Genome and protein evolution in eukaryotes

Richard R Copley*, Ivica Letunic and Peer Bork

The past year has seen the completion of the genome sequence of the flowering plant *Arabidopsis thaliana* and the initial sequence reports of the human genome. The availability of completely sequenced eukaryotic genomes from disparate phylogenetic lineages has opened the door to comparative analyses and a better understanding of the evolutionary processes shaping genomes. Complex many-to-many relationships between genes from different species appear to be the norm, suggesting that transfer of detailed functional annotation will not be straightforward. In addition to expansion and contraction of gene families, new genes evolve from recombination of pre-existing domains, although some domain families do appear to have evolved recently and to be specific to restricted phylogenetic lineages. The overall picture is of a huge diversity of gene content within eukaryotic genomes, reflecting different functional demands in different species.

Addresses
European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany
*e-mail: copley@embl-heidelberg.de

Current Opinion in Chemical Biology 2001, 6:39–45

1367-5931/01/$ – see front matter

Published online 29 November 2001

Introduction

Despite the publication of the genomes of *Saccharomyces cerevisiae* in 1996 [1], *Caenorhabditis elegans* in 1998 [2] and, more recently, *Drosophila* [3], the relative abundance of completed prokaryotic genome sequences has inevitably focussed comparative studies of genome evolution on prokaryotes. Now, the pivotal additions of the first plant genome, that of *Arabidopsis thaliana* [4••] and the rough drafts of the human genome [5••,6••], have paved the way for a more profound understanding of the factors shaping eukaryotic genome evolution. As more complete eukaryotic genome sequences rapidly become available, the results of comparative analyses are likely to become more informative, but progress is already being made in understanding the evolution of the gene and protein content of genomes.

Eukaryotic gene numbers

Perhaps the most basic relevant measure of a genome is how many genes it contains. Gene prediction in eukaryotes is notoriously difficult; it is not even a straightforward business to decide how to count [7] and, in humans at least, there is considerable discrepancy between different gene prediction methods [8]. Clearly, phenomena such as alternative splicing and post-translational modifications of genes must also be taken into account. Accurate prediction of alternative gene splicing is not yet available, although its contribution to genome complexity is expected to be very significant [9], and regulation of transcription and translation will need to be better understood to fully get to grips with genome complexity.

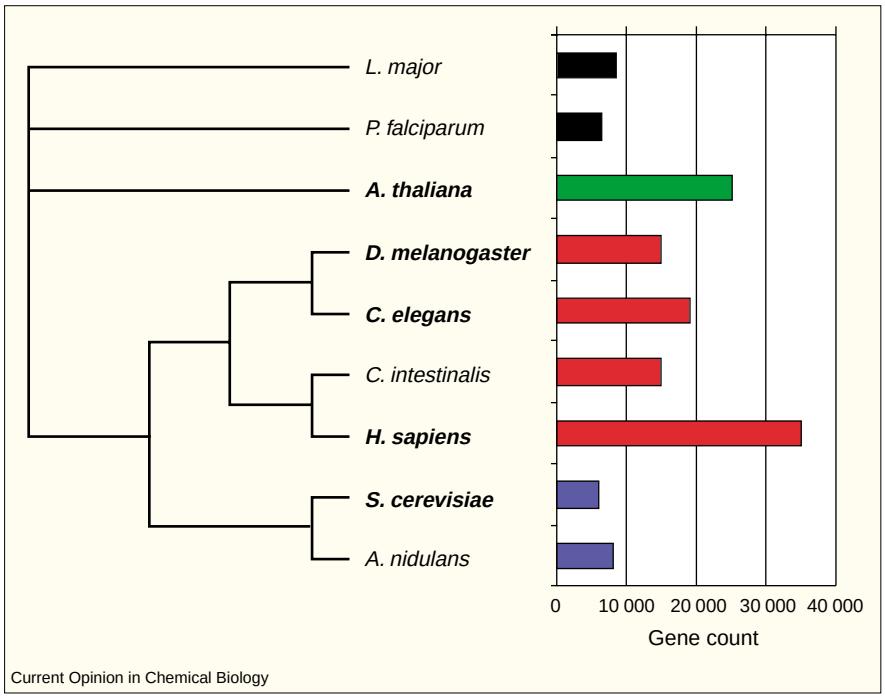
Although exact numbers of genes are uncertain, some gross trends are apparent. Gene numbers of around 15 000 for *Drosophila* [10] and the urochordate *Ciona intestinalis* [11], and 19 000 in *C. elegans* [2] (with experimental evidence for at least 17 300 [12]), contrast with around 25 000 in *Arabidopsis* [4••] and 30 000–40 000 in humans [5••,6••]. *S. cerevisiae* has roughly 6000 genes (although multicellular ascomycetous fungi probably contain a third more [13]), *Leishmania major* about 8500 [14] and the malaria parasite *Plasmodium falciparum* about 6500 [15,16]. Interpreted in the light of eukaryotic phylogeny [17,18], these numbers suggest that the ancestral eukaryote had around 6000 genes (in line with large prokaryotic genome sizes), independent expansions of gene numbers have taken place in plants and metazoa, with the ancestral metazoan having something of the order of 15 000 genes, and that gene numbers have further doubled (at least) in the lineage leading to vertebrates (see Figure 1).

