Orthologs, Paralogs, and Evolutionary Genomics[1]

Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894;
email: koonin@ncbi.nlm.nih.gov

Annu. Rev. Genet.
2005. 39:309–38

First published online as a Review in Advance on August 30, 2005

The *Annual Review of Genetics* is online at http://genet.annualreviews.org

doi: 10.1146/annurev.genet.39.073003.114725

0066-4197/05/1215-0309$20.00

Key Words

homolog, ortholog, paralog, pseudoortholog, pseudoparalog, xenolog

Abstract

Orthologs and paralogs are two fundamentally different types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication. Orthology and paralogy are key concepts of evolutionary genomics. A clear distinction between orthologs and paralogs is critical for the construction of a robust evolutionary classification of genes and reliable functional annotation of newly sequenced genomes. Genome comparisons show that orthologous relationships with genes from taxonomically distant species can be established for the majority of the genes from each sequenced genome. This review examines in depth the definitions and subtypes of orthologs and paralogs, outlines the principal methodological approaches employed for identification of orthology and paralogy, and considers evolutionary and functional implications of these concepts.

Contents

INTRODUCTION 310
HISTORY, DETAILED DEFINITIONS, AND CLASSIFICATION OF ORTHOLOGS AND PARALOGS 311
A Super-Brief History of Homology 311
Orthology and Paralogy: Definitions and Complications . 312
IDENTIFICATION OF ORTHOLOGS AND PARALOGS: PRINCIPLES AND TECHNIQUES 316
EVOLUTIONARY PATTERNS OF ORTHOLOGY AND PARALOGY 321
Coverage of Genomes in Clusters of Orthologs 321
One-to-One Orthologs and Inparalogs 322
Orthologous Clusters and the Molecular Clock 323
Xenologs, Pseudoorthologs, and Pseudoparalogs 324
Protein Domain Rearrangements, Gene Fusions/Fissions, and Orthology 327
FUNCTIONAL CORRELATES OF ORTHOLOGY AND PARALOGY 330
GENERAL DISCUSSION 331
Orthology and Paralogy as Evolutionary Inferences and the Homology Debates 331
Generalized Concepts of Orthology and Paralogy 332
CONCLUSIONS 333

INTRODUCTION

One of the most fascinating aspects of modern genomics is the radical change it brings to evolutionary biology. The availability of multiple, complete genomes of diverse life forms for comparative analysis provides a qualitatively new perspective on homologous relationships between genes. By comparing the sequences of all genes between genomes from different taxa and within each genome, it is, in principle, possible to reconstruct the evolutionary history of each gene in its entirety (within the set of sequenced genomes). This, in turn, will allow a deeper understanding of the general trends in the evolution of genomic complexity and lineage-specific adaptations. Gene histories must be presented in the form of scenarios that comprise several types of elementary events (55, 64, 84). The elementary events of gene evolution can be classified as follows, roughly in the order of relative contribution to the evolutionary process: (*i*) vertical descent (speciation) with modification; (*ii*) gene duplication, also followed by descent with modification; (*iii*) gene loss; (*iv*) horizontal gene transfer (HGT); and (*v*) fusion, fission, and other rearrangements of genes. Vertical descent and duplication might be considered the primary events of genome evolution and have been well recognized in the pregenomic era. In contrast, the crucial evolutionary importance of gene loss, HGT, and gene rearrangements was among the major, fundamental generalizations of the emerging evolutionary genomics (13, 14, 16, 50, 51, 57, 77, 78).

Along with the notion of elementary evolutionary events, all descriptions of evolution of genes, gene ensembles, and, ultimately, complete gene repertoires of organisms rest on certain key concepts of evolutionary biology, primarily, the definitions of homologs, orthologs, and paralogs. Developed long ago by evolutionists, these related concepts and terms have reemerged and have become the subject of intense debate and numerous misunderstandings with the advent of molecular evolution and, subsequently, evolutionary genomics (24, 25, 40, 46, 76, 80, 81, 97). The aversion of some biologists to ideas and terms deriving from evolutionary biology is reflected in the peculiar word

"homologuephobia," albeit used in a tongue-in-cheek manner (80).

Homology, the most general definition, designates a relationship of common descent between any entities, without further specification of the evolutionary scenario. Accordingly, the entities related by homology, in particular, genes, are called homologs. The other two key terms define subcategories of homologs. Orthologs are genes related via speciation (vertical descent), whereas paralogs are genes related via duplication (23). The combination of speciation and duplication events, along with HGT, gene loss, and gene rearrangements, entangle orthologs and paralogs into complex webs of relationships. Correct, coherent usage of these terms would be of certain importance if only to provide clarity to the descriptions of genome evolution. However, beyond semantics, these concepts have distinct and important evolutionary and functional connotations.

In this review, I discuss the intricacies of the definitions of orthologs and paralogs, including several derivative categories of homologs and the respective terms, approaches used for identification of orthologs and paralogs, and the functional implications of orthologous and paralogous relationships.

HISTORY, DETAILED DEFINITIONS, AND CLASSIFICATION OF ORTHOLOGS AND PARALOGS

A Super-Brief History of Homology

The term homolog was introduced by Richard Owen in 1843 to designate "the same organ in different animals under every variety of form and function." Owen clearly distinguished homologs from analogs, which he defined as a "part or organ in one animal which has the same function as another part or organ in a different animal" (72). He attributed homologies to the existence of the same "archetype" (structural plan) in all vertebrates but, not being an evolutionist, did not consider the notion of common origin of homologous organs (74). Homology had been immediately reinterpreted after the publication of Darwin's *Origin of Species* (8). Darwin himself never used the term homology, but less than a year after the publication of the *Origin*, Huxley, in his review of Darwin's work, invoked homology as evidence of evolution (37).

