OPEN ACCESS Freely available online

Correspondence

Dog as an Outgroup to Human and Mouse

Gerton Lunter

In a recent contribution to *PLoS Computational Biology,* Cannarozzi, Schneider, and Gonnet published evidence that rodents form an outgroup to human and dog [1], in disagreement with several recent studies suggesting that the dog is an outgroup to the primate–rodent clade [2,3]. The authors' arguments rest on a variety of analyses of human, mouse, and dog genes, using opossum to root the phylogeny. Here I argue that despite the large number of characters used in this study, their results may well be erroneous. I then provide new and, I believe, conclusive evidence in favour of the current consensus phylogeny, and I briefly review other recent studies that support this conclusion.

The problem of determining the evolutionary relationship between all extant mammals has a long history. Traditionally, morphological features were used to group "like" mammals together in a tree, purportedly reflecting their phylogeny. More recently, molecular data have generally confirmed these inferences, but have also led to surprising revisions. While sequence analysis is more objective than morphology, it nevertheless emerged that it has its own set of issues, and some phylogenies remain contentious. In [1], Cannarozzi et al. suggested that this contention extends to the phylogeny of human, mouse, and dog, and inferred a phylogeny of these species that disagrees with a recently emerging consensus. Here I challenge their findings, providing new evidence in support of the consensus phylogeny, and suggest that their results may have been biased by long branch attraction (LBA), a known issue in molecular phylogenetic inference.

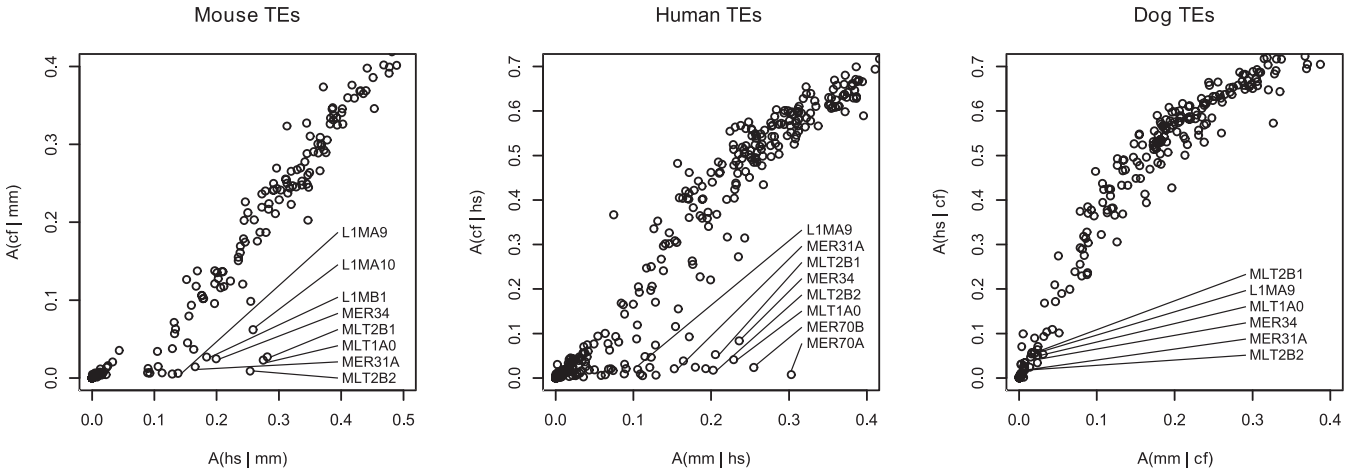
It is well-known that phylogenetic inferences can be biased, and may be inaccurate even with strong bootstrap or posterior support. Felsenstein showed that in parsimony analyses, long branches in the phylogeny tend to attract one another [4]. In contrast to what the authors claim, maximum likelihood methods, although less vulnerable, are similarly affected by LBA [5], particularly when small numbers of taxa are used [6]. This methodological bias has led to various erroneous inferences, such as the now-discredited claim that "the guinea pig is not a rodent" [7,8]. Perhaps counterintuitively, the effect of LBA does not diminish with increasing amounts of sequence data. To quote from a review, "spurious conclusions are often derived from an over-credibility of enormous numbers of nucleotide or amino acid characters (e.g., complete genomes) when combined with poor taxon sampling" [9].

