Letter

Orthology, paralogy and proposed classification for paralog subtypes

The conceptual underpinning of the terms 'orthology' and 'paralogy' has been the subject of several recent publications [1–4]. The renewed interest in these descriptors of the evolutionary relationships among genes is not surprising given the need for unambiguous definitions in the fast-growing field of comparative and evolutionary genomics and the widespread confusion about the exact meanings of some key terms, (e.g. [5–7]). Many researchers seem to believe that orthologs are simply genes (proteins) with the same function in different organisms, whereas paralogs are simply homologs within one organism. This does not agree with the original definitions of orthology and paralogy given by [8] (see also [9] for an overview) and could easily lead to confusion. We therefore find it important to clarify these terms in some detail, and also wish to further reduce ambiguity by introducing two new terms for subtypes of paralog.

The original definition of orthologs is two genes from two different species that derive from a single gene in the last common ancestor of the species (e.g. HB and WB in Fig. 1). Paralogs are defined as genes that derive from a single gene that was duplicated within a genome. The latter definition does not specify that paralogs can only be found in a single organism, and hence genes in different organisms that arose from gene duplication in an ancestral genome are also paralogs according to the definition.

Several other aspects of orthologous and paralogous relationships between genes have emerged as important in evolutionary genomics. Figure 1 illustrates how multiple genes can simultaneously be orthologs of another gene, in this case HA* can be said to be 'co-orthologs' of WA* (where HA* indicates all genes whose name starts with HA, *etc.*) Co-orthologs are thus paralogs produced by duplications of orthologs subsequent to a given speciation event (also called lineage-specific expansions of paralogous families), which is commonly observed between distantly related species [10–12]. This special type of paralog needs a qualifier to distinguish it from paralogs that resulted from an ancestral (relative to the given speciation event) duplication and, consequently, are not (co)orthologous to a given gene in the second species (e.g. HA* and WB in Fig. 1).

We here suggest two terms that are derived by analogy to terms used in phylogenetics, 'outgroup' and 'ingroup', which denote anciently and recently branching lineages, respectively. Relative

http://tig.trends.com 0168-9525/02/$ – see front matter PII: S0168-9525(02)02793-2