How can we account for these variations in gene number? The fact that the number for plants is so high, and that *C. elegans* seems to have around a third more genes than *Drosophila* suggests that, beyond multicellularity, there is not a stately progression of gene count with what might be regarded as intuitive measures of complexity. Where do the new genes come from, and what do they do?

Genome and gene duplication

New genes evolve from old genes. One way in which the raw material can be supplied is via a whole genome duplication event. In the past, this idea had wide acceptance, with evidence of increased numbers of *hox* genes, interpreted in the light of the expectation of around 80 000 genes in humans, being taken as evidence of two rounds of genome duplication in the early history of the vertebrates. Decreased human gene counts, and systematic studies based on individual gene families have given ample reason to question this hypothesis. In a pertinent recent paper, Hughes *et al.* [19•] examined the phylogeny of 42 gene families that included two or more members on the *hox*-bearing chromosomes, and in 32 found evidence against simultaneous duplication with the *hox* clusters. Moreover, neither of the two initial analyses of the human genome found strong evidence for whole-genome duplication, although large block duplications do occur [5••,6••]. An additional genome duplication event has been proposed for ray-finned fish [20]. Again, the evidence from different sources is contradictory [21–24]. The case for a genome duplication in the lineage leading to *Arabidopsis* appears strong, and is supported by independent lines of evidence

Figure 1



Estimated gene numbers in a variety of eukaryotes, alongside a proposed phylogeny of these species. References for gene number estimates are given in the text. See [17] for a discussion of deep-branching eukaryote relationships Species names in bold represent estimates from complete genome data. The bars are colour coded according to taxanomic group. Black, deep branching eukaryotes; green, plant; red, metazoa; blue, fungi.

[25,26•]. Wolfe [27•] has recently reviewed the arguments for and against genome-wide duplication, and suggested that even if such an event had occurred in the lineage leading to vertebrates, it will no longer appear to have had a major impact in shaping the vertebrate proteome.

Recently duplicated genes are likely to have the same functions, and thus it is possible that one of the duplicate copies will be lost. The fate of duplicated gene pairs has been analysed by Lynch and Conery [26•] and others [28•], by examining the patterns of synonymous and non-synonymous mutations in closely related gene pairs. Using this information it was possible to estimate when gene duplication events occurred, and thus the likely rates of gene duplication and loss.

Genes that do not acquire new functions are likely to become pseudogenes. Pseudogenes are detected in genome sequences by the presence of interruptions such as frameshifts or stop codons in the putative translation of a gene, or by an absence of introns in a copy of a gene that has paralogues with introns ('processed' pseudogenes). Other features that could be used to identify pseudogenes are the absence of promoters, or relaxed selection as evidenced by the ratio of synonymous to non-synonymous nucleotide substitutions. Gerstein and co-workers [29] surveyed the pseudogene population of *C. elegans*, looking for interrupted reading frames or evidence of processing, and found evidence that there was one pseudogene for every eight genes. This compares with around 20% for human chromosome 22 [30], and around 3000 processed pseudogenes detected in the Celera genome sequence [6••]. There does not appear to be a clear correlation between the size of a pseudogene population for a given protein family and the number of real representatives of that family [29].