The recently emerging consensus on mammalian phylogeny based on molecular data is surprisingly different from the traditional, morphological phylogeny [2,3]. It proposes four mammalian cohorts, including the Laurasiatheria (of which the dog lineage is part), which separated from the Euarchontoglires about 85–95 million years ago (Mya) [10]. The subsequent speciation separating the Euarchontoglires into Glires (including rodents) and Euarchonta (which includes primates) occurred roughly 80 Mya. The difference is small compared with the total branch length to opossum (180 + 90 My), so that a relatively small bias would suffice to bring about a topology change. As the mouse genome sequence has been evolving fast relative to those of human and dog [11], its branch is expected to be affected by LBA to the opossum branch, which would result in the reported grouping.

These considerations throw some doubt on both the parsimony and maximum likelihood analyses. What about the genome rearrangement argument? After all, genome rearrangements are large-scale but relatively infrequent events, so that the parsimony approximation might be justified. However, the opossum genome had not yet been assembled, and the authors had to resort to chicken, which diverged ~310 Mya from the mammalian lineage, considerably earlier than the opossum did. Moreover, there is strong evidence for hotspots of breakage [12] and breakpoint reuse [13], discounting the "random breakage" model. The use of (nuclear) gene orderings to analyze rearrangements further exacerbates these issues, as it affords little power to resolve breakpoints and artificially increases inhomogeneities in breakage rates, because of large and highly variable intergenic distances. For these reasons, the parsimony approximation may well be invalid, which makes LBA a concern for the genome rearrangement analysis, too.

I thus considered whether the reported tree might be incorrect. To investigate the issue, I used a simple (and, to my knowledge, novel) summary statistic based on the distribution of transposable elements (TEs) in pairwise alignments, which does not require an outgroup genome to root the phylogeny. If a family of TEs is specific to lineage x when compared with y, each occurrence in x is expected to be located opposite a gap in a whole-genome alignment of species x to y. In contrast, if the family is ancestral to x and y, a proportion of TEs will have survived in both species and will align. To quantify the evidence for these alternatives, I defined a statistic $A(y|x)$ (for "ancestralness") as the proportion of nucleotides from a particular TE family in species x that is aligned to a secondary species y. This statistic is near-zero if a family of TEs is specific to x, and non-zero if it is ancestral to the species split. For an outgroup x and a particular family of TEs, the statistics $A(y|x)$ are thus expected to be consistent across ingroup species y (either zero, or non-zero, for all). In contrast, for an ingroup species, some TE families may be ancestral with respect to another ingroup, but lineage-specific when compared with the outgroup. Provided such TE families exist, this would then determine the topology of the phylogeny.

The results (Figure 1 and Table 1) show clear support for the rodent–primate grouping. For example, the MLT2B2 long terminal repeat element is clearly ancestral in the human-to-mouse and mouse-to-human comparisons ($A > 0.20$), but is highly lineage-specific in the other comparisons, each of which include the dog ($A < 0.03$ for all). This pattern can be explained if dog is assumed to be an outgroup to both human and mouse, and that the element has been active primarily between the two speciation events. The same pattern was observed for several other TE families (MLT1A0, MLT2B1, L1MA9, L1MB1, L1MC1, MER31A, MER21B, MER34), while no examples supporting alternative groupings were found. Unlike analyses based on nucleotide characters, TE-based



doi:10.1371/journal.pcbi.0030074.g001

Figure 1. Evidence for the ((Human, Mouse), Dog) Phylogeny

Shown are the ancestralness A(*y*|*x*) for a range of TEs families in a species *x* (mouse, mm; human, hs; dog, cf), compared with the two remaining auxiliary species *y*. Data are shown for all TEs that were present in at least 500 copies covering 50 kb or more in species *x*. When *x* is the outgroup, the ancestralness is expected to be consistent across auxiliary ingroup species *y*, while for ingroup species, some TE families may be ancestral ($A > 0$) for the second ingroup, but lineage-specific ($A \approx 0$) for the outgroup. All three scatter plots support dog as the outgroup species.

