathogen Database Resources (<http://eupathdb.org/eupathdb/>); GenBank; Genetic Sequence Database of the National Institutes of Health (<http://www.ncbi.nlm.nih.gov/genbank/>); GeneDB, Sanger Institute Pathogen Genomics Data

^{*}SysIDs: SysIDs are found in differently named fields in each database: EuPathDB (GeneID), GeneDB (Systematic Name or Gene), TDRTargets (Name or Identifier/ Accession) and UniPro (Accession Number). They are provided in GenBank Entrez Gene in the 'ocus tag' or 'gene symbol' field.

© 2011 Elsevier Ltd. All rights reserved.

Corresponding author: Parsons, M. (marilyn.parsons@seattlebiomed.org).

^{*}SysIDs: SysIDs are found in differently named fields in each database: EuPathDB (GeneID), GeneDB (Systematic Name or Gene), TDRTargets (Name or Identifier/ Accession) and UniPro (Accession Number). They are provided in GenBank Entrez Gene in the 'ocus tag' or 'gene symbol' field.

Resource (<http://www.genedb.org/Homepage>); TDRTargets, Tropical Disease Research Chemogenomics Resource (<http://tdrtargets.org/>); UniProt, Universal Protein Resource (<http://www.uniprot.org>).

Gene names change with research advances

Our knowledge about the function(s) of individual gene products is dynamic and cumulative, enabling progressive improvement of the annotation of the parasite genomes. An unanticipated byproduct of this improvement is the difficulty in tracking gene function in the literature (Box 1 for examples). Historically, many gene products were named according to their apparent molecular mass, leading to a plethora of names such as p34, MP52, or p67. Because orthologs in different species often have different molecular masses, and unrelated proteins can have similar molecular masses, these uninformative historical names have tended to be replaced by more descriptive names as research has revealed specific gene functions. However, different groups working on the same genes often assigned different names, as was done for RNA-editing proteins. In addition, different names can be used for orthologs across species, and the same name can be used for genes with very different functions (Box 1). The ensuing ambiguity leads to confusion; worse, important information can be overlooked, leading to futile studies. Tracking these myriad names requires investigators to trace back through years of literature to identify the sequence or accession number that allows one to determine definitively which gene product is being discussed.

Box 1

Examples of nomenclature challenges in the literature

Different names for the same gene product within a species

Trypanosoma brucei Tb09.160.2970 encodes a mitochondrial RNA-editing ligase 1 which has been variously named KREL1, REL1, TbMP52, LC-7a and band IV [3].

Plasmodium falciparum PF13_0201 encodes a protein that has been published as TRAP (thrombospondin-related anonymous protein) [4] and SSP2 (sporozoite surface protein 2) [5].

The same names for gene products with different functions

T. brucei Tb11.02.4910 and Tb11.02.4930 gene products are both published as VSP1 (vacuolar soluble pyrophosphatase), but have different stage-specific expression profiles [6].

T. brucei Tb11.01.7730 and Tb11.01.5570 gene products have both been named p34, but the former is a TFIIF basal transcription factor subunit [7] and the latter is an RNA binding protein, which has also been named NRBD1 [8,9].

Different names for orthologs in related species

Leishmania major LmjF.30.0370 and *T. brucei* Tb927.6.1780 encode orthologous protein kinases, respectively named *MPK12* and *MAPK4* [10].

Plasmodium yoelii PY00454 and *Plasmodium berghei* PBAN-KA_100630 gene products have been published as PPLP1 (perforin-like protein 1) [11] and SPECT2 (sporozoite micronemal protein essential for cell traversal) [12], respectively.

