Minireview

Consistency of genome-based methods in measuring Metazoan evolution

Evgeny M. Zdobnov[a], Christian von Mering[a], Ivica Letunic[a], Peer Bork[a,b,*]

[a] *EMBL Heidelberg, Meyerhofstrasse 1, D-69117 Heidelberg, Germany*
[b] *MDC Berlin-Buch, Robert-Roessle-Strasse 10, Germany*

Accepted 5 April 2005

Available online 18 April 2005

Edited by Gáspár Jékely

Abstract Seven distinct genome-wide divergence measures were applied pairwise to the nine sequenced animal genomes of human, mouse, rat, chicken, pufferfish, fruit fly, mosquito, and two nematode worms (*Caenorhabditis briggsae* and *Caenorhabditis elegans*). Qualitatively, all of these divergence measures are found to correlate with the estimated time since speciation; however, marked deviations are observed in a few lineages. The distinct genome divergence measures also correlate well among themselves, indicating that most of the processes shaping genomes are dominated by neutral events. The deviations from the clock-like scenario in some lineages are observed consistently by several measures, implicitly confirming their reliability.

Keywords: Phylogeny; Metazoan evolution; Genome divergence

1. Introduction

The *Encyclopedia Britannica* defines phylogeny as "the history of the evolution of a species or group, especially in reference to lines of descent and relationships among broad groups of organisms". Nowadays, phylogenetic methods mostly rely on molecular data, assessing the similarities in protein or DNA sequences and looking for the most parsimonious scenarios capable of explaining the data. Qualitatively, the aim is to resolve species genealogies, and quantitatively, the aim is to date the speciation events. The assumption that molecular evolution rates are relatively constant over time became known as the molecular clock hypothesis [1,2]. The molecular divergence measures have to be calibrated against available fossil data to scale the genetic distance into time (reviewed e.g., in [3,4]).

Although the reliability of the molecular clock hypothesis can be questioned (discussed below) it is widely used to "illuminate" the evolutionary history of life [5–7]. It has been recognized that the use of multiple gene families is more robust for deciphering phylogenies [3,8], although it is not clear how the data from a number of genes evolving under different rates should be integrated. Consequently, the complete or nearly complete sequencing of the genomes from many species has led to the development of new approaches that exploit the information from genomes as a whole to reconstruct phylogenies, rather than relying on levels of sequence identity within selected gene families, e.g., the small subunit ribosomal RNA. The proposed methods measure the number of shared orthologous genes or shared gene families between genomes [9–11], or count occurrences of different protein domain combinations [12,13] as evolutionary characters for phylogenetic tree reconstruction.

Here, we apply distinct pairwise genome-wide divergence measures to the nine sequenced animal genomes of human, mouse, rat, chicken, pufferfish, fruit fly, mosquito, and two nematode worms (*Caenorhabditis briggsae* and *Caenorhabditis elegans*) (Fig. 1). Namely, for each pair of organisms we compute: (1) the median protein identity of shared orthologous genes, (2) the fraction of introns remaining in the same positions in these orthologous genes, (3) the fraction of orthologous exons having protein coding insertions or deletions, (4) the sequence identity of well aligned regions of 18S ribosomal RNA, (5) the conservation of genomic gene arrangements (synteny), (6) the variation in the neighboring protein domain architectures, and (7) the fraction of homologues recognized as orthologues.

2. Divergence measures and evolutionary time

2.1. Mutations accumulate with time

Qualitatively, all genome-wide divergence measures tested here correlate with time elapsed since speciation (as estimated from the fossil record, Fig. 1). This general applicability of the molecular clock hypothesis confirms that most of the processes shaping genomes are to a large extent dominated by neutral events [14]. When examined in detail, however, marked deviations from a simple, rate-constant clock model become apparent (Fig. 1). Despite all the associated uncertainties in dating evolutionary events (for example, see the recent debate on Metazoan timing [15–18]) a conservative approach to estimate the reliability of such predictions concluded that molecular time estimates remain a useful tool in evolutionary biology [19], estimating 15–20% accuracy for most of the molecular time estimates in the 10–100 MYA range. Here, we used dates for speciation of *A. gambiae* and *D. melanogaster* from [7,20]; of *C. briggsae* and *C. elegance* from [21]; and the others from [5,6]. Although the dates have significant inherent uncertainties (not shown in Fig. 1) and should be taken with caution, it is clear that the higher divergence in the insect and worm lineages

*Corresponding author. Fax: +49 6221 387 517.
E-mail address: bork@embl.de (P. Bork).