Leaping forward a century, the distinction between orthologs and paralogs and the terms themselves were introduced by Walter Fitch in 1970 in a now classic paper (23). However, in the early 1960s, these concepts were considered in a sufficiently clear form, albeit with the use of different and somewhat awkward wording, in the prescient work of Zuckerkandl & Pauling, which laid the foundations of molecular evolution as a discipline (104, 105). A parallel line of relevant developments involved theoretical and empirical studies of gene duplications and their role in evolution. Although the idea of duplication and its contribution to evolutionary innovation was already present in Fisher's classic work of 1928 (22), the coherent concept was developed in Ohno's famous 1970 book *Evolution by Gene Duplication* (69). Ohno argued that gene duplication, i.e., formation of paralogous genes, is the main process responsible for the emergence of functional novelty during evolution whereby one of the newborn paralogs escapes the pressure of preexisting constraints (purifying selection) and becomes free to evolve a new function. In a subsequent section, I briefly discuss the modern theoretical developments that provide a better-supported, more nuanced perspective on the role of gene duplication in evolution. These advances notwithstanding, Ohno's principal message has certainly withstood the test of time and got a second wind through the discovery of omnipresent duplications in genomes.

For 20 years after Fitch developed the notions of orthologs and paralogs, these definitions quietly stayed within the domain of evolutionary biology. A search of the PubMed database (National Center for Biotechnology

Homologs: genes sharing a common origin

Orthologs: genes originating from a single ancestral gene in the last common ancestor of the compared genomes

Paralogs: genes related via duplication

Information, NIH, Bethesda) shows that between 1971 and 1990, the term ortholog(ue) was used 35 times and the term paralog(ue) only 5 times. This is the low bound of the usage of these terms because many old issues of biological journals, including *Systematic Zoology* which published Fitch's article, are not in PubMed. Nevertheless, these numbers clearly show that orthologs and paralogs were hardly in vogue in the pregenomic era. Indeed, according to the Science Citation Index (**http://isiwok.cit.nih.gov/portal.cgi**), Fitch's article was cited only 48 times in the first 20 years of its postpublication life.

It all changed almost overnight in 1995, when the first two complete genome sequences of cellular life forms, the bacteria *Haemophilus influenzae* and *Mycoplasma genitalium*, were released (26, 28). **Figure 1** shows the striking increase in the usage of the terms "ortholog" and "paralog" during the past 14 years, illustrating the conspicuous change of fortune that genomics brought about for these concepts. Suffice it to say that, in the current version of the PubMed database (which includes publications in all areas of biology and medicine as well as chemistry and physics), more than 1 in every 1000 papers includes "ortholog(ue)" in its title or abstract.

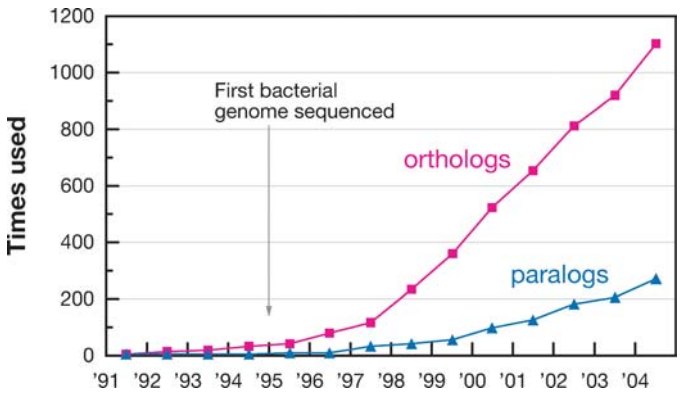


Figure 1

The time dynamics of the usage of the terms "ortholog" and "paralog". The PubMed database was searched using the Entrez search engine with the following queries: "ortholog or orthologs or orthologue or orthologues" and "paralog or paralogs or paralogue or paralogues" to accommodate both the American and the British spelling of the terms.

The advent of complete genomes necessitated a new language in which to discuss the relationships between genomes meaningfully, i.e., the parlance of evolutionary genomics. I would posit that orthology is the keystone definition of evolutionary genomics and paralogy is the paramount, complementary notion (the graphs in **Figure 1** suggest the primacy of orthologous relationships by showing that the term ortholog is used several times more often than paralog). Indeed, in order to make any conclusions regarding evolution of genomes, one first must establish, as precisely as possible, the correspondence between genes in the compared genomes, i.e., orthologous relationships that are inextricably intertwined with paralogous relationships. These constitute the framework on which any evolutionary anomalies and unique events can be mapped. In the following section, I discuss the computational approaches developed to disentangle orthologous and paralogous relationships. Here, I present a general, theoretical breakdown of evolutionary situations in which orthology and paralogy reveal their various faces.

