References

1 Kimbrell, D.A. and Beutler, B. (2001) The evolution and genetics of innate immunity. *Nat. Rev. Genet.* 2, 256–267
2 Otvos, L. Jr (2000) Antibacterial peptides isolated from insects. *J. Pept. Sci.* 6, 497–511
3 Cociancich, S. *et al.* (1993) Insect defensin, an inducible antibacterial peptide, forms voltage-dependent channels in *Micrococcus luteus*. *J. Biol. Chem.* 268, 19239–19245
4 Maget-Dana, R. and Ptak, M. (1997) Penetration of the insect defensin A into phospholipid monolayers and formation of defensin A-lipid complexes. *Biophys. J.* 73, 2527–2533
5 Matsuyama, K. and Natori, S. (1990) Mode of action of sapecin, a novel antibacterial protein of *Sarcophaga peregrina* (flesh fly). *J. Biochem. (Tokyo)* 108, 128–132
6 Risso, A. (2000) Leukocyte antimicrobial peptides: multifunctional effector molecules of innate immunity. *J. Leukoc. Biol.* 68, 785–792
7 Hanzawa, H. *et al.* (1990) 1H nuclear magnetic resonance study of the solution conformation of an antibacterial protein, sapecin. *FEBS Lett.* 269, 413–420
8 Bonmatin, J.M. *et al.* (1992) Two-dimensional 1H NMR study of recombinant insect defensin A in water: resonance assignments, secondary structure and global folding. *J. Biomol. NMR* 2, 235–256
9 Cornet, B. *et al.* (1995) Refined three-dimensional solution structure of insect defensin A. *Structure* 3, 435–448
10 Yang, Y.S. *et al.* (2000) Solution structure and activity of the synthetic four-disulfide bond Mediterranean mussel defensin (MGD-1). *Biochemistry* 39, 14436–14447
11 Froy, O. and Gurevitz, M. (1998) Membrane potential modulators: a thread of scarlet from plants to humans. *FASEB J.* 12, 1793–1796
12 Mitta, G. *et al.* (2000) Mytilin B and MGD2, two antimicrobial peptides of marine mussels: gene structure and expression analysis. *Dev. Comp. Immunol.* 24, 381–393
13 Nakajima, Y. *et al.* (2002) Antibacterial peptide defensin is involved in midgut immunity of the soft tick, *Ornithodoros moubata*. *Insect Mol. Biol.* 11, 611–618
14 Dimarcq, J.L. *et al.* (1994) Characterization and transcriptional profiles of a *Drosophila* gene encoding an insect defensin. A study in insect immunity. *Eur. J. Biochem.* 221, 201–209
15 Cho, W.L. *et al.* (1997) Molecular characterization of a defensin gene from the mosquito, *Aedes aegypti*. *Insect Biochem. Mol. Biol.* 27, 351–358
16 Eggleston, P. *et al.* (2000) Genomic organization and immune regulation of the defensin gene from the mosquito, *Anopheles gambiae*. *Insect Mol. Biol.* 9, 481–490
17 Munks, R.J. *et al.* (2001) Regulation of midgut defensin production in the blood-sucking insect *Stomoxys calcitrans*. *Insect Mol. Biol.* 10, 561–571
18 Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling – a review. *Gene* 238, 103–114
19 Sudhof, T.C. *et al.* (1985) The LDL receptor gene: a mosaic of exons shared with different proteins. *Science* 228, 815–822
20 Long, M. (2001) Evolution of novel genes. *Curr. Opin. Genet. Dev.* 11, 673–680
21 Patthy, L. (1994) Introns and exons. *Curr. Opin. Struct. Biol.* 4, 383–392

0168-9525/$ - see front matter
doi:10.1016/j.tig.2003.10.010

Letters

Eukaryotic genes in *Mycobacterium tuberculosis*? Possible alternative explanations

Rhoda J. Kinsella and James O. McInerney

Bioinformatics and Pharmacogenomics Laboratory, Biology Department, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

Estimates concerning the incidence of interspecies gene transfer vary considerably. In 2002, *Trends in Genetics* reported an analysis of the pathogenic bacterium *Mycobacterium tuberculosis* indicating that 19 genes that are currently resident in this species were acquired from various eukaryotes. We have re-examined these data using a combination of larger databases and more detailed phylogenetic analyses. We found that for all 19 genes alternative hypotheses must be considered.