Lineage-specific gene gain and loss

The concepts of orthology and paralogy have been of great value in genome annotation. Orthologues are genes that are related by a speciation event, whereas paralogues are genes that have duplicated within a genome. The significance of the distinction lies in the fact that orthologues are likely to have the same function. Although for many cases, unambiguous assignments of orthology can be made [5••,6••], the increased gene counts seen within many crown group eukaryotes complicate the use of orthologues to predict functions. If a gene in one organism has multiple copies in another, it is not clear that any will share exactly the same function. Cases of many-to-one orthologous relationships, or many-to-many relationships are likely to represent instances of new function arising and/or of sub-functionalization, where the original function of the gene is partitioned over multiple new copies [31]. Four lamprey *Dlx* genes, for instance, have overall expression patterns similar to six mammalian homologues, although the common ancestor of lampreys and mammals is believed to have had only two such genes [32]. In more dramatic cases of duplication, new functions are likely to be acquired: for instance, in nuclear hormone receptors specific to *C. elegans* ([5••], see Figure 39 in this reference), and in a similar case, steroid receptors found in humans, but not lampreys, have evolved new functions [33].

At the other extreme to the orthology/paralogy distinction, some studies simply count relative numbers of a given protein domain or sequence family in a given organism, seeking to highlight cases where the number in one organism is dramatically different to that in another. This can highlight gross changes in domain usage. It does not, however, draw sufficient attention to the precise relationships between these domains. Even if two organisms have comparable numbers of a given domain, the evolutionary (and, by presumption, functional) links between those domains are not necessarily straightforward. What is ultimately necessary is detailed study of individual families, to tease out the evolutionary relationships within them.

Complex evolutionary relationships between characteristically eukaryotic genes are widespread. Remm and Sonnhammer [34] found that out of 189 groups of *C. elegans* proteins predicted to contain two or more transmembrane domains, for 174, putative human–worm orthology could be assigned, with around 30% of these consisting of simple one-to-one relationships, and 30% many-to-many relationships. The usefulness of the orthology concept depends on the gene family, the phylogenetic distance being considered, and the extent to which functions are conserved. Basic helix–loop–helix (bHLH) transcription factors, for instance, can be divided into 44 families: 36 of these have only animal members, four are found only in plants, and two are found in fungi, suggesting the presence of the domain in the last common ancestor of these eukaryotes, but later adaptation for different functions [35]. Of the 38 total animal families however, 35 are conserved in flies, nematodes and humans, suggesting conserved roles specific to animal development [35]. The last common ancestor of plants and animals probably had one serpin-like gene — now there are distinct monophyletic families with multiple members in plants, insects, nematodes and vertebrates, although none in fungi [36]. The case of the plant receptor kinases, transmembrane proteins involved in signalling, is another dramatic example of differential expansion of a family, with acquisition of new functions. Phylogenetic analysis of the kinase domain suggests that around 600 of these genes found in *Arabidopsis* arose from a single kinase present in the ancestor of plants and metazoa, and that the metazoan Pelle (IRAK) kinases represent the extant metazoan descendants [37].

Family expansions are also seen over much smaller phylogenetic distances. In such cases, chromosomal locations of genes aid the interpretation of evolutionary relationships. Comparing human chromosome 19 with related regions in mouse shows independent expansions of different genes containing the *Krüppel* associated box (KRAB) domain [38•]. The novel function, if any, of these new genes is not clear.

Rapid gene loss and gain can obscure the relationships between genes and confound evolutionary analyses. When insufficient data are available, contraction in one lineage can be mistaken for expansion in the other. Rodent eosinophil-associated RNase genes appeared to have undergone rapid expansions in different rodent lineages [39]. Re-analysis of these relationships including pseudogenes suggested instead that the results were better interpreted in terms of rapid gene sorting, with gene duplication and subsequent deactivation ('pseudo-izaton') of different genes in different lineages [40].