0014-5793/$30.00
doi:10.1016/j.febslet.2005.04.006

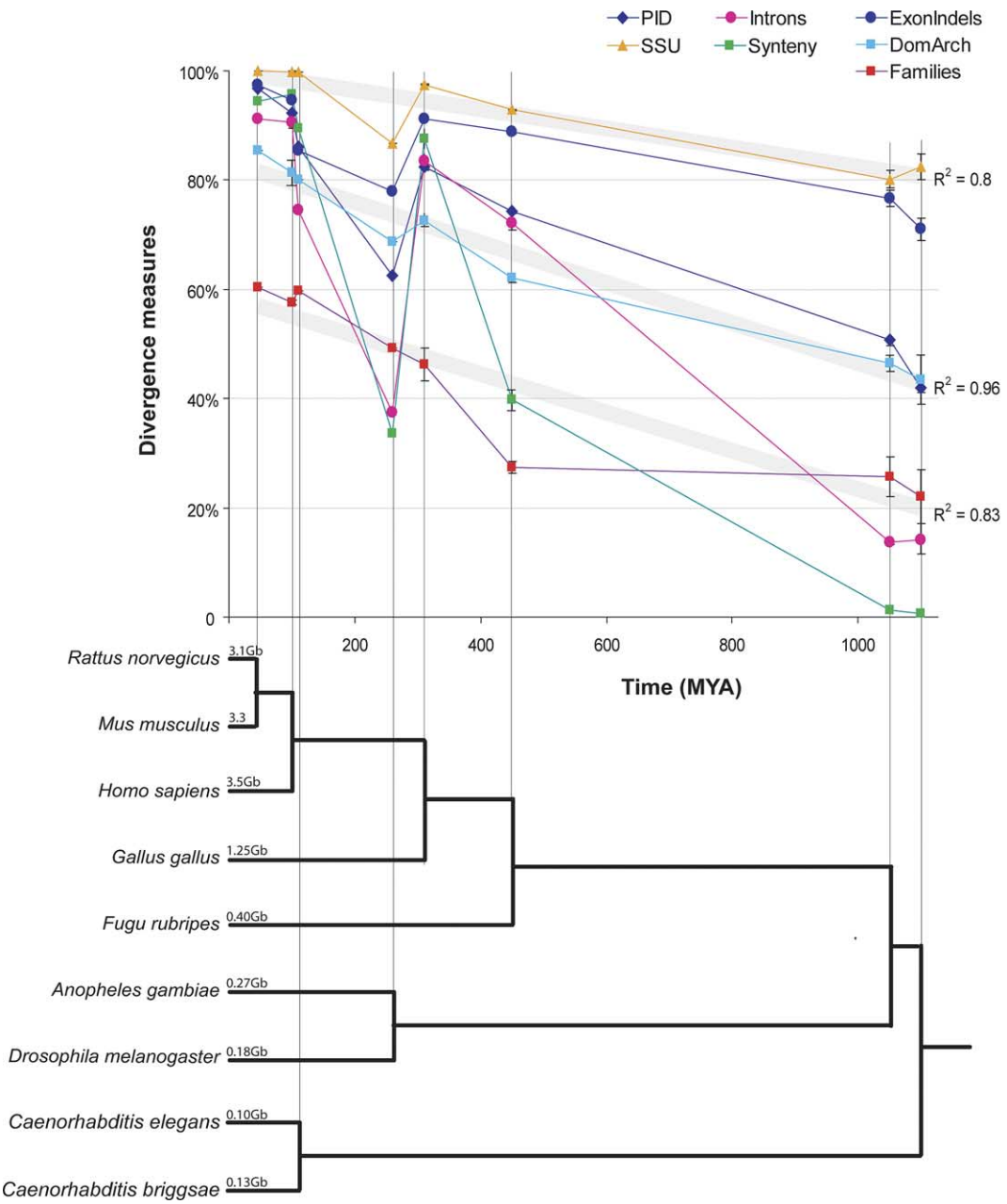

Fig. 1. Divergence of metazoan genomes with respect to time, plotted for the following pairwise divergence measures: (1) median protein identity of shared orthologous genes (PID), (2) fraction of introns remaining in the same positions in these orthologous genes (Introns), (3) fraction of orthologous exons having protein coding insertions or deletions (ExonIndels), (4) sequence identity of well aligned regions of 18S ribosomal RNA (SSU), (5) conservation of genomic gene arrangements (Synteny), (6) variation in the neighboring protein domain architectures (DomArch), (7) fraction of homologues recognized as orthologues (Families). The current opinion on the species phylogeny and the estimated time elapsed since the last common ancestor we used are shown below the time axis.

cannot be accommodated by these dating uncertainties, and thus indicate indeed different rates of genome evolution in different lineages. This general but rather 'sloppy' molecular clock seems to be also true for prokaryotes, where the clock-like behavior has been shown for 70% of several hundred protein coding genes in orthologous gene clusters from the three major bacterial lineages [22].