Orthology and Paralogy: Definitions and Complications

Orthologs are genes derived from a single ancestral gene in the last common ancestor of the compared species. This short, simple definition includes two distinct, explicit statements that are important to rationalize; furthermore, it does not include other requirements that might seem natural but are not actually intrinsic to orthology. First, the requirement of a single ancestral gene is central to the concept of orthology. Once the ancestral genome is shown to have contained two paralogous genes that gave rise to the genes in question, it will be incorrect to consider the latter orthologs, even if, on some occasions, there may be the appearance of orthology (see below). Second, the definition specifies the presence of an ancestral gene in the last common ancestor of the compared species rather than

in some arbitrary, more ancient ancestor. Of course, this definition assumes the existence of a distinct common ancestor of the compared species, a proposition sometimes challenged for prokaryotes owing to the high incidence of HGT (see discussion below). This is restrictive and might exclude cases where genes behave like orthologs in the evolutionary and functional senses. Nevertheless, we shall stick to the above definition of orthology for the present discussion. An important statement that might at first seem natural is not included in the definition of orthology: There is no requirement that orthology is a one-to-one relationship. The ensuing discussion shows that such a restriction would have been artificial and meaningless. Of even greater import is the connection between orthology and biological function. It should be emphasized that the above definition has nothing to do with function. However, a crucial property of orthologs, which is both theoretically plausible and empirically supported, is that they typically perform equivalent functions in the respective organisms (I avoid the phrase "identical functions" because, in different biological contexts, functions cannot be literally the same). In a subsequent section, I discuss in some detail both the available evidence of the functional equivalency of orthologs and some notable exceptions. Emphasized here is the asymmetry of the relationship between orthology and function: Orthologs most often have equivalent functions, but the reverse statement is much weaker. Situations when equivalent functions are performed by non-orthologous (often, non-homologous) proteins are common enough as captured in the notion of non-orthologous gene displacement, i.e., recruitment of non-orthologous genes for the equivalent, essential functions in different organisms (47, 52). Therefore it is unadvisable (to put it mildly) to speak of "functional orthologs" whereby functional equivalency is taken as a proxy for orthology; of course, phrases such as "orthologous genes with the same function" are quite legitimate (disregarding, as a subtlety, the above distinction between equivalent and identical functions).

Paralogs are genes related via duplication. Note the generality of this definition, which does not include a requirement that paralogs reside in the same genome or any indication as to the age of the duplication leading to the emergence of paralogs (some of these duplications occurred at the earliest stages of life's evolution but the respective genes nevertheless qualify as paralogs). As in the case of orthology, the definition of paralogy does not refer to biological function, but there are major functional connotations. Generally, paralogs perform biologically distinct, even if mechanistically related, functions. Functional differentiation of paralogs is a complex subject that has been addressed in numerous theoretical and empirical studies; a brief synopsis is given in a subsequent section.

Figure 2 shows a hypothetical phylogenetic tree of a gene family that consists of

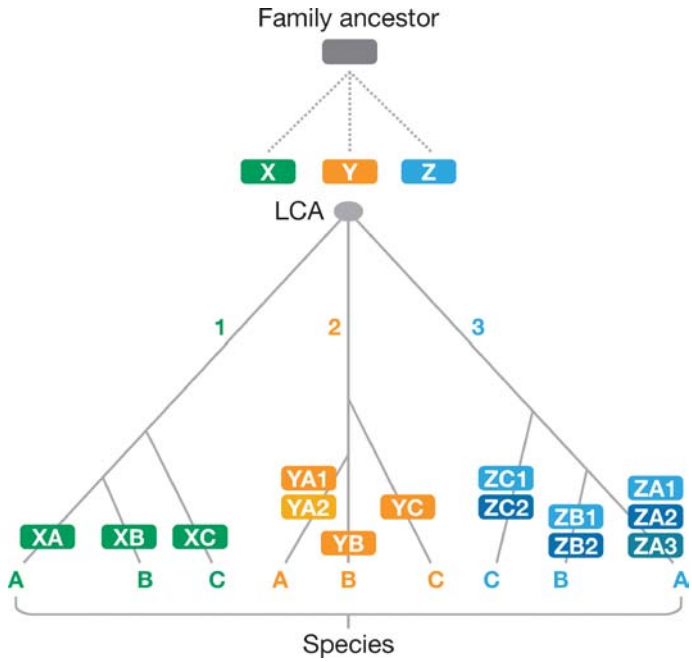


Figure 2

A hypothetical phylogenetic tree illustrating orthologous and paralogous relationships between three ancestral genes and their descendants in three species. LCA, last common ancestor (of the compared species).

Co-orthologs: two or more genes in one lineage that are, collectively, orthologous to one or more genes in another lineage due to a lineage-specific duplication(s)

Outparalogs: paralogous genes resulting from a duplication(s) preceding a given speciation event

Inparalogs: paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event

three branches, each illustrating a distinct case of orthologous-paralogous relationships. Note first of all, that under the evolutionary scenario illustrated by the tree in **Figure 2**, the common ancestor of the entire family existed prior to the last common ancestor of all three compared species. The latter already encoded three paralogous genes from the given family, which became the progenitors of the three branches of the tree. Thus, each gene in branch 1 is a paralog of each gene in branches 2 and 3, and vice versa. Branch 1 corresponds to a straightforward case whereby evolution from the last common ancestor involved nothing but vertical inheritance. Accordingly, the genes in different species are orthologous to each other and, moreover, show a one-to-one orthologous relationship. However, this is only a specific and not the most common form of orthology (at least when large sets of species are analyzed). Branch 2 shows, in addition to the pattern of vertical inheritance, a lineage-specific duplication in species A. This simple situation, nevertheless, requires the introduction of a new group of terms. Since the duplication occurred in a single lineage after the radiation of the analyzed species, the paralogs in species A fit the definition of orthology with respect to all other genes in this branch. Accordingly, genes YA1 and YA2 are coorthologs (85) of the genes YB and YC. In branch 3, the situation is further complicated by lineage-specific duplications in each of the species. Thus, genes ZA1-3 are, collectively, co-orthologs of genes ZB1-2, etc. The scheme in **Figure 2** also points to different classes of paralogs with respect to speciation events. Relative to each speciation event, it makes sense to define outparalogs (alloparalogs), which evolved via ancient duplication(s) preceding the given speciation event (genes X, Y, and Z in **Figure 1**), and inparalogs (symparalogs), which evolved more recently, subsequent to the speciation event (e.g., genes YA1 and YA2 relative to the radiation of species A and B). Even more complex evolutionary scenarios emerge when duplications are associated with internal branches of a phylogenetic tree (rather than with the terminal branches corresponding to species) such that certain gene sets are inparalogs relative to one speciation event and outparalogs relative to another.