Gene loss is apparent in the fungal lineage leading to *S. cerevisiae* [41•]. Koonin and co-workers [42•] have studied cases of gene loss and rapid divergence of sequences between the fungi *Schizosaccharomyces pombe* and *S. cerevisiae*. Since divergence from a common ancestor, they estimate that around 300 genes have been lost in *S. cerevisiae*, with a further 300 undergoing rapid sequence divergence. Many genes appear to be lost as functionally linked groups, suggesting that analysing patterns of gene loss may allow some level of function prediction. As more eukaryotic genome sequences become available, such techniques are likely to gain importance.

Horizontal transfer versus gene loss

One of the more striking conclusions of the publicly funded human genome project's analysis of the rough draft of the human genome, was that the presence of many genes was best explained by horizontal gene transfer from bacteria into the vertebrate lineage [5••]. This argument was based on the phylogenetic distribution of the closest matches to the genes in question — in all cases, the best matches were to bacterial sequences. Further analyses have convincingly questioned the inference of horizontal transfer [43–45], although for some cases such transfer into the germline of metazoans remains a possibility [46]. These works favour an explanation in terms of multiple gene loss events in eukaryotic gene evolution, and suggest that such widespread gene loss may be more common than previously imagined.

Innovation in proteins

Genes are translated into proteins, which fold to give a functioning three-dimensional structure. Protein structures often reveal the presence of distinct domains — that is, regions of compact three-dimensional structure that can have distinct evolutionary histories but conserved functions. Genome sequences, via their predicted proteins, provide a rich source for understanding how domains evolve and recombine, leading to new functions.

New domain arrangements

A simple first approach to rationalizing phenotypic complexity is to look for protein families that are only found within particular phylogenetic lineages, and then attempt to correlate the presence or absence of these families with particular biological systems. In such studies, it is important to distinguish protein families from domain families. A protein family is defined by having a particular arrangement of domains; new protein families can arise from the shuffling and rearrangement of domains within a

genome. Data from comparisons of human, worm, fly and yeast suggest that increasing complexity of protein architecture correlates with increasing complexity of an organism [5••], although such results are potentially biased by an over-representation of identified human domains in current domain databases.

New domain architectures can arise by a process dubbed 'domain accretion' — over time, an ancestral gene acquires DNA encoding new domains, and thus the protein product gradually becomes more complex, with only part of the gene being related by descent to the original [47]. Examples can be seen in chromatin-associated proteins [5••] and proteins involved in apoptosis [48]. Comparison of domain architectures in such ways relies on having accurate gene predictions. It is easy for a gene prediction to miss a domain, which makes it look as though an orthologous sequence has instead gained a domain. Original annotation of the *C. elegans* genome did not include the TIR domain in the single Toll-like receptor that has recently been identified [49,50], for instance, making it appear as though such receptors are innovations that appeared after the divergence of nematodes from other lineages. Moreover, inferring loss of a domain in one lineage rather than gain in the other relies on accurate phylogenies. The interpretation of accretion from *C. elegans* to *Drosophila* and human (rather than loss in *C. elegans*), for instance, depends crucially on the existence of a coelomate clade, which other lines of evidence question [51–53].

Domain loss is also a possibility. Fungal zuotin proteins, for example, are believed to have lost MYB domains, as these are present in orthologous sequences from plants, metazoa [49,54], and *Leishmania* (RR Copley, unpublished data). An alternative, though less parsimonious, scenario is that both domain architectures were present in an ancestral eukaryote, but have been differentially lost in extant organisms. Another possibility is the existence of a plant/fungal clade rather than an animal/fungal clade, in conflict with current views of eukaryotic phylogeny [18], and the sequences of the genes themselves [54]. MRNA-capping enzymes in metazoa are found with a phosphatase-like domain fused at the N-terminus. This architecture is also found in plants, though not in fungi. Such a scenario is similar to the one described for fungal zuotins — however, in this case, the sequence of the equivalent enzyme in *P. falciparum* resembles the architecture of the fungal sequences [55], suggesting that either two separate fusion events have occurred, or that fungi are better regarded as an outgroup of plants and animals. As always, however, new data can change the perspective: mRNA-capping enzyme appears to be independently fused with adenylate-kinase-like domains in trypanosomes, and the fungus *Candida albicans* encodes a sequence similar to the phosphatase-like domain found in the metazoan proteins (RR Copley, unpublished data), suggesting a particularly complex evolutionary history for these genes.