2.2. Faster divergence rate in insects and worms

Remarkably, the deviations from the clock-like scenario in Fig. 1 are often consistently observed across several measures, most likely because some of the factors influencing divergence rates act globally, and not only specifically on one type of object measured. For example, all but one of the measures seem to indicate that diptera evolve at a much faster rate than vertebrates (Fig. 1). This has been previously observed for a number of measures, including protein sequences [23], rRNA divergence [24] and chromosomal rearrangements [25]. We show that other measures follow this trend, such as the fraction of retained introns, or the degree of shared protein domain architectures with exception of the fraction of homologues recognized as orthologues. A similar deviation is

seen in the nematode lineage, in terms of conservation of sequence in orthologous proteins (measured by both sequence identity and insertions/deletions), introns, and genomic neighborhood (synteny). While the rise of the orthologous gene fraction in worms is a likely artifact of the incomplete *C. briggsae* gene set obtained from Ensembl [26] as suggested by the different numbers quoted in the *C. briggsae* genome analysis paper [21], the rRNA sequence divergence is in line with the predicted speciation time. The reliability of the molecular clock hypothesis has been questioned long before [14,27–30]; recently reviewed in [4], and the genomic measures used here consistently indicate that it does not hold in its exact sense, i.e. the rate of genome evolution is non-uniform in different species lineages, despite a good general correlation with time.

2.3. Neutral divergence of functionally constrained sequences

Although there is a clear correlation of all genome-based divergence measures with time, it is also apparent from Fig. 1 that measures under a high load of functional constraints dissipate slower over time. For example, the less functionally constrained gene order and intron conservation are the fastest evolving characters. In contrast, ribosomal RNA and small insertions or deletions (indels) in orthologous protein coding exons are probably the most functionally constrained, and thus they are the slowest-evolving characters (Fig. 1). This can be explained in the frame of the neutral theory of genome sequence evolution [14,28], as more functionally constrained characters have less capacity to accommodate nearly neutral mutations and thus many mutations are discarded by purifying selection. Consequently, different rates of accumulation of different types of mutations define windows of their applicability for phylogenetic reconstruction. For example, neutrally evolving non-genic DNA sequences can mutate beyond recognition already at distances such as human and chicken [31] or fruit fly and mosquito [32]. On the other hand, for highly functionally constrained objects it might be hard to recognize those characters that can still be neutrally mutated and to what extent they are already saturated with mutations.

3. Consistency of distinct divergence measures

3.1. Correlation of measures with different functional constraints

Apart from the correlation of all measures with time, they also correlate well with each other. For example, considering the similarity of orthologous protein coding genes as a baseline for comparing metazoan genomes (the median identity of mutually best matching proteins, marked as PID in Fig. 2A), unambiguous groupings consistent with the suggested phylogenetic tree are observed (Fig. 2A). If, in turn, we plot the other pairwise divergence measures against the protein identity we can see their agreement in each of the tree branching events (Fig. 2B). The 'clouds' of dots indicate the precision of the estimates as they result from different pairwise comparisons at a given divergence point. All the measures display strong correlation among themselves, as exemplified by their correlation with protein identity (see R^2 values along the Fig. 2 legend). The slopes of the linear approximations indicate the relative rates of their divergence, showing that genomic neighborhood (synteny) and intron conservation is much less constrained