There have been many claims of horizontal gene transfer (HGT) events [1,2] that have subsequently been shown to be artifacts of the analysis method used [3–5]. In the absence of more direct evidence, the most convincing method of generating hypotheses of HGT involves the construction and analysis of phylogenetic tree topologies. There are alternative approaches such as those that analyze the results of BLAST [6] database searches. However, these approaches have been shown to occasionally rely on *ad hoc* assumptions concerning the completeness of database coverage and the constancy of the rate of evolution and do not adequately account for gene loss [5].

Corresponding author: James O. McInerney (james.o.mcinerney@may.ie).

In addition, estimates indicate that only a small proportion of the diversity of prokaryotes is known [7], which makes the task of HGT inference more difficult.

Gamieldien and co-workers recently proposed that 19 genes in *Mycobacterium tuberculosis* were acquired from eukaryotes [8]. The authors carried out a series of analyses and, on the basis of either BLAST results that only returned tubercule bacilli and eukaryotic homologues or phylogeny reconstruction, hypotheses of HGT were constructed.

We wished to explore whether or not there were methodological biases that might have caused Gamieldien *et al.* to reach their conclusions. Specifically, we wished to examine issues related to the incompleteness of data sampling, methods for constructing HGT hypotheses and possible alternative explanations of the data.

Sequence analysis

The 19 reported HGT genes were selected and their protein homologues were identified with BLASTP [6], using the default values and setting a cut off E-value of 10^{-7}. The databases that were used for this analysis included Swiss-Prot Release 41.4 (124 464 entries) and

Table 1. Alternative explanations for the presence of proposed eukaryotic genes in *Mycobacterium tuberculosis*[a]

Gene	Product	Alternative explanation
Rv1373	Hydroxysteroid sulfotransferase	Closest relatives are found in γ-proteobacteria
Rv3451	Cutinase	No intermixing of prokaryotic and eukaryotic genes
Rv3452	Cutinase	No intermixing of prokaryotic and eukaryotic genes
Rv1758	Cutinase	No intermixing of prokaryotic and eukaryotic genes
Rv1984c	Cutinase	No intermixing of prokaryotic and eukaryotic genes
Rv0386	Adenylyl cyclase	Form monophyletic group with other bacteria, including α-proteobacteria and actinobacteria
Rv0891c	Adenylyl cyclase	Form monophyletic group with other bacteria, including α-proteobacteria and actinobacteria
Rv1358	Adenylyl cyclase	Form monophyletic group with other bacteria, including α-proteobacteria and actinobacteria
Rv1359	Adenylyl cyclase	Form monophyletic group with other bacteria, including α-proteobacteria and actinobacteria
Rv2488c	Adenylyl cyclase	Form monophyletic group with other bacteria, including α-proteobacteria and actinobacteria
Rv2577	Purple acid-phosphatase	Forms monophyletic group with genes from four other bacteria (gram positive and gram negative)
Rv3548	17-β hydroxyestradiol dehydrogenase	Resides in clade consisting of many other bacterial sequences
Rv0148	17-β hydroxyestradiol dehydrogenase	Groups with a *Caulobacter crescentus* sequence near a clade comprising bacteria and eukaryotes
Rv1106c	Steroid dehydrogenase	Groups with *Mycobacterium leprae* and *Nocardia* genes
Rv2790c	Sterol carrier protein	Groups with another tuberculosis gene outside a clade of eukaryotes
Rv2483c	Lysophosphatidic acid acetyltransferase	Groups with many other prokaryotic sequences
Rv0764c	Cytochrome-P450 lanosterol demethylase	Resides outside eukaryotic clade
Rv2590	Long-chain fatty-acid CoA ligase	Fusion protein with complex history; it is not clear if this is the result of HGT
Rv1834	Epodide hydrolase	Groups with other prokaryotes to the exclusion of the eukaryotes

[a]Abbreviation: HGT, horizontal gene transfer.