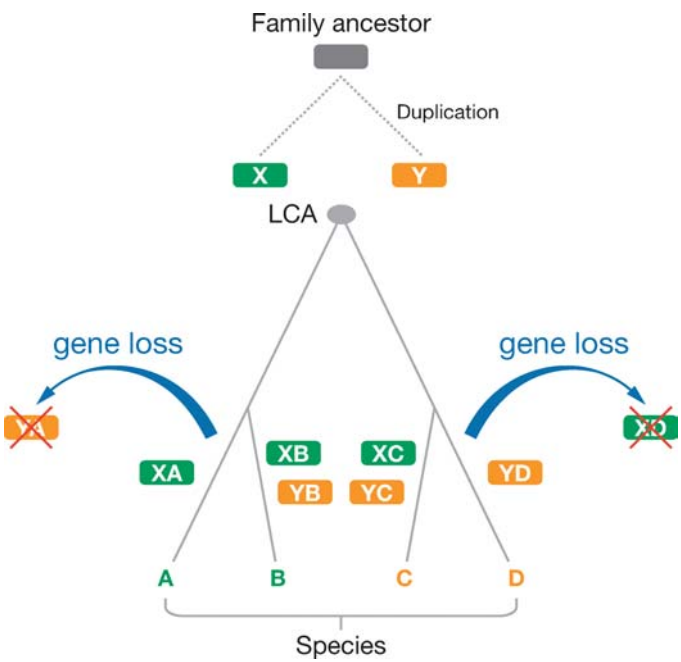


Figure 3

A hypothetical phylogenetic tree illustrating emergence of pseudoorthologs via lineage-specific gene loss.

The terms coorthologs, inparalogs, and outparalogs are relatively new (85) and so far have not been widely adopted. Nevertheless, they seem to be helpful for concise and accurate description of the evolutionary process and functional diversification of genes.

Let us now examine the effect of lineage-specific gene loss on the orthologous and paralogous relationships between genes; **Figure 3** shows a hypothetical example of differential gene loss in two lineages obscuring these relationships. This hypothetical scenario starts with two genes (X and Y) that are outparalogs relative to the included speciation event. Subsequently, gene Y is lost in species A, whereas gene X is lost in species D; the species B and C retain both paralogs (**Figure 3**). By comparing species A and D in

isolation, we might conclude that gene XA is the ortholog of gene YD. Under the scenario in **Figure 3**, this conclusion is obviously false: These genes are paralogs because they evolved from two paralogous genes of the last common ancestor of the compared species (outparalogs). By analyzing the entire set of representatives of the given gene family in the four species and applying the parsimony principle, we can infer the correct evolutionary scenario and, accordingly, draw the conclusion that the genes XA and YD are, actually, outparalogs relative to the divergence of species A and D; these genes only mimic orthology and, for convenience, we may call them pseudoorthologs. Such conclusions are interesting not only from the evolutionary standpoint but may also have substantial functional implications.

Now we must consider the effects of HGT on the observed orthologous and paralogous relationships between genes. As shown in **Figure 4*a***, two species (A and B) may have homologous genes of which one is ancestral for the given lineage but the other has been acquired via HGT from an outside source C, displacing the ancestral gene. This phenomenon has been dubbed xenologous gene displacement [XGD; (51)]. In an even more complex case, both genes might have been acquired from different sources. In a perfunctory analysis, such a pair of genes (XA and XB in **Figure 4*a***) would mimic orthologs. Obviously, however, these genes do not fit the above definition of orthology because they do not come from a single ancestral gene in the last common ancestor of the compared species. To designate such pseudoorthologs acquired from different sources, the rarely used but natural and useful term xenologs has been proposed (32, 33, 76). As with pseudoorthology caused by lineage-specific gene loss, distinguishing xenology from true orthology may be hard, if possible at all, in a pairwise genomic comparison. However, when multiple genomes are analyzed such that we can, if only roughly, identify the origin of each gene, xenologs and true orthologs become distinguishable.

A related but distinct situation transpires when a species (B) acquires via HGT a gene homologous to a gene that it already has, without the latter gene being displaced (**Figure 4*b***). The result of such an event could reasonably be described as pseudoparalogy (such that XB1 and XB2 are pseudoparalogs) because the homologous genes in species B are not paralogs under the simple definition given above: They have not evolved via gene duplication at any stage of evolution.