New domains

New domain architectures in different lineages can be readily detected using sequence-based methods. Cases of invention of protein domains are harder to be sure about. Although many domains appear, at the sequence level, to be specific to particular phylogenetic lineages, it can be difficult to ascertain whether they have progenitors in other organisms. Statistically significant sequence similarity can be lost relatively quickly, but homology can still be inferred if two domains share a common structure. In such cases of rapid sequence divergence, the domain is likely to have acquired a new function, as in the case of the ephrin ectodomain, which is structurally similar to cupredoxins, and was inferred to be related to them, despite little sequence similarity [56]. As more sequences and structures become available, new light can be shed on evolutionary origins of domains. For instance, Grishin [57] has demonstrated that the MH1 domain of SMADs, a family of transcription factors specific to metazoans, is probably homologous to His–Me finger endonucleases, which are found in all kingdoms of life.

The sequence-based families of ephrins and SMADs are found only in metazoans, but structural analysis suggests that they may have precursors in other kingdoms of life. In contrast, the α-helical structure of the Frizzled domain, another apparently metazoan-specific extracellular signalling domain, was found not to be similar to any known folds [58]. Koonin *et al.* [47] have suggested that many new α-helical domains have evolved from coiled-coil structures that are particularly abundant in eukaryotes. Other scenarios cannot be ruled out: Grishin [59] has proposed a plausible route by which an all β class fold could, via a succession of intermediates, evolve into an all α-helical type. Given enough time, new folds will evolve from old folds, and the absence of intermediate forms in protein structure databases will give the appearance of *de novo* invention of fold types.

Contrasting organization of eukaryotic and prokaryotic genomes

Aside for the complement of genes carried, what can we learn about higher order aspects of genome evolution? Is eukaryotic complexity evident at other levels? The genomes of bacteria show evidence of organisation into operons. With the comparison of increasing numbers of bacterial genome sequences, it is possible to identify where gene order is apparently conserved between smultiple species (although whether this always represents conservation, distinct from convergence, is perhaps a moot point). Such conservation is typically found when the proteins all have some functional association, such as operating in a particular pathway or protein complex. This has been used to develop new methods for protein function prediction [60,61]. Attempts to find similar types of organization within eukaryotic genomes have not been fruitful. The arrangement of genes within higher eukaryotic genomes does not appear to be shaped by conserved

higher-order regulatory structures such as operons [62], although microarray data reveal adjacent genes with correlated expression in yeast [63].

Within the vertebrate lineage, even zebrafish and humans show considerable numbers of syntenies [21]. The extent to which these have adaptive significance (are some syntenies preferentially conserved?) or simply reflect the shorter time periods since divergence remains to be seen, although some kinds of order are apparent. For instance, organization is seen at the level of the sex chromosomes [64]. Not only are male-specific genes found on the Y chromosome in mice, but also on the X. This X-linkage is believed to reduce the effect of alleles that would be deleterious in females, while maximizing the advantage for males, during the evolution of such genes. Another kind of higher order is seen in the clustering of highly expressed genes in certain chromosomal domains [65].

As more closely related eukaryotic species are sequenced, we will be better placed to address issues of eukaryotic genome organisation.

Conclusions

The differences between the eukaryotic genomes sequenced so far are striking. Although the case for a conserved eukaryotic core set of genes has been confirmed, it represents only a small proportion of the larger genomes. As a consequence, the relationships between many genes in the completed genomes defy simple summary. This should not be surprising, given the extensive differences in lifestyle between eukaryotes. What we can say is that gene duplication and gene loss seem to be pervasive themes shaping eukaryotic gene content. In addition to these processes, new genes are created by the recombination of old domains, and new domains (or at any rate, dramatically different ones) appear from time to time. In general, we expect loss of a gene or domain to be less common than gain. New genes or domains are free to acquire new functions, but it does not seem likely that pre-existing genes are free to lose their functions.