studies are not expected to suffer from LBA, because the size of TEs allows for reliable homology assignments (if well-anchored alignments are used), and the marked differences between the TE insertion and small deletion processes means that back mutations are rare. It thus appears that the dog lineage is basal to the primate and rodent lineages.

Numerous recent studies support this conclusion. When many taxa are analyzed simultaneously, the dog consistently appears as an outgroup to human and mouse, when using either nuclear or mitochondrial DNA [2,3,9,14–16]. Studies of rare genomic changes (which are less vulnerable to LBA) consistently support this grouping. For example, by rooting the phylogeny using the consensus sequence of TEs, the evolutionary distance between the speciation events was estimated to be 0.024 substitutions per site [11]. In another study, two of the TE families found here, MLT1A0 and L1MA9, were identified as clear examples supporting the rodent–primate grouping [17], and a recent analysis of several single TE insertions provides additional support [18], as does a method that uses multiple alignments of TEs to infer phylogenies in very similar ways to ours [19]. Rare indels at homologous positions in otherwise well-conserved protein-coding genes also support this phylogeny [20]. Finally, a large cluster of PRAME genes that is absent in chicken and dog, but present in homologous locations in human and mouse, again support the same grouping [21].

Taken together with the possible influence of LBA on the analysis of Cannarozzi et al. [1], it appears unjustified to continue to consider the phylogeny of primates, rodents, and canines as contentious. ■

Table 1. Ancestralness of TE Families in the Six Pairwise Comparisons between Human, Dog, and Mouse

TE Family	A(hs|cf)	A(mm|cf)	A(cf|hs)	A(mm|hs)	A(cf|mm)	A(hs|mm)
MLT1A0	0.05	0.02	0.04	0.23	0.02	0.27
MLT2B2	0.03	0.01	0.02	0.20	0.01	0.25
L1MA9	0.02	0.01	0.02	0.11	0.01	0.14
MER31A	0.03	0.01	0.02	0.15	0.01	0.17
MER21B	0.05	0.03	0.05	0.11	0.04	0.16
MER34	0.05	0.02	0.08	0.24	0.02	0.20
MLT2B1	0.06	0.02	0.05	0.21	0.03	0.28
L1MB1	0.07	0.02	0.07	0.13	0.03	0.18
L1MC1	0.06	0.02	0.09	0.10	0.05	0.15

Listed are all cases where the element was (1) present in all three genomes in at least 500 copies covering at least 50 kb, (2) lineage-specific in at least four comparisons (arbitrarily defined as $A < 0.1$), and (3) ancestral in at least one (defined as $A > 0.1$) All examples point to dog as the outgroup species. Gap assignments were obtained from TBA (Threaded Blockset Aligner) whole-genome alignments [22], and RepeatMasker (unpublished) was used for TE annotations. (Note that the cutoff of 0.1 was chosen for purposes of illustration only; more examples supporting the same conclusion can be identified in Figure 1.)

A, ancestralness; cf, dog; hs, human; mm, mouse.

doi:10.1371/journal.pcbi.0030074.t001

Gerton Lunter (gerton.lunter@dpag.ox.ac.uk)
University of Oxford,
Oxford, United Kingdom

Acknowledgments

The author thanks Martin Goodson, Leo Goodstadt, Chris Ponting, and Caleb Webber for discussions, and one anonymous referee for providing background information.