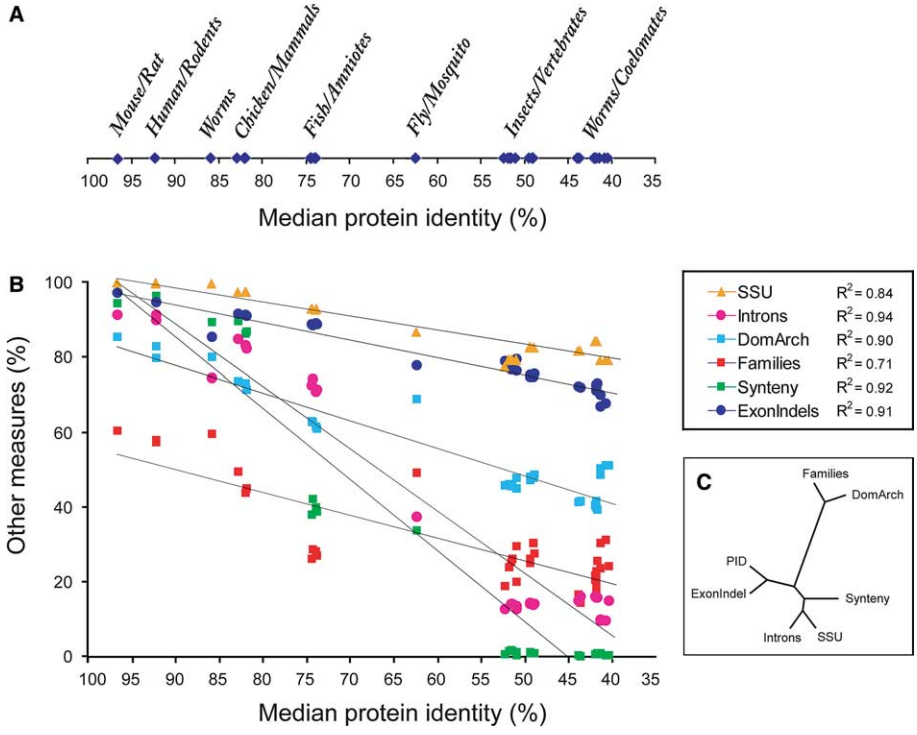

Fig. 2. (A) Distribution of median protein identity of orthologous genes shared between each pair of the Metazoan genomes considered. (B) Correlation between the orthologue median identity and the other genome divergence measures shown in Fig. 1. Each dot corresponds to a particular pair of organisms (see panel A for reference). The corresponding linear coefficients of determination (R-squared) are shown on the legend. (C) Methods relation tree as computed by unweighted pair-group average clustering of pairwise Pearson correlation coefficients.

than the evolution of protein coding or rRNA genes. Using the Pearson correlation coefficient among the measures as their relative distance, their relations could be expressed as a tree (Fig. 2C). This tree of different genome divergence measures groups together, as expected, protein identity (PID) and insertions/deletions in coding exons (ExonIndels) as well as variation in the neighboring protein domain architectures (DomArch) and fraction of homologues recognized as orthologues (Families). Unexpectedly, it also groups together an intron-based measure (fraction of introns remaining in the same positions in orthologous genes) and a structural RNA-based measure (identity of well aligned regions of 18S ribosomal RNA). A priori it is unclear why the two types of mutations (intron gain/loss vs. rRNA sequence change) should be rate-correlated. However, it may be relevant that they both might require a double change (co-variation) in distant sequence positions that would keep or make a functional base-pairing (stems in rRNA, and the two splice junctions, respectively).

3.2. Current limits of the measures: coelomata or ecdysozoa

The traditional topology of the animal phylogenetic tree based on comparative anatomy joins together animals with a true body cavity (Coelomates, such as arthropods and chordates), whereas animals that have a pseudocoelome, such as nematodes, and those without a coelome, such as flatworms, are considered more basal [33]. This hypothesis has been questioned on the basis of 18S ribosomal RNA analysis, which clustered arthropods and nematodes in a clade of molting animals termed Ecdysozoa [34]. The ecdysozoan scenario gained a wide popularity being further supported by independent phylogenetic analysis of 18S RNA [35,36] and by combined analysis of 18S and 28S rRNA sequences [37]. Apparently, the ecdysozoan topology was recovered only when certain species of nematodes, which evolve slowly, were included in the analysis. Contrary to this, the evolution of protein coding sequences provides a clear support for the Coelomata scenario [13] on the bases of analysis of over 500 sets of orthologous proteins. However, this type of evidence has been questioned [12] as it could be biased toward grouping arthropods with chordates by the systematic high rate of character loss in the nematodes. We followed the Coelomata hypotheses for the animal tree topology in Fig. 1 as it was only contradicted by the analysis of 18S rRNA (labeled SSU for small subunit) and the fraction of retained orthologous introns. As it has been noted above, we also see a strong signal that insect proteins are much more like vertebrate ones (Fig. 2), while all the other measures (both pro and con), in our opinion, do not provide strong enough signal to definitively resolve the Coelomata versus Ecdysozoa hypotheses. More genomes at the right phylogenetic distance would have to be sequenced in order to increase the resolution of the measures.