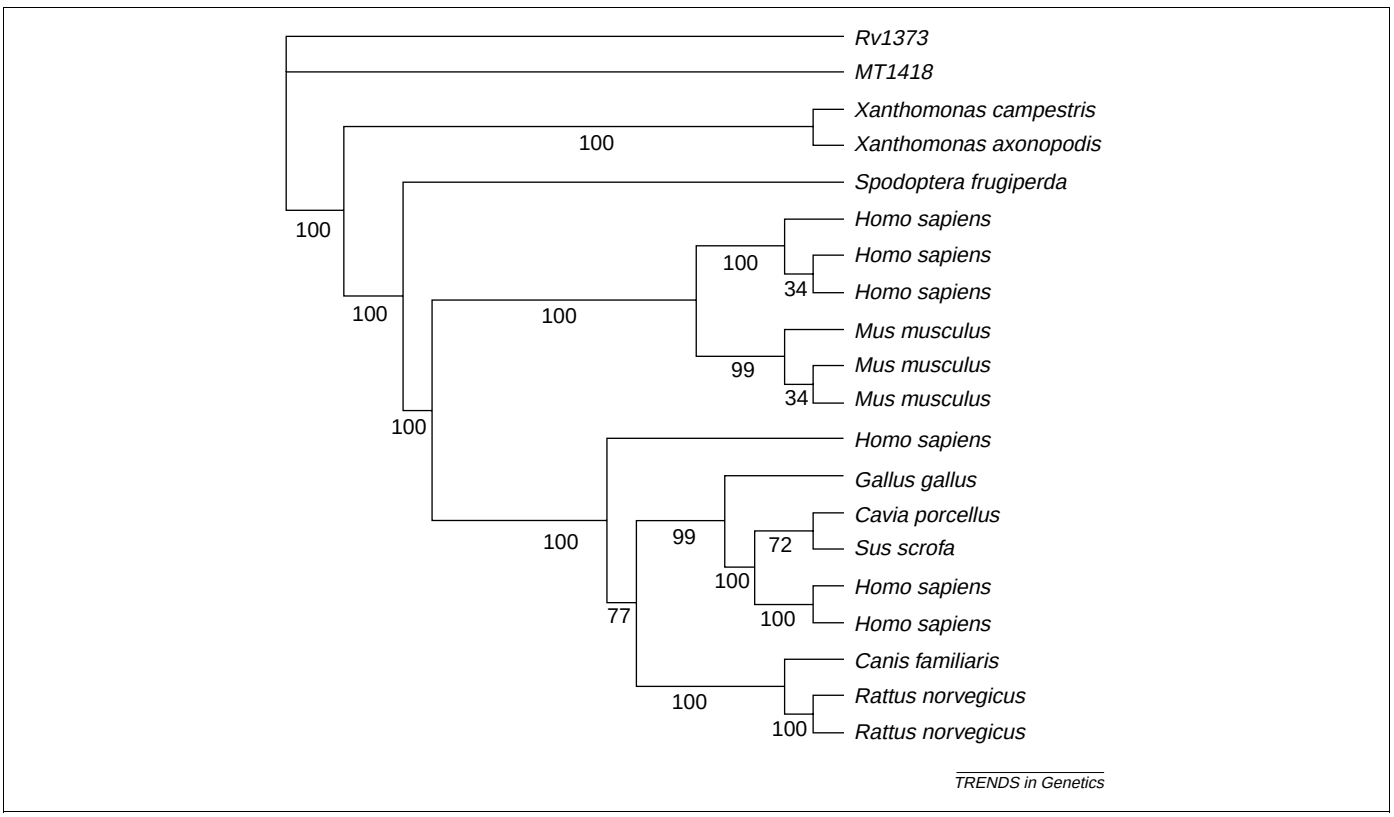

Figure 1. Phylogenetic tree for the hydroxysteroid sulfotransferase gene. The *Mycobacterial* genes are *Rv1373* and *MT1418*. The numbers on the internal branches are clade credibility values (i.e. the degree of confidence) obtained from the majority rule consensus method in Phylogenic Analysis Using Parsimony (PAUP*) [13]. A value of 100 indicates complete confidence.

TrEMBL Release 23.6 (828 210 entries). Homologous sequences were retrieved from the database and aligned using CLUSTALW 1.81 [9], using the default values. These alignments were corrected for obvious ambiguities using SeAl version 2.0.a9 (http://evolve.zoo.ox.ac.uk/software/Se-Al/main.html)†.