Xenologous gene displacement: displacement of a gene in a given lineage with a member of the same orthologous cluster from a distant lineage (xenolog)

Pseudoparalogs: homologous genes that come out as paralogs in a single-genome analysis but actually ended up in the given genome as a result of a combination of vertical inheritance and HGT

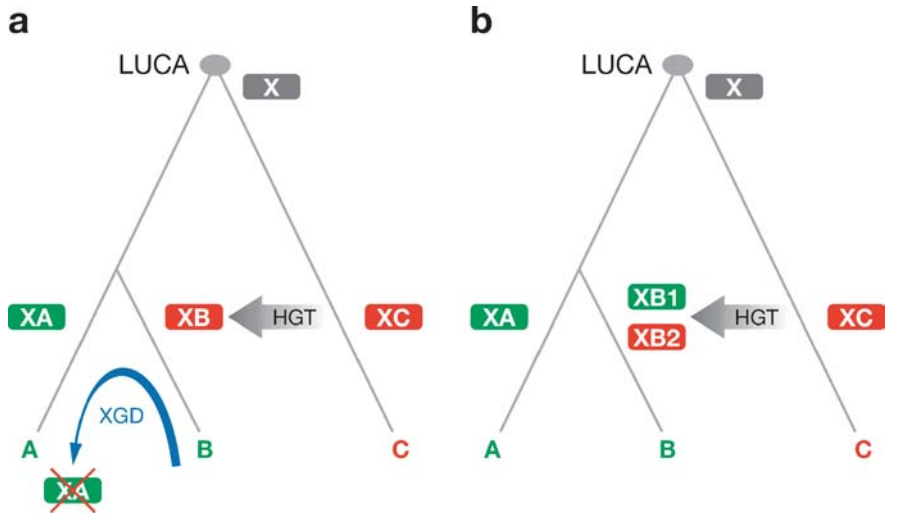


Figure 4

Effect of horizontal gene transfer on orthology and paralogy. (*a*) A hypothetical evolutionary scenario with HGT leading to xenology. (*b*) A hypothetical evolutionary scenario with HGT leading to pseudoparalogy. LUCA, Last Universal Common Ancestor (of all extant life forms).

As discussed below, there are situations, particularly those that involve endosymbiosis, where pseudoparalogy caused by HGT is common.

The concept of orthology is further complicated by another phenomenon that is common in genome evolution, gene fusion, as well as the complementary process of gene fission (54, 83, 100). The impact of such evolutionary events on orthology is that different parts (often encoding distinct domains) of genes in one species are orthologous to different genes in another species. Thus, a new type of orthologous relationship emerges whereby a gene ceases to be the atomic unit of orthology. Further implications of this shift in meaning are considered in the final section of this review.

Table 1 summarizes the meaning and area of applicability of each term introduced in this section. This system of definitions and terms is more complex than the simple dichotomy of orthology and paralogy, with subcategories of paralogs (in- and outparalogs) as well as more specialized notions, xenologs, pseudoorthologs, and pseudoparalogs, additionally introduced. The advantage, I believe, is that this system seems to be complete, i.e., includes definitions that apply to every logically imaginable case of homologous relationships between genes. We now turn to the empirical manifestations and implications of this system in evolutionary genomics.

IDENTIFICATION OF ORTHOLOGS AND PARALOGS: PRINCIPLES AND TECHNIQUES

As soon as genome comparison became a practically important task, the questions arose as to how should one delineate the set of corresponding genes (or, inaccurately, but intuitively, "the same" genes in different species), i.e., orthologs. Indeed, evolutionary classification of genes, which can only be based on the concepts of orthology and paralogy, becomes a must as soon as several genome

Table 1 Homology: terms and definitions

Homologs	**Genes sharing a common origin**
Orthologs	Genes originating from a single ancestral gene in the last common ancestor of the compared genomes.
Pseudoorthologs	Genes that actually are paralogs but appear to be orthologous due to differential, lineage-specific gene loss.
Xenologs	Homologous genes acquired via XGD by one or both of the compared species but appearing to be orthologous in pairwise genome comparisons.
Co-orthologs	Two or more genes in one lineage that are, collectively, orthologous to one or more genes in another lineage due to a lineage-specific duplication(s). Members of a co-orthologous gene set are inparalogs relative to the respective speciation event.
Paralogs	**Genes related by duplication**
Inparalogs (symparalogs)	Paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event (defined only relative to a speciation event, no absolute meaning).
Outparalogs (alloparalogs)	Paralogous genes resulting from a duplication(s) preceding a given speciation event (defined only relative to a speciation event, no absolute meaning).
Pseudoparalogs	Homologous genes that come out as paralogs in a single-genome analysis but actually ended up in the given genome as a result of a combination of vertical inheritance and HGT.

sequences become available.[1] Individual characterization of each gene in each genome rapidly turns into a gargantuan task for computational analysis and completely impractical experimentally as more genomes are sequenced (12, 30). Comparative genomics can be feasible and meaningful only if the number of distinct entities to be analyzed is substantially reduced by introducing a rational classification of genes. A natural way to do so is to delineate sets of orthologous (including co-orthologous) genes. The extent to which this is going to help in genome analysis critically depends on the nature of the relationships between genomes of different species that only can be deciphered empirically. To take one extreme, if all genes in compared genomes formed perfect clusters of one-to-one orthologs, the number of entities to study would be equal to the number of genes in each genome and would remain constant regardless of how many new genomes were sequenced. Should that be the case, the entire enterprise of comparative-evolutionary genomics would be straightforward to the point of being trivial. On the other end of the spectrum, should the number of identifiable clusters of orthologs be small compared with the total number of genes in genomes, comparative genomics would be in serious trouble.