The overall picture gained from comparison of the currently available eukaryotic genomes is of a dynamic gene content. A clearer picture of these dynamics will emerge as more eukaryotic genome sequences become available. Sequences from organisms such as *Plasmodium falciparum* and *Leishmania major* will give us a broader picture of the basic conservation and variability of the genomes of eukaryotes, and wider genomic coverage of metazoa, including representatives from invertebrate chordates, will enable us to better understand the innovation and change that has led to humans.

Acknowledgement

IL is funded by the Louis-Jeantet Foundation.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: Life with 6000 genes. *Science* 1996, 274:546, 563-567.

2. The *C. elegans* Sequencing Consortium: Genome sequence of the nematode ***C. elegans***: a platform for investigating biology. *Science* 1998, 282:2012-2018.

3. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: The genome sequence of ***Drosophila melanogaster***. *Science* 2000, 287:2185-2195.

4. •• The Arabidopsis Genome Initiative: Analysis of the genome sequence of the flowering plant ***Arabidopsis thaliana***. *Nature* 2000, 408:796-815.
The first complete genome sequence of a plant, and so, according to many current molecular phylogenies, the most divergent eukaryote from human sequenced so far. As a plant, *Arabidopsis* is presumed to have evolved multicellularity independently of the metazoa. The analysis described here is an overview of differences and conserved features in comparison to other eukaryotes.

5. •• Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: Initial sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.
This paper, and the following one, presents the first attempts at making sense of the human genome as a whole. Both are exceptionally wide-ranging, touching on all aspects of human genome evolution, and present valuable snapshots of our current understanding.

6. •• Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al.*: The sequence of the human genome. *Science* 2001, 291:1304-1351.
See annotation to [5••].

7. Bork P, Copley R: The draft sequences. Filling in the gaps. *Nature* 2001, 409:818-820.

8. Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, Kay SA, Schultz PG, Cooke MP: A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* 2001, 106:413-415.

9. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* 2000, 474:83-86.

10. Gopal S, Schroeder M, Pieper U, Sczyrba A, Aytekin-Kurban G, Bekiranov S, Fajardo JE, Eswar N, Sanchez R, Sali A *et al.*: Homology-based annotation yields 1,042 new candidate genes in the ***Drosophila melanogaster*** genome. *Nat Genet* 2001, 27:337-340.

11. Simmen MW, Leitgeb S, Clark VH, Jones SJ, Bird A: Gene number in an invertebrate chordate, ***Ciona intestinalis***. *Proc Natl Acad Sci USA* 1998, 95:4437-4440.

12. Reboul J, Vaglio P, Tzellas N, Thierry-Mieg N, Moore T, Jackson C, Shin-i T, Kohara Y, Thierry-Mieg D, Thierry-Mieg J *et al.*: Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in ***C. elegans***. *Nat Genet* 2001, 27:332-336.

13. Kupfer DM, Reece CA, Clifton SW, Roe BA, Prade RA: Multicellular ascomycetous fungal genomes contain more than 8000 genes. *Fungal Genet Biol* 1997, 21:364-372.

14. Myler PJ, Stuart KD: Recent developments from the Leishmania genome project. *Curr Opin Microbiol* 2000, 3:412-416.

15. Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T *et al.*: The complete nucleotide sequence of chromosome 3 of ***Plasmodium falciparum***. *Nature* 1999, 400:532-538.

16. Wahlgren M, Bejarano MT: A blueprint of 'bad air'. *Nature* 1999, 400:506-507.

17. Philippe H, Germot A, Moreira D: The new phylogeny of eukaryotes. *Curr Opin Genet Dev* 2000, 10:596-601.

18. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF: A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 2000, 290:972-977.

19. • Hughes AL, da Silva J, Friedman R: Ancient genome duplications did not structure the human ***hox***-bearing chromosomes. *Genome Res* 2001, 11:771-780.
Based on phylogenetic analyses of genes occurring on the *hox* gene cluster bearing chromosomes, this paper argues against a whole genome duplication event. It goes on to show that, given these phylogenies are accurate, independent gene duplication and translocation could represent a more parsimonious explanation than simultaneous duplication.

20. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL *et al.*: Zebrafish ***hox*** clusters and vertebrate genome evolution. *Science* 1998, 282:1711-1714.

21. Woods IG, Kelly PD, Chu F, Ngo-Hazelett P, Yan YL, Huang H, Postlethwait JH, Talbot WS: A comparative map of the zebrafish genome. *Genome Res* 2000, 10:1903-1914.

22. Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS: Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res* 2000, 10:1890-1902.

23. Robinson-Rechavi M, Marchand O, Escriva H, Laudet V: An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. *Curr Biol* 2001, 11:R458-R459.

24. Robinson-Rechavi M, Marchand O, Escriva H, Bardet PL, Zelus D, Hughes S, Laudet V: Euteleost fish genomes are characterized by expansion of gene families. *Genome Res* 2001, 11:781-788.

25. Vision TJ, Brown DG, Tanksley SD: The origins of genomic duplications in ***Arabidopsis***. *Science* 2000, 290:2114-2117.

26. • Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. *Science* 2000, 290:1151-1155.
This paper presents an attempt to use genomic data to rationalize the fates of duplicated genes. By using the ratio of synonymous to non-synonymous substitutions within the DNA sequences of duplicated gene pairs, timings of duplication events are proposed, and from this, assuming constant rates of duplication, decay rates are inferred. Recently duplicated genes are found to be under relaxed selective constraints, but the vast majority of duplicated genes are silenced within a few million years. Concerns have been raised about aspects of the analysis [28•], but it is clearly stimulating debate about these important problems.

27. • Wolfe KH: Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2001, 2:333-341.
A very clear review of the arguments involved in proposals of whole genome duplications.

28. • Long M, Thornton K, Zhang L, Gaut BS, Vision TJ, Lynch M, Conery JC: Gene duplication and evolution. *Science* 2001, 293:1551a-.
See annotation to [26•].

29. Harrison PM, Echols N, Gerstein MB: Digging for dead genes: an analysis of the characteristics of the pseudogene population in the ***Caenorhabditis elegans*** genome. *Nucleic Acids Res* 2001, 29:818-830.

30. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ *et al.*: The DNA sequence of human chromosome 22. *Nature* 1999, 402:489-495.

31. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 1999, 151:1531-1545.

32. Neidert AH, Virupannavar V, Hooker GW, Langeland JA: Lamprey ***Dlx*** genes and early vertebrate evolution. *Proc Natl Acad Sci USA* 2001, 98:1665-1670.

33. Thornton JW: Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad Sci USA* 2001, 98:5671-5676.

34. Remm M, Sonnhammer E: Classification of transmembrane protein families in the ***Caenorhabditis elegans*** genome and identification of human orthologs. *Genome Res* 2000, 10:1679-1689.

35. Ledent V, Vervoort M: The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis. *Genome Res* 2001, 11:754-770.

36. Irving JA, Pike RN, Lesk AM, Whisstock JC: Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Res* 2000, 10:1845-1864.

37. Shiu S, Bleecker AB: Receptor-like kinases from ***Arabidopsis*** form a monophyletic gene family related to animal receptor kinases. *Proc Natl Acad Sci USA* 2001, 98:10763-10768.

38. • Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P, Lucas S, Land M, Terry A, Ecale Zhou CL, Rash S *et al.*: Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* 2001, 293:104-111.
This study makes clear the value of comparing closely related genome sequences and, in particular, the light that mouse can shed on human genome evolution. Orthologous gene pairs are, in general, found to be arranged syntenically. Single copy genes are found to be conserved, but tandem clusters of genes are not conserved between the two species, indicating that they are products of recent gene duplication events.

39. Singhania NA, Dyer KD, Zhang J, Deming MS, Bonville CA, Domachowske JB, Rosenberg HF: Rapid evolution of the ribonuclease A superfamily: adaptive expansion of independent gene clusters in rats and mice. *J Mol Evol* 1999, 49:721-728.

40. Zhang J, Dyer KD, Rosenberg HF: Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. *Proc Natl Acad Sci USA* 2000, 97:4701-4706.