Robust hypotheses of phylogenetic relationships were generated using the corrected alignments the Jones amino acid substitution model [10] and the Bayesian framework implemented in the MrBayes 2.01 software (MB) [11]. Following 500 000 generations of Markov Chain Monte Carlo (MCMC) sampling of tree space with trees saves from every hundredth sample, the burn-in phase [12] of 499 trees was discarded and we generated majority rule consensus trees from the remaining trees using Phylogenetic Analysis Using Parsimony (PAUP*) [13]. Clade credibility values were obtained from the majority rule consensus method implemented in PAUP*. We also examined neighbor-joining [14] trees and found that these did not differ significantly.

Alternative hypothesis

All phylogenetic trees and greater detail on the results are available at: http://bioinf.may.ie/comparative/trees/. We find that in no case is the hypothesis of HGT better than an alternative hypothesis (Table 1). Specifically, many of the original hypotheses are the result of incomplete sampling: for example, the *M. tuberculosis* gene family published by Gamieldien *et al.* (Figure 1). In our analysis, there are four available bacterial genes from three different species, both gram positive and gram negative, and there is a single internal branch that separates them from the eukaryotes. Although it is unusual to see such sparseness in the distribution of this gene, it is much too premature to suggest that this is the result of an interdomain HGT event.

Conclusion

We believe that it is not sufficient to propose HGT on the basis of E-values obtained from BLAST hits or on the basis that there are no other bacterial orthologues available. The only hypotheses of HGT that are likely to remain strongly supported are those where a sequence is found in the genome of a species, and phylogenetic analysis of a large number of orthologues of this gene places it within a group of sequences from distantly related species. Further statistical support of the significance of this relationship is also needed.

Several issues confound the identification of HGT events. First, only a tiny portion of all bacterial genomes has been sequenced. Second, protein family evolution is notoriously difficult to interpret and conflicting or unexpected phylogenetic trees do not always imply HGT. Reconciling species and gene phylogenies can often be carried out without having to invoke HGT as an explanation [15]. Paralogy, gene loss, incomplete sampling and our naivety concerning the nature of life on earth all contribute to our lack of understanding of the complicated pattern of gene family evolution. Reconstructing the evolutionary history of gene families is fraught with difficulties and until we have a better grasp of the evolutionary history of life on this planet, hypotheses of HGT must be constructed with extreme caution.

Acknowledgements

This work was supported by the Health Research Board, Higher Education Authority, Program for Research in Third Level Institutes, Enterprise Ireland, IBM and NUI Maynooth. R.J.K. is the recipient of a Daniel O'Connell fellowship.

References

1 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Sogin, M. (1997) History assignment: when was the mitochondrion founded? *Curr. Opin. Genet. Dev.* 7, 792–799
3 Hirt, R.P. *et al.* (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl. Acad. Sci. U. S. A.* 96, 580–585
4 Stanhope, M.J. *et al.* (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 411, 940–944
5 Salzberg, S.L. *et al.* (2001) Microbial genes in the human genome: lateral transfer or gene loss? *Science* 292, 1903–1906
6 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
7 Ward, D.M. *et al.* (1990) 16S rRNA sequences reveal numerous uncultured inhabitants in a natural community. *Nature* 345, 63–65
8 Gamieldien, J. *et al.* (2002) Eukaryotic genes in *Mycobacterium tuberculosis* could have a role in pathogenesis and immunomodulation. *Trends Genet.* 18, 5–8
9 Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680
10 Jones, D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282
11 Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755
12 Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574
13 Swofford, D.L. (2003) PAUP. Phylogenetic analysis using parsimony (and other methods) (Version 4), *Sinauer Associates, Sunderland, Massachusetts*
14 Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425
15 Page, R.D.M. and Cotton, J.C. (2000) GeneTree: a tool for exploring gene family evolution. In *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families* (Sankoff, D. and Nadeau, J., eds), pp. 525–536, Kluwer Academic Publishers

0168-9525/$ - see front matter
doi:10.1016/j.tig.2003.10.001

† These alignments and additional information are available on request from the authors.