Since orthology and paralogy are definitions that are inextricably coupled to certain types of evolutionary events (speciation and duplication, respectively), the classical scheme for identifying orthologs involves phylogenetic analysis and, in particular, the procedure generally known as tree reconciliation (20, 63, 73, 102). Under this approach, the topology of a gene tree is compared with that of the chosen species tree and the two are reconciled on the basis of the parsimony principle, by postulating the minimal possible number of duplication and gene-loss events in the evolution of the given gene. The reconciled tree is expected to reflect orthologous relationships. However, genome-wide application of this approach is effectively precluded by both fundamental and practical difficulties. The principal obstacle faced by tree reconciliation (and any other phylogenetic approach) as a strategy for ortholog identification is the prevalence of HGT, especially in prokaryotes. Strictly speaking, the wide spread of HGT invalidates the very notion of a species tree, allowing, at best, the use of various forms of consensus trees for multiple genes as surrogates of the species tree (15, 16, 98). Furthermore, even when a particular tree topology is taken as the species tree, the possibility of HGT of the analyzed gene undermines the concept of reconciliation because the topologies of the two trees are likely to be genuinely different. At a more practical level, even when HGT is not considered to be a major factor, as in the evolution of eukaryotes, both the species tree and many gene trees are fraught by uncertainties and artifacts. Even more practically, fully automated construction and analysis (with appropriate reliability tests) of trees for all genes in sequenced genomes is a major challenge for software engineering and is expensive computationally.

Further in this section, I discuss several attempts at explicit phylogenetic classification of orthologs and paralogs. However, given the substantial difficulties faced by these approaches, most genome-wide studies to date employ simplifications and shortcuts. The simple but crucial assumption that underlies such "surrogate" approaches is that the sequences of orthologous genes (proteins) are more similar to each other than they are to any other genes (proteins) from the compared

[1]Note that the first efforts to delineate sets of orthologous genes shared by relatively large genomes were undertaken years before the appearance of complete genome sequences of cellular life forms. Indeed, this was done as soon as the first pair of related genomes containing many (on the order of 100–200) genes became available, namely, when the genome of varicella-zoster virus was sequenced in 1986 and compared with the previously sequenced genome of another herpesvirus, Epstein-Barr virus (10). Subsequently, a core set of conserved (orthologous) genes was delineated for several herpesviruses (35). However, in these studies, the conceptual basis of comparative genomics was not considered explicitly and neither were the terms orthologs and paralogs used.

Pseudoorthologs: genes that actually are paralogs but appear to be orthologous due to differential, lineage-specific gene loss

genomes, i.e., they form symmetrical best hits (SymBets). Conversely, it is assumed that SymBets are most likely to be formed by orthologs, suggesting a very simple and straightforward method for identification of orthologs. For brevity, I call these two assumptions taken together the SymBet hypothesis. It seems highly plausible, based on the definition of orthology and the notion that orthologs typically occupy the same functional niche, that the SymBet hypothesis is true, at least statistically. Consider the alternative: A gene from one genome is most similar not to its ortholog but to a paralog from another genome. As shown in **Figure 5*a***, this requires a substantial difference in the rates of evolution of paralogs sufficient to offset the divergence of paralogs prior to speciation such that, using the notation of **Figure 5*a***, $AS_x > SD_x + SD_y + BS_y$ (A and B are two species; S stands for speciation and D for duplication; AS_x and BS_y are the amounts of divergence accumulated, respectively, by the genes XA and YB after speciation; SD_x and SD_y are the amounts of divergence accumulated by the paralogous genes x and y after duplication but prior to the speciation). It is generally unlikely that one paralog evolves so much faster than another, given that paralogs retain similar functions. Although asymmetry in the evolution of paralogs has been detected in some studies, typically, it is relatively small. Furthermore, even should that be the case, only the first assumption of the SymBet hypothesis will be violated. In other words, if one paralog evolves much faster than the other, this could lead to a false negative under the SymBet method for ortholog detection (a pair of orthologs missed) but not to a false positive (no erroneous detection of orthologs). Lineage-specific gene duplications producing inparalogs are likely to be a more common cause of violation of the first assumption of the SymBet hypothesis; these duplications may obscure orthologous relationships leading to false negatives (**Figure 5*b***). It seems that the only realistic situations when the second assumption of the SymBet hypothesis can be violated are the cases of pseudoorthology and xenology (**Figures 3** and **4**). Pseudoorthologs and xenologs will likely form a SymBet and produce a false positive under this approach to orthology detection. However, even though formally, neither pseudoorthologs nor xenologs are orthologs, there is a subtle difference between these two situations. Unlike pseudoorthologs, which never can be considered orthologous given their origin by duplication, xenologs come across as orthologs in comparisons of genomes from the donor and recipient lineages (i.e., XA is the ortholog of XC in **Figure 4**). Accordingly, xenology is likely to be a more reliable predictor of gene function that pseudoorthology (see discussion below).

Given these uncertainties, empirical results on the prevalence of SymBets in genome comparisons are important to assess the level of one-to-one orthology between genomes that is critical for evolutionary and functional genomics. **Table 2** shows the number of SymBets between prokaryotic genomes separated by varying evolutionary distances. These results clearly demonstrate that a one-to-one orthologous relationship is a major rather than a minor pattern in genome evolution. For relatively closely related species, e.g., different γ-Proteobacteria, the fraction of probable one-to-one orthologs identified as SymBets typically is >0.5. Predictably, the fraction of genes that produce SymBets drops with evolutionary distance but remains substantial (20%–30% of the genes) even between bacteria and archaea. This fraction also depends on the total number of genes in a genome such that small genomes with few paralogs (e.g., *Mycoplasma genitalium*) show a high level of one-to-one orthology even with distant species. Detection of SymBets is arguably the simplest method for the identification of probable orthologs that is most suitable for closely related genomes but also serves well at greater evolutionary distances for the specific purpose of detecting one-to-one orthologs.