SysIDs provide a solution

With the completion and annotation of the genome sequences for many protozoan pathogens, unique SysIDs have been used to unequivocally identify each predicted gene, even when its nucleotide coordinates change within a chromosome or contig sequence. Thus, the SysIDs are stable and consistent between the major genome databases used by protozoan parasite researchers (such as EuPathDB, GeneDB, TDRTargets and UniProt, each of which contains multiple species-specific databases)*. Although SysIDs occasionally change as genome assemblies improve, a list of previous SysIDs remains associated with each gene (as aliases) to enable record retrieval based on current or past identifiers. Therefore, to avoid the gene identity confusion rampant in the current literature we propose that researchers identify all genes and gene products by their ‘official’ SysIDs in every manuscript submitted for publication and in all datasets submitted to public databases. Not only does inclusion of SysIDs facilitate the work of researchers, but this also allows database curators to unambiguously and rapidly extract gene-specific functional information, thereby making best use of limited curation resources. The format or placement of the listing could be according to the preference of the journal concerned, but at first mention in the text and/or in a table within the Methods section would be convenient for readers. Because abstracts are frequently used for text mining, listing SysIDs within an abstract should also enhance the ability of genome annotators to extract relevant information for curation purposes [1]. We further propose that journals require the listing of the SysID when available. Should a SysID not be available for the gene of interest, it would be helpful to indicate the SysID of the ortholog in a closely related species.

Finally, we suggest that all researchers visit the relevant genome database(s) and ensure that all the synonyms for the genes that they study are listed, as well as any explanatory information about naming. A convenient mode to provide such information is to use the ‘User Comment’ option, which is available on all EuPathDB gene pages. These comments post immediately and are searchable in routine text queries. The comments are reviewed by the genome annotators/curators and usually result in a change in gene name, alias, and/or product description at the ‘official’ source genome database (such as GeneDB), which is then passed back to EuPathDB at the next synchronization. These simple steps will streamline the ability to correlate genes with function, save time and effort, and enhance the genome databases used by all in the field. Although not as poetic or descriptive as the names we might conceive for the genes we pursue, SysIDs abide [2].

Acknowledgments

The authors acknowledge the support from the National Institute of Allergy and Infectious Diseases grant 5U01AI075641 and Bill and Melinda Gates Foundation Grant #50097. The authors are solely responsible for the content.

References

1. Gerstein M, et al. Structured digital abstract makes text mining easy. *Nature*. 2007; 447:142. [PubMed: 17495904]
2. Coen, J.; Coen, E. Working Title Films. 1998. The Big Lebowski.
3. Stuart KD, et al. Complex management: RNA editing in trypanosomes. *Trends Biochem Sci*. 2005; 30:97–105. [PubMed: 15691655]
4. Robson KJ, et al. A highly conserved amino-acid sequence in thrombospondin, properdin and in proteins from sporozoites and blood stages of a human malaria parasite. *Nature*. 1988; 335:79–82. [PubMed: 3045563]
5. Rogers WO, et al. Characterization of *Plasmodium falciparum* sporozoite surface protein 2. *Proc Natl Acad Sci USA*. 1992; 89:9176–9180. [PubMed: 1409621]
6. Lemercier G, et al. A pyrophosphatase regulating polyphosphate metabolism in acidocalcisomes is essential for *Trypanosoma brucei* virulence in mice. *J Biol Chem*. 2004; 279:3420–3425. [PubMed: 14615483]
7. Lee JH, et al. Transcriptionally active TFIIF of the early-diverged eukaryote *Trypanosoma brucei* harbors two novel core subunits but not a cyclin-activating kinase complex. *Nucleic Acids Res*. 2009; 37:3811–3820. [PubMed: 19386623]
8. Zhang J, et al. Developmental regulation of two nuclear RNA binding proteins, p34 and p37, from *Trypanosoma brucei*. *Mol Biochem Parasitol*. 1998; 92:79–88. [PubMed: 9574912]
9. De Gaudenzi J, et al. RNA-binding domain proteins in Kinetoplastids: a comparative analysis. *Eukaryot Cell*. 2005; 4:2106–2114. [PubMed: 16339728]
10. Wiese M. Leishmania MAP kinases – Familiar proteins in an unusual context. *Int J Parasitol*. 2007; 37:1053–1062. [PubMed: 17548090]
11. Kaiser K, et al. A member of a conserved *Plasmodium* protein family with membrane-attack complex/perforin (MACPF)-like domains localizes to the micronemes of sporozoites. *Mol Biochem Parasitol*. 2004; 133:15–26. [PubMed: 14668008]
12. Ishino T, et al. A *Plasmodium* sporozoite protein with a membrane attack complex domain is required for breaching the liver sinusoidal cell layer prior to hepatocyte infection. *Cell Microbiol*. 2005; 7:199–208. [PubMed: 15659064]