References

1. Cannarozzi G, Schneider A, Gonnet G (2007) A phylogenomic study of human, dog, and mouse. PLoS Comput Biol 3: e2.
2. Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, et al. (2001) Parallel adaptive radiations in two major clades of placental mammals. Nature 409: 610–614.
3. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, et al. (2001) Molecular phylogenetics and the origins of placental mammals. Nature 409: 614–618.
4. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27: 401–410.
5. Chang JT (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci 134: 189–215.
6. Waddell PJ (1995) Statistical methods of phylogenetic analysis: Including Hadamard conjugations, LogDet transforms, and maximum likelihood. Auckland: Massey University.
7. Graur D, Hide WA, Li WH (1991) Is the guinea-pig a rodent? Nature 351: 649–652.
8. D'Erchia AM, Gissi C, Pesole G, Saccone C, Arnason U (1996) The guinea-pig is not a rodent. Nature 381: 597–600.

9. Bergsten J (2005) A review of long-branch attraction. Cladistics 21: 163–193.
10. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A 100: 1056–1061.
11. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803–819.
12. Webber C, Ponting CP (2005) Hotspots of mutation and breakage in dog and human chromosomes. Genome Res 15: 1787–1797.
13. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, et al. (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309: 613–617.
14. Eizirik E, Murphy WJ, O'Brien SJ (2001) Molecular dating and biogeography of the early placental mammal radiation. J Hered 92: 212–219.
15. Reyes A, Gissi C, Catzeflis F, Nevo E, Pesole G, et al. (2004) Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. Mol Biol Evol 21: 397–403.
16. Kitazoe Y, Kishino H, Okabayashi T, Watabe T, Nakajima N, et al. (2005) Multidimensional vector space representation for convergent evolution and molecular phylogeny. Mol Biol Evol 22: 704–715.
17. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788–793.
18. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, et al. (2006) Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol 4: e91.
19. Bashir A, Ye C, Price AL, Bafna V (2005) Orthologous repeats and mammalian phylogenetic inference. Genome Res 15: 998–1006.
20. Poux C, van Rheede T, Madsen O, de Jong WW (2002) Sequence gaps join mice and men: Phylogenetic evidence from deletions in two proteins. Mol Biol Evol 19: 2035–2037.
21. Birtle Z, Goodstadt L, Ponting C (2005) Duplication and positive selection among hominin-specific PRAME genes. BMC Genomics 6: 120.
22. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708–715.

Citation: Lunter G (2007) Dog as an outgroup to human and mouse. PLoS Comput Biol 3(4): e74. doi:10.1371/journal.pcbi.0030074

Abbreviations: LBA, long branch attraction; Mya, million years ago; TE, transposable element

Dr. Gerton Lunter is with the Department of Physiology, Anatomy, and Genetics, Medical Research Council Functional Genetics Unit, University of Oxford, Oxford, United Kingdom. E-mail: gerton.lunter@dpag.ox.ac.uk

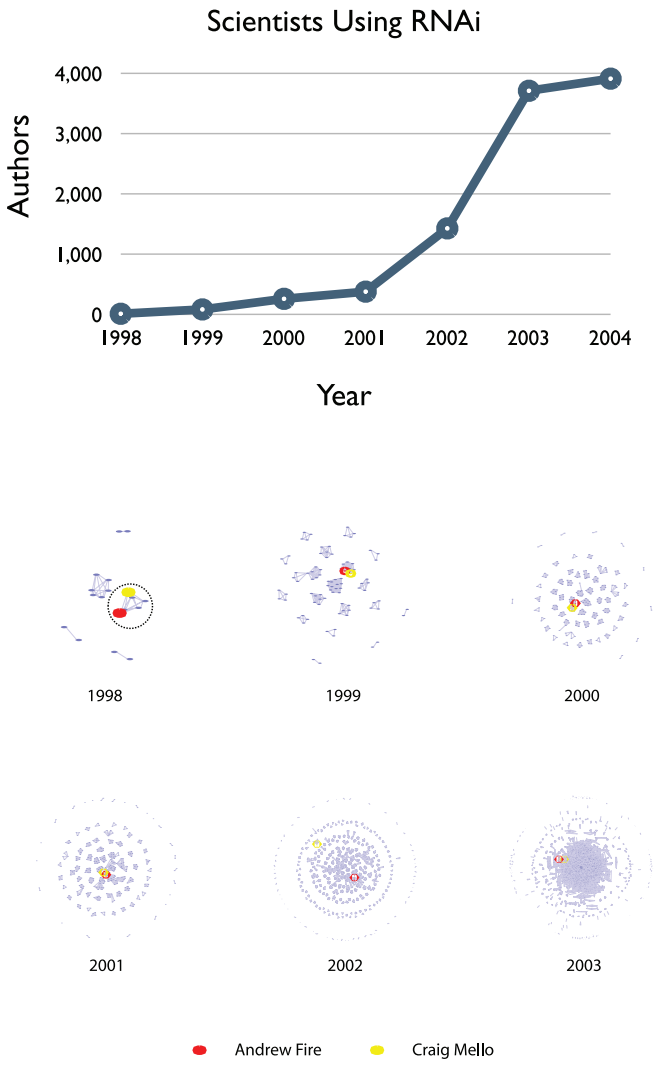
Funding. This research was supported by the Medical Research Council United Kingdom, and in part by the US National Science Foundation under grant PHY99–07949.