The demonstration that numerous genes in sequenced genomes produced SymBets

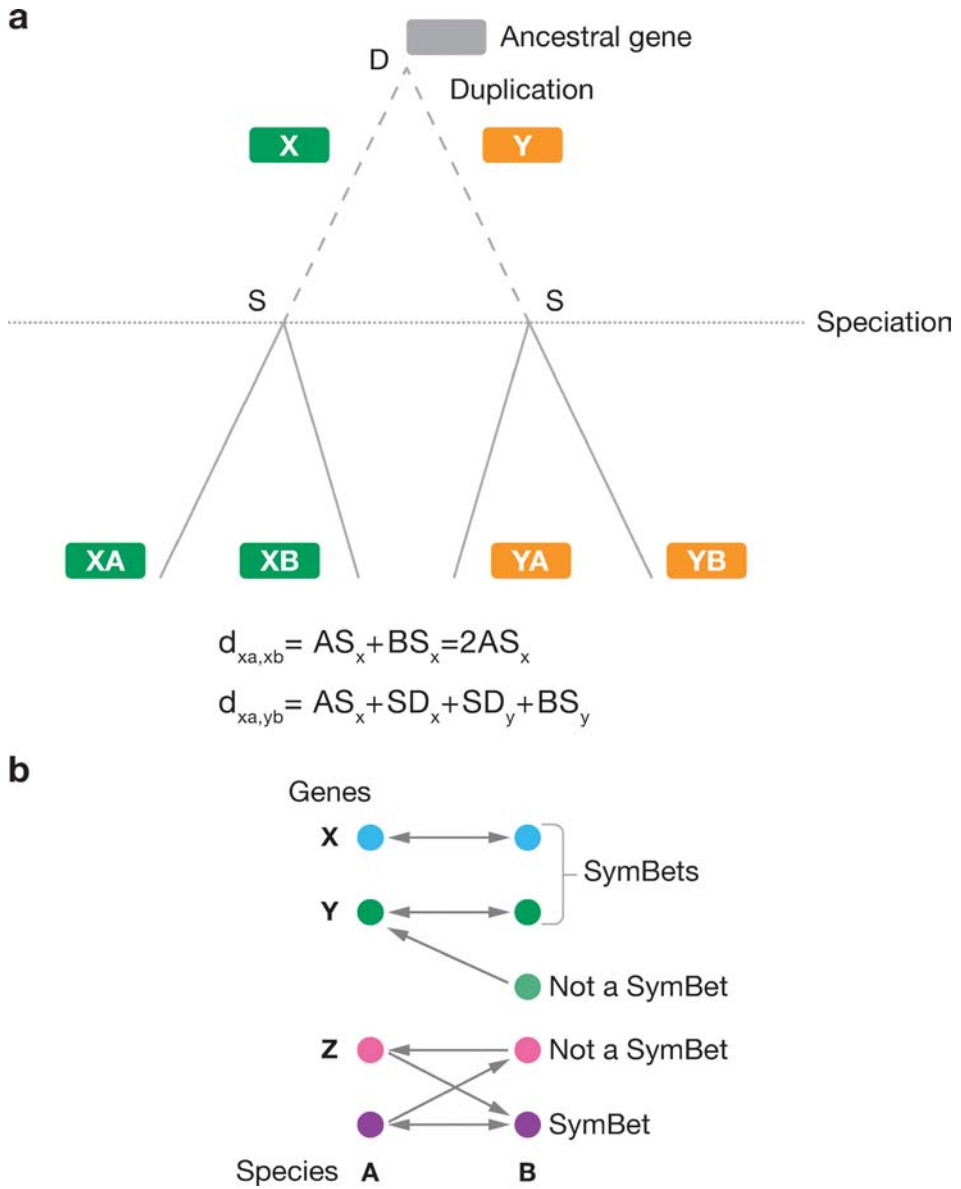


Figure 5

Orthology and genome-specific best hits. (*A*) An evolutionary scheme illustrating the connection between orthology and symmetrical best hits (SymBets). X and Y represent two paralogous genes. The branch lengths in the tree are taken to reflect evolutionary distances between the compared genes, and the formulas for the distances between orthologs and paralogs are given. A molecular clock is assumed for the evolution of orthologs but not paralogs. A and B are two species; D indicates a duplication event and S indicates a speciation event; AS_x and BS_y are the amounts of divergence accumulated, respectively, by the genes XA and YB after speciation; SD_x and SD_y are the amounts of divergence accumulated by the paralogous genes x and y after duplication but prior to the speciation. (*B*) An evolutionary scheme illustrating violation of the SymBet relationship caused by a lineage-specific duplication. Arrows designate best hits; circles of similar shades represent inparalogs. X, Y, and Z designate three cases of (co)orthologous relationships: one-to-one (X), one-to-many (Y) and many-to-many (Z).