3.3. Factors influencing mutation rates

There are several hypotheses on how differences in generation times, metabolic rates, effective population sizes, reproductive strategies, etc. in different taxonomic groups can explain the observed differences in molecular evolution rates. For example, the greater the effective population size the stronger the effect of purifying selection, and thus the slower the apparent rate of the molecular clock [38]. This relation was recently used to suggest a new theory of genome complexity evolution by Lynch and Conery [39]. It states that the genome complexity emerges as a secondary effect of accumulation of nearly neutral genomic 'junk' when purifying selection is relaxed during population bottlenecks, rather than being a result of positive selection. This has received some support recently by the observation that the gain and loss of introns is weakly correlated with the gain and loss of genes [40] – this correlation is remarkable because these two items are supposedly at opposite ends of the spectrum of functional selection. Our data (focused on animals and based on more genomes) roughly confirm this correlation, although it is obvious that there are also strong deviations in the detailed picture of gene/intron gains and losses (Fig. 3). Nevertheless, the correlated behaviour of introns and genes does lead to a weak correlation of introns per gene and genes per genome (Fig. 3). However,

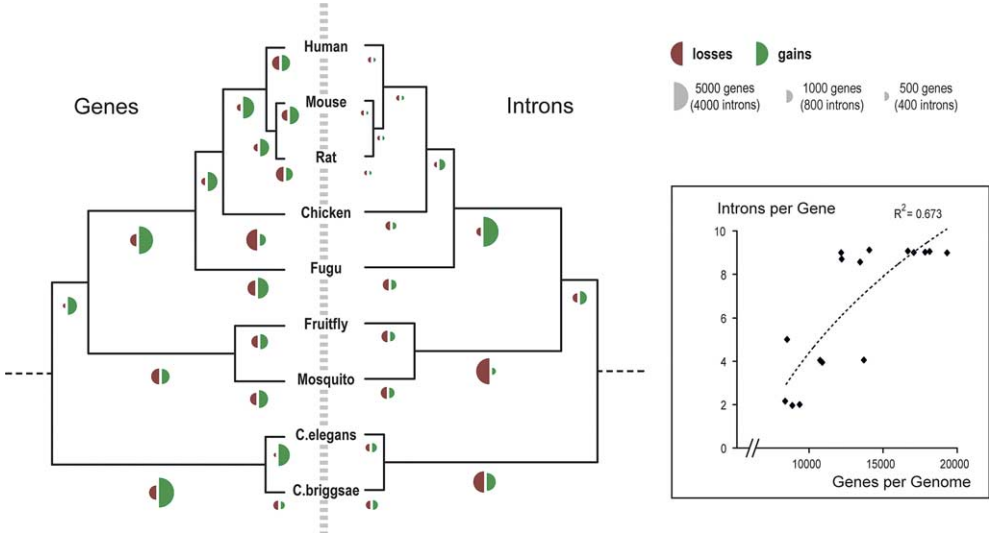

Fig. 3. Reconstructed history of genes and introns in animal evolution. Parsimony was used to infer roughly the contents of ancestral genomes, and gains/losses of both genes and introns are plotted. The inferred ancestral, as well as the present-day genomes show a weak correlation when comparing genes per genome vs. introns per gene.

population size is certainly not the only factor influencing genome evolution; a specific attempt to find different rates of molecular evolution between social and non-social lineages, which differ significantly in effective population size, showed no consistent pattern [41]. The proposed influence of other processes associated with explosive radiations such as body size, morphological rate, speciation rate, and ecological diversification on the rate of molecular evolution has also yet to be confirmed [42,43].

4. Conclusions

Most of the processes that shape genomes appear to be to a large extent dominated by neutral events. As a consequence, several distinct genome divergence measures, not a priori related, roughly correlate with time and among themselves. Despite all the controversies regarding the applicability of the molecular clock hypothesis in dating evolutionary events, the picture of a sloppy clock obtained using a particular gene or a set of genes extends to whole genome based measures as shown here. On the other hand, there are marked deviations from the clock-like model and some disagreement between different measures with respect to some of the divergence points. The somewhat unexpected general consistency of the methods implies further that the genome-wide measures used here seem accurate enough to capture major trends in metazoan evolution.

5. Materials and methods

All proteins and gene exon-intron structures were obtained from Ensembl (ftp.ensembl.org, [26]). The following gene sets were used: *H. sapiens* – v19.34a; *M. musculus* – v19.30; *R. norvegicus* – v19.3a; *G. gallus* – v22.1.1; *T. rubripes* – v21.2c.1; *D. melanogaster* – v19.3a; *A. gambiae* – v19.2a; *C. elegans* – v19.102; *C. briggsae* – v19.25. Orthologous genes were inferred through Smith-Waterman [44] all-against-all similarity searches at the level of predicted proteins, defined as reciprocally best matching genes in pairwise comparisons (e.g., for calculating median protein identity, or identification of synteny), while orthologous groups shared among several organisms were defined through identification of reciprocal triangles as described earlier [31,32,45].