Competing interests. The author has declared that no competing interests exist.

RNAi Development

Mark Gerstein, Shawn M. Douglas

The Nobel Prize in Physiology or Medicine in 2006 was awarded to Craig Mello and Andrew Fire for the development of essentially a new field, RNA Interference or RNAi. Because this field sprung from a singular discovery made very recently, we can track its growth in precise detail in the biological and literature databases. In particular, in the figure we show the results for searching PubMed for the term "RNA Interference" in each of the years from 1998 to 2003. (This was done with a tool called PubNet that allows one to visualize the networks generated by arbitrary queries against the National Center for Biotechnology Institute's PubMed database [1].) The top subgraph simply shows that the term



doi: 10.1371/journal.pcbi.0030080.g001

first appeared in 1998, when RNAi was definitively characterized in *Caenorhabditis elegans;* then there was a rapid increase in the number of authors using the term, particularly around 2001. The bottom subgraph shows authors (represented by nodes) who are linked together when they published together in a given year. It dramatically illustrates that in 1998 there were a small number of distinct author clusters; one of these, highlighted by a dotted line in the figure, corresponded to the classic effort of Fire and Mello in *Nature* [2], describing the phenomena of degradation of double-stranded RNA. In subsequent years, one can see that Fire and Mello continued to publish together as a collaborative unit, but many additional groups of investigators appeared, with the number of new groups increasing very rapidly from 1999 to 2001. However, in 2002, Fire and Mello separated and became part of two disconnected publication clusters. Finally, in 2003, one sees a new phenomenon: there were so many authors in the field that they all merged into a huge mega-cluster. Fire and Mello were again connected in the framework of this cluster. That is, there were so many authors in the RNAi field that everyone was linked to everyone else through at least one co-

publication event. In a sense, one is witnessing a "social phase transition": in just five years, a singular discovery spread through the scientific community, progressed through a "tipping point," and became commonplace. ■

Mark Gerstein (Mark.Gerstein@yale.edu)
Program in Computational Biology and Bioinformatics,
Department of Molecular Biophysics and Biochemistry,
and Department of Computer Science,
Yale University,
New Haven, Connecticut, United States of America

Shawn M. Douglas
Genetics Department,
Harvard Medical School,
Boston, Massachusetts, United States of America

References

1. Douglas SM, Montelione GT, Gerstein M (2005) PubNet: A flexible system for visualizing literature-derived networks. Genome Biol 6: R80. Available: http://pubnet.gersteinlab.org. Accessed 22 March 2007.
2. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, et al. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391: 806–811.

Citation: Gerstein M, Douglas SM (2006) RNAi development. PLoS Comput Biol 3(4): e80. doi:10.1371/journal.pcbi.0030080

Funding: The authors received no specific funding for this study.

Competing interests: The authors declare that there are no competing interests.