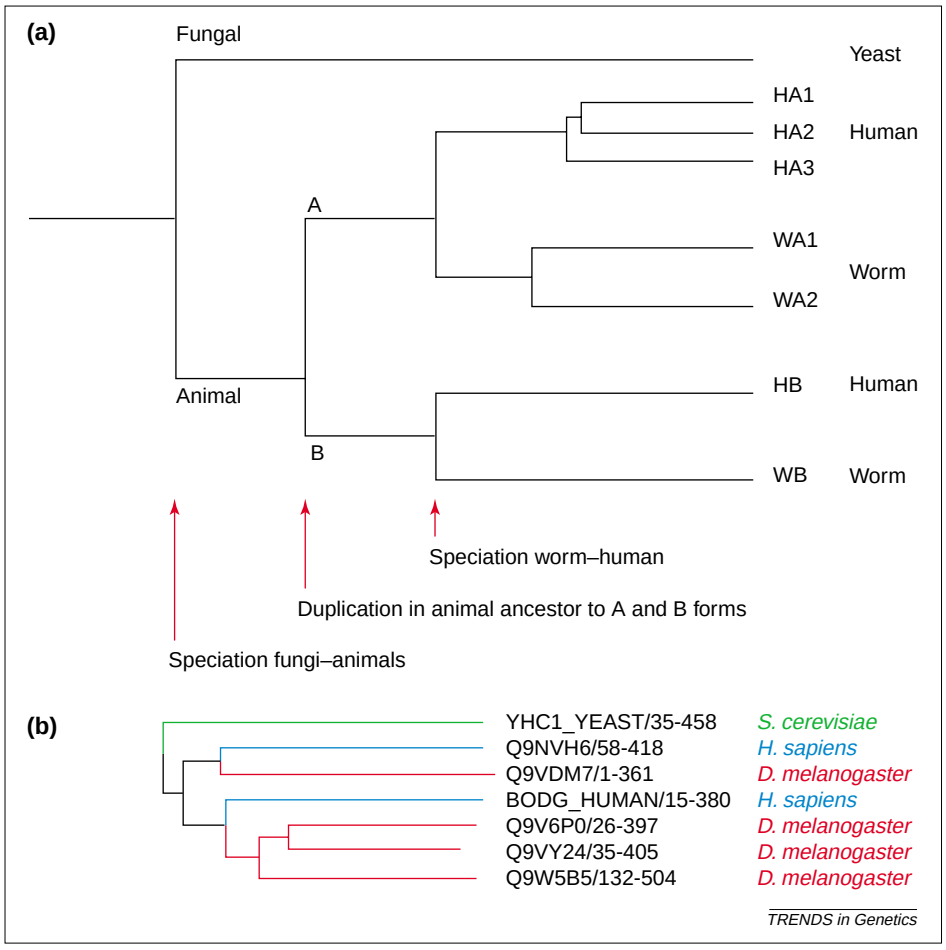

Fig. 1. The definition of inparalogs and outparalogs. (a) Consider an ancient gene inherited in the yeast, worm and human lineages. The gene was duplicated early in the animal lineage, before the human–worm split, into genes A and B. After the human–worm split, the A form was in turn duplicated independently in the human and worm lineages. In this scenario, the yeast gene is orthologous to all worm and human genes, which are all co-orthologous to the yeast gene. When comparing the human and worm genes, all genes in the HA* set are co-orthologous to all genes in the WA* set. The genes HA* are hence 'inparalogs' to each other when comparing human to worm. By contrast, the genes HB and HA* are 'outparalogs' when comparing human with worm..However, HB and HA*, and WB and WA* are inparalogs when comparing with yeast, because the animal–yeast split pre-dates the HA*–HB duplication. (b) Real-life example of inparalogs: γ-butyrobetaine hydroxylases. The points of speciation and duplication are easily identifiable. The alignment is a subset of Pfam:PF03322 and the tree was generated by neighbor-joining in Belvu. All nodes have a bootstrap support exceeding 95%.

to a given speciation event, paralogs derive either from an ancestral duplication and do not form orthologous relationships, or they derive from a lineage-specific duplication, giving rise to co-orthologous relationships. The logical terms therefore seem to be, respectively, 'outparalog' and 'inparalog', explicitly denoting that they are subtypes of paralogs and when they branched relative to the given speciation event. We would also consider more classical terms, such as 'alloparalog' for outparalog and 'symparalog' for inparalog (by analogy to allopatric and sympatric speciation), but will not use them further here for the sake of consistency.

Therefore, our definition of 'inparalogs' is: paralogs in a given lineage that all evolved by gene duplications that happened *after* the radiation (speciation) event that separated the given lineage from the other lineage under consideration. Our definition of 'outparalogs' is: paralogs in the given lineage that evolved by gene duplications that happened *before* the radiation (speciation) event.

With more and more complete genome sequences becoming available, the genomics community is becoming aware that 'homology' is not a sufficiently well-defined term to describe the evolutionary relationships between genes. Emphasis is instead shifting towards identifying orthologs, which are evolutionary and, typically, functional counterparts in different species. Conversely, analysis of paralogs, particularly inparalogs, is important for detecting lineage-specific adaptations. This is particularly relevant for identifying functions of human genes by studying orthologs in model organisms. A real-life example of in- and outparalogs between human and fly γ-butyrobetaine hydroxylases is shown in Fig. 1b.

We hope that adopting the terms inparalog and outparalog leads to an increase in clarity in genomic and evolutionary publications and help avoid misleading statements on evolutionary relationships between genes.

Acknowledgements
We thank Walter Fitch, Roy Jensen and Lennart Philipson for helpful discussions.

Erik L.L. Sonnhammer*
Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177 Stockholm, Sweden.
*e-mail: Erik.Sonnhammer@cgb.ki.se

Eugene V. Koonin
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA.

References

1 Petsko, G.A. (2001) Homologuephobia. *Genome Biol.* 2, comment1002.1-1002.2
2 Jensen, R.A. (2001) Orthologs and paralogs – we need to get it right. *Genome Biol.* 2, interactions1002
3 Koonin, E.V. (2001) An apology for orthologs – or brave new memes. *Genome Biol.* 2, comment1005.1–1005.2
4 Theissen, G. (2002) Secret life of genes. *Nature* 415, 741
5 Gerlt, J.A. and Babbitt, P.C. (2001) Can sequence determine function? *Genome Biol.* 1, reviews0005.1–0005.10
6 Fabrega, C. *et al.* (2001) An aminoacyl tRNA synthetase whose sequence fits into neither of the two known classes. *Nature* 411, 110–114
7 Xie, T. and Ding, D. (2000) Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale. *Gene* 261, 305–310
8 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113
9 Fitch, W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.* 16, 227–231
10 Jordan, I.K. *et al.* (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* 11, 555–565
11 Lespinet, O. *et al.* (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12, 1048–1059
12 Remm, M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052

Published online: 30 October 2002