For the comparison of intron positions, a set of 1148 core-orthologues was derived by selecting orthologous groups that covered all organisms, and contained between 9 and 12 genes per group. Multiple genes per organism were allowed (cases where recent duplications have occurred within the group), in order to achieve a sufficiently large set of core-orthologues; in such cases only one of the duplicated genes was chosen, at random, when analyzing intron positions. To detect orthologous introns within a set of orthologous genes, predicted proteins were aligned using ClustalW [46], and intron positions were mapped onto the alignment. Introns were considered orthologous if they had the same phase and were within 4 amino acids (12 nucleotides) from each other. Orthologous exons were derived from the same gene set. To create a very stringent set, bordering orthologous introns were required on both exon sides, with no additional introns in between. Exon insertion-deletion numbers were acquired by comparing lengths of aligned orthologous exons.

Identity of 18S ribosomal RNA was calculated over their structural alignment downloaded from http://www.psb.ugent.be/rRNA/ssu/ [47] and filtered for well conserved columns using gBlocks server (http://molevol.ibmb.csic.es/Gblocks_server/, [48]).

Genomic synteny blocks were identified using SyntQL (Zdobnov, unpublished) as described earlier in [31,32], by looking for a conserved neighborhood of orthologous gene pairs but allowing up to 4 intervening genes and micro-rearrangements inside otherwise orthologous chromosomal loci.

To identify known protein domains we scanned the respective proteomes for characteristic HMM profile signatures from Pfam [49] and SMART [50] databases using the HMMER (S. Eddy, http://hmmer.wustl.edu/) software and applying corresponding domain specific cut-offs. The extent of proteome divergence through protein domain shuffling was estimated by counting all unique domain combinations consecutive on the protein sequences.

The fraction of homologues recognized as orthologues was calculated as a fraction of the pairwise orthologues over the sum of both gene sets exhibiting at least 60 bit homology score to these orthologues (a variant strategy, counting only orthologues with identified domains with respect to all proteins containing at least one of these domains gives a very similar picture).

For measuring gene gain/loss, each orthologous group was counted as a single gain in one ancestral organism whose descendants are needed to cover all of the proteins in the group; orthologous groups were potentially counted multiple times as losses (depending on their pattern of species coverage), assuming a parsimonious scenario with as few losses as possible in order to accommodate the observed pattern. An identical procedure was applied to all intron positions in the protein alignment of each orthologous group to estimate intron gains and losses in various taxonomic branches.

To estimate the number of genes in extant genomes, for each genome we counted all the genes present in the orthologous groups (i.e., having at least one recognizable orthologue in any of the other genomes). In addition, we considered those genes that had paralogy support within the genome – but only if the similarity at the protein level was found to be sufficiently strong to rule out most cases of fragmentation, pseudogenes or gene-prediction artifacts. Specifically, similarity within the genome had to be above 200 bits in Smith-Waterman searches, covering at least 200 amino acids in each of the proteins, and the similarity within the genome had to be higher than the similarity towards any protein in any of the other genomes.

References

[1] Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G. and Little, E. (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271, 470–477.
[2] Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. J. Theor. Biol. 8, 357–366.
[3] Thorne, J.L. and Kishino, H. (2002) Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51, 689–702.
[4] Bromham, L. and Penny, D. (2003) The modern molecular clock. Nat. Rev. Genet. 4, 216–224.
[5] Kumar, S. and Hedges, S.B. (1998) A molecular timescale for vertebrate evolution. Nature 392, 917–920.
[6] Hedges, S.B. (2002) The origin and evolution of model organisms. Nat. Rev. Genet. 3, 838–849.
[7] Wiegmann, B.M., Yeates, D.K., Thorne, J.L. and Kishino, H. (2003) Time flies, a new molecular time-scale for brachyceran fly evolution without a clock. Syst. Biol. 52, 745–756.
[8] Rokas, A., Williams, B.L., King, N. and Carroll, S.B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.
[9] Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. Proc. Natl. Acad. Sci. USA 95, 5849–5856.
[10] Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content. Nat. Genet. 21, 108–110.
[11] Korbel, J.O., Snel, B., Huynen, M.A. and Bork, P. (2002) SHOT: a web server for the construction of genome phylogenies. Trends Genet. 18, 158–162.
[12] Copley, R.R., Aloy, P., Russell, R.B. and Telford, M.J. (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. Evol. Dev. 6, 164–169.
[13] Wolf, Y.I., Rogozin, I.B. and Koonin, E.V. (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14, 29–36.
[14] Ohta, T. (1987) Very slightly deleterious mutations and the molecular clock. J. Mol. Evol. 26, 1–6.