Table 2 Symmetrical best hits between selected prokaryotic genomes[a]

	Ec	Yp	Hp	Bs	Mg	Aa	Tm	Mj	Ma	Ta
Ec-4289		0.584	0.456	0.305	0.527	0.519	0.428	0.24	0.151	0.308
Yp-4083	2385		0.432	0.28	0.525	0.496	0.423	0.218	0.144	0.275
Hp-1566	714	677		0.403	0.442	0.396	0.321	0.181	0.211	0.176
Bs-4100	1251	1144	631		0.648	0.495	0.465	0.239	0.153	0.306
Mg-480	253	252	212	311		0.469	0.515	0.235	0.281	0.238
Aa-1553	806	771	615	768	225		0.449	0.279	0.33	0.256
Tm–1846	808	780	503	858	247	697		0.245	0.294	0.265
Mj-1770	425	385	284	423	113	434	433		0.489	0.362
Ma-4540	649	589	330	627	135	513	543	866		0.415
Ta-1478	455	406	260	453	114	378	392	535	614	

[a]The bottom half of the table shows the number of SymBets for each pair of genomes and the upper half shows the fraction of proteins in the smaller genome that form SymBets with the putative orthologs from the larger genome. Species abbreviations are as follows. Proteobacteria: Ec, *Escherichia coli*; Yp, *Yersinia pestis*; Hp, *Helicobacter pylori:* Gram-positive bacteria: Bs, *Bacillus subtilis*; Mg, *Mycoplasma genitalium*: deep-branching, hyperthermophilic bacteria: Aa, *Aquifex aeolicus*; Tm, *Thermotoga maritima*; archaea: Mj, *Methanocaldococcus jannaschii*; Ta, *Thermoplasma acidophilum*; Ma, *Methanosarcina acetivorans*. For each species, the total number of protein-coding genes is indicated.

even between relatively distant species (90) made it clear that construction of a genome-wide evolutionary classification of orthologous and paralogous genes was a feasible task even if the SymBet approach itself is insufficient for this purpose owing to the prevalence of inparalogs. To my knowledge, such a classification was first implemented in the system of the so-called Clusters of Orthologous Groups of proteins (COGs) (89). The phrase "orthologous groups," which has been subsequently criticized as "terminology muddle" (71), was intended to emphasize that the system captured not only one-to-one orthologous relationships but also coorthologous relationships between inparalogs. The idea behind the COG approach was to generalize and extend the notion of a genome-specific best hit. First, the requirement for the reciprocity of best hits (as in SymBets) was abandoned because of the in-paralog problem (see **Figure 5*b***). Second, the notion of a genome-specific best hit was extended to multiple genomes such that the algorithm sought to identify clusters of consistent best hits. More specifically, the COG construction procedure is based on the assumption that any set of at least three proteins from relatively distant genomes that are more similar to each other than they are to any other proteins from the same genomes are most likely bona fide orthologs. This prediction holds even if sequence similarity between some of the compared proteins is relatively low and, accordingly, even fast-evolving genes can be incorporated into the COGs. Briefly, the procedure for COG construction consists of the following steps. (*i*) An all-against-all comparison of protein sequences encoded in multiple genomes (typically using the BLAST program). (*ii*) Detection and clustering of obvious inparalogs, i.e., proteins from the same genome that are more similar to each other than they are to any proteins from other species. (*iii*) Identification of triangles of mutually consistent, genome-specific best hits such that clusters of inparalogs detected at step 2 are treated as single entities. (*iv*) Merging triangles with a common side to form COGs.

The COG approach neatly delineates clusters of probable orthologs that include inparalogs in relatively few lineages. However, the procedure tends to err toward overlumping in the case of large protein families that include a complex mix of in- and outparalogs. Additional complications emerge in the case of

multidomain proteins that also may artificially bridge unrelated COGs. Several other approaches for identification of orthologs, based on either specially designed clustering procedures or on explicit phylogenetic analysis, have been developed to overcome these problems and better disentangle orthologs and paralogs. In particular, the INPARANOID procedure developed by Sonnhammer and coworkers identifies clusters of orthologs, including co-orthologous sets of inparalogs, for pairs of genomes, by first detecting Sym-Bets and then incorporating additional inparalogs according to developed statistical criteria (67, 82). High accuracy of identification of inparalogs seems to be achievable with this approach, but the inability to handle multiple genomes simultaneously is a serious limitation. Another method for ortholog detection developed by the same group involves comparison of gene trees with species trees, with the goal of direct identification of orthologs (86). The parts of the gene tree that have the same topology as the species tree are inferred to include orthologs. In principle, this and similar phylogenomic [i.e., applying phylogenetic analysis on genome scale (19)] methods are supposed to provide the strongest and most direct evidence of orthology. A fundamental drawback is, however, the uncertainty of the species tree in the case of prokaryotes due to the prevalence of HGT (this does not appear to be a problem in the case of eukaryotes). In addition, the method is computationally expensive and sensitive to tree construction artifacts. Nevertheless, this approach has been applied to the eukaryotic subset of the Pfam database of protein families, yielding numerous (inferred) orthologous domains (87). A very similar automated phylogenomic procedure for inference of orthologs has been developed by Zmasek & Eddy (103). In a more recent development, a Bayesian probabilistic technique has been introduced to assign probabilities to the orthology identifications (3). A major effort to identify orthologs and paralogs has been undertaken by Perrière and coworkers who constructed databases of homologous genes from vertebrates (HOVERGEN) and bacteria (HOBACGEN) in which families of homologs (each consisting of a mix of orthologs and paralogs) are accompanied by phylogenetic trees (18, 79). Very recently, tools have been developed for tree reconciliation in the framework of the database, with the goal of identifying sets of orthologs (17). The effectiveness of these methods on genome scale remains to be assessed.

In summary, although phylogenomic methods, in principle, should be best suited for deciphering orthologous and paralogous relationships, in practice, these approaches so far have not matured enough to produce a comprehensive collection of orthologous-paralogous clusters covering multiple species. Such collections have been constructed only with methods based on sequence similarity and the notion of genome-specific best hits. Clusters produced with these approaches are by no means error