41. • Braun EL, Halpern AL, Nelson MA, Natvig DO: Large-scale comparison of fungal sequence information: mechanisms of innovation in ***Neurospora crassa*** and gene loss in ***Saccharomyces cerevisiae***. *Genome Res* 2000, 10:416-430.
This paper, and [42•], highlights the role that gene loss has played in the evolution of the *S. cerevisiae* genome. Both compare data from an incomplete fungal genome (*Neurospora crassa* or *S. pombe*) with the finished genome sequence of *S. cerevisiae*, and hence identify genes that are absent in *S. cerevisiae*. In neither case are the absolute numbers of genes lost that large. Aravind *et al.* emphasize that genes seem to be coeliminated in functionally linked groups. More such comparisons, of other closely related pairs of species, will play an important role in better understanding the dynamics of eukaryotic genome evolution.

42. • Aravind L, Watanabe H, Lipman DJ, Koonin EV: Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci USA* 2000, 97:11319-11324.
See annotation to [41•].

43. Salzberg SL, White O, Peterson J, Eisen JA: Microbial genes in the human genome: lateral transfer or gene loss? *Science* 2001, 292:1903-1906.

44. Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR: Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* 2001, 411:940-944.

45. Roelofs J, Van Haastert PJ: Genes lost during evolution. *Nature* 2001, 411:1013-1014.

46. Andersson JO, Doolittle WF, Nesbo CL: Genomics. Are there bugs in our genome? *Science* 2001, 292:1848-1850.

47. Koonin EV, Aravind L, Kondrashov AS: The impact of comparative genomics on our understanding of evolution. *Cell* 2000, 101:573-576.

48. Aravind L, Dixit VM, Koonin EV: Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science* 2001, 291:1279-1284.

49. Copley RR, Ponting CP, Schultz J, Bork P: Sequence analysis of multidomain proteins: past perspectives and future directions. *Adv Protein Chem* 2001, in press.

50. Pujol N, Link EM, Liu LX, Kurz CL, Alloing G, Tan M, Ray KP, Solari R, Johnson CD, Ewbank JJ: A reverse genetic analysis of components of the Toll signaling pathway in ***Caenorhabditis elegans***. *Curr Biol* 2001, 11:809-821.

51. Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA: Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 1997, 387:489-493.

52. Mushegian AR, Garey JR, Martin J, Liu LX: Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res* 1998, 8:590-598.

53. de Rosa R, Grenier JK, Andreeva T, Cook CE, Adoutte A, Akam M, Carroll SB, Balavoine G: ***Hox*** genes in brachiopods and priapulids and protostome evolution. *Nature* 1999, 399:772-776.

54. Braun EL, Grotewold E: Fungal zuotin proteins evolved from mida1-like factors by lineage-specific loss of MYB domains. *Mol Biol Evol* 2001, 18:1401-1412.

55. Ho CK, Shuman S: A yeast-like mRNA capping apparatus in ***Plasmodium falciparum***. *Proc Natl Acad Sci USA* 2001, 98:3050-3055.

56. Toth J, Cutforth T, Gelinas AD, Bethoney KA, Bard J, Harrison CJ: Crystal structure of an ephrin ectodomain. *Dev Cell* 2001, 1:83-92.

57. Grishin NV: Mh1 domain of Smad is a degraded homing endonuclease. *J Mol Biol* 2001, 307:31-37.

58. Dann CE, Hsieh JC, Rattner A, Sharma D, Nathans J, Leahy DJ: Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. *Nature* 2001, 412:86-90.

59. Grishin NV: Fold change in evolution of protein structures. *J Struct Biol* 2001, 134:167-185.

60. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* 2001, 11:356-372.

61. Huynen M, Snel B, Lathe W III, Bork P: Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 2000, 10:1204-1210.

62. Huynen MA, Snel B, Bork P: Inversions and the dynamics of eukaryotic gene order. *Trends Genet* 2001, 17:304-306.

63. Cohen BA, Mitra RD, Hughes JD, Church GM: A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* 2000, 26:183-186.

64. Wang PJ, McCarrey JR, Yang F, Page DC: An abundance of X-linked genes expressed in spermatogonia. *Nat Genet* 2001, 27:422-426.

65. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA *et al.*: The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 2001, 291:1289-1292.