[15] Shields, R. (2004) Pushing the envelope on molecular dating. Trends Genet. 20, 221–222.
[16] Reisz, R.R. and Muller, J. (2004) Molecular timescales and the fossil record: a paleontological perspective. Trends Genet. 20, 237–241.
[17] Hedges, S.B. and Kumar, S. (2004) Precision of molecular time estimates. Trends Genet. 20, 242–247.
[18] Graur, D. and Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet. 20, 80–86.
[19] Glazko, G.V., Koonin, E.V. and Rogozin, I.B. (2005) Molecular dating: ape bones agree with chicken entrails. Trends Genet. 21, 89–92.
[20] Gaunt, M.W. and Miles, M.A. (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol. Biol. Evol. 19, 748–761.
[21] Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D.H., Fulton, L.A., Fulton, R.E., Griffiths-Jones, S., Harris, T.W., Hillier, L.W., Kamath, R., Kuwabara, P.E., Mardis, E.R., Marra, M.A., Miner, T.L., Minx, P., Mullikin, J.C., Plumb, R.W., Rogers, J., Schein, J.E., Sohrmann, M., Spieth, J., Stajich, J.E., Wei, C., Willey, D., Wilson, R.K., Durbin, R. and Waterston, R.H. (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol. 1, E45.
[22] Novichkov, P.S., Omelchenko, M.V., Gelfand, M.S., Mironov, A.A., Wolf, Y.I. and Koonin, E.V. (2004) Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J. Bacteriol. 186, 6575–6585.
[23] Caccone, A. and Powell, J.R. (1990) Extreme rates and heterogeneity in insect DNA evolution. J. Mol. Evol. 30, 273–280.
[24] Friedrich, M. and Tautz, D. (1997) An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. Mol. Biol. Evol. 14, 644–653.
[25] Ranz, J.M., Casals, F. and Ruiz, A. (2001) How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. Genome Res. 11, 230–239.
[26] Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez- Suarez, X.M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C. and Birney, E. (2005) Ensembl 2005. Nucleic Acids Res. 33 (Database Issue), D447–D453.
[27] Gaut, B.S., Muse, S.V., Clark, W.D. and Clegg, M.T. (1992) Relative rates of nucleotide substitution at the rbcL locus of monocotyledonous plants. J. Mol. Evol. 35, 292–303.
[28] Kimura, M. and Takahata, N. (1983) Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. Proc. Natl. Acad. Sci. USA 80, 1048–1052.
[29] Fitch, W.M. and Langley, C.H. (1976) Protein evolution and the molecular clock. Fed. Proc. 35, 2092–2097.
[30] Langley, C.H. and Fitch, W.M. (1974) An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3, 161–177.
[31] Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., Dodgson, J.B., Chinwalla, A.T., Cliften, P.F., Clifton, S.W., Delehaunty, K.D., Fronick, C., Fulton, R.S., Graves, T.A., Kremitzki, C., Layman, D., Magrini, V., McPherson, J.D., Miner, T.L., Minx, P., Nash, W.E., Nhan, M.N., Nelson, J.O., Oddy, L.G., Pohl, C.S., Randall-Maher, J., Smith, S.M., Wallis, J.W., Yang, S.P., Romanov, M.N., Rondelli, C.M., Paton, B., Smith, J., Morrice, D., Daniels, L., Tempest, H.G., Robertson, L., Masabanda, J.S., Griffin, D.K., Vignal, A., Fillon, V., Jacobbson, L., Kerje, S., Andersson, L., Crooijmans, R.P., Aerts, J., van der Poel, J.J., Ellegren, H., Caldwell, R.B., Hubbard, S.J., Grafham, H., Kierzek, A.M., McLaren, S.R., Overton, I.M., Arakawa, H., Beattie, K.J., Bezzubov, Y., Boardman, P.E., Bonfield, J.K., Croning, M.D., Davies, R.M., Francis, M.D., Humphray, S.J., Scott, C.E., Taylor, R.G., Tickle, C., Brown, W.R., Rogers, J., Buerstedde, J.M., Wilson, S.A., Stubbs, L., Ovcharenko, I., Gordon, L., Lucas, S., Miller, M.M., Inoko, H., Shiina, T., Kaufman, J., Salomonsen, J., Skjoedt, K., Wong, G.K., Wang, J., Liu, B., Wang, J., Yu, J., Yang, H., Nefedov, M., Koriabine, M., Dejong, P.J., Goodstadt, L., Webber, C., Dickens, N.J., Letunic, I., Suyama, M., Torrents, D., von Mering, C., Zdobnov, E.M., Makova, K., Nekrutenko, A., Elnitski, L., Eswara, P., King, D.C., Yang, S., Tyekucheva, S., Radakrishnan, A., Harris, R.S., Chiaromonte, F., Taylor, J., He, J., Rijnkels, M., Griffiths-Jones, S., Ureta-Vidal, A., Hoffman, M.M., Severin, J., Searle, S.M., Law, A.S., Speed, D., Waddington, D., Cheng, Z., Tuzun, E., Eichler, E., Bao, Z., Flicek, P., Shteynberg, D.D., Brent, M.R., Bye, J.M., Huckle, E.J., Chatterji, S., Dewey, C., Pachter, L., Kouranov, A., Mourelatos, Z., Hatzigeorgiou, A.G., Paterson, A.H., Ivarie, R., Brandstrom, M., Axelsson, E., Backstrom, N., Berlin, S., Webster, M.T., Pourquie, O., Reymond, A., Ucla, C., Antonarakis, S.E., Long, M., Emerson, J.J., Betran, E., Dupanloup, I., Kaessmann, H., Hinrichs, A.S., Bejerano, G., Furey, T.S., Harte, R.A., Raney, B., Siepel, A., Kent, W.J., Haussler, D., Eyras, E., Castelo, R., Abril, J.F., Castellano, S., Camara, F., Parra, G., Guigo, R., Bourque, G., Tesler, G., Pevzner, P.A., Smit, A., Fulton, L.A., Mardis, E.R. and Wilson, R.K. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695–716.
[32] Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R., Christophides, G.K., Thomasova, D., Holt, R.A., Subramanian, G.M., Mueller, H.M., Dimopoulos, G., Law, J.H., Wells, M.A., Birney, E., Charlab, R., Halpern, A.L., Kokoza, E., Kraft, C.L., Lai, Z., Lewis, S., Louis, C., Barillas-Mury, C., Nusskern, D., Rubin, G.M., Salzberg, S.L., Sutton, G.G., Topalis, P., Wides, R., Wincker, P., Yandell, M., Collins, F.H., Ribeiro, J., Gelbart, W.M., Kafatos, F.C. and Bork, P. (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science 298, 149–159.
[33] Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R. and Raff, R.A. (1988) Molecular phylogeny of the animal kingdom. Science 239, 748–753.
[34] Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A. and Lake, J.A. (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387, 489–493.
[35] Peterson, K.J. and Eernisse, D.J. (2001) Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol. Dev. 3, 170–205.
[36] Giribet, G., Distel, D.L., Polz, M., Sterrer, W. and Wheeler, W.C. (2000) Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. Syst. Biol. 49, 539–562.
[37] Mallatt, J. and Winchell, C.J. (2002) Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. Mol. Biol. Evol. 19, 289–301.
[38] Ohta, T. (1995) Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. 40, 56–63.
[39] Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. Science 302, 1401–1404.
[40] Koonin, E.V. (2004) A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. Cell Cycle 3, 280–285.
[41] Bromham, L. and Leys, R. (2005) Sociality and the rate of molecular evolution. Mol. Biol. Evol..
[42] Bromham, L. (2003) Molecular clocks and explosive radiations. J. Mol. Evol. 57 (Suppl 1), S13–S20.
[43] Bromham, L. and Woolfit, M. (2004) Explosive radiations and the reliability of molecular clocks: island endemic radiations as a test case. Syst. Biol. 53, 758–766.

[44] Smith, T.F., Waterman, M.S. and Fitch, W.M. (1981) Comparative biosequence metrics. J. Mol. Evol. 18, 38–46.
[45] Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Rogozin, I.B., Smirnov, S., Sorokin, A.V., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. 5, R7.
[46] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (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.
[47] Wuyts, J., Van de Peer, Y., Winkelmans, T. and De Wachter, R. (2002) The European database on small subunit ribosomal RNA. Nucleic Acids Res. 30, 183–185.
[48] Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.
[49] Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) The Pfam protein families database. Nucleic Acids Res. 32, D138–D141.
[50] Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004) SMART 4.0: towards genomic data integration. Nucleic Acids Res. 32, D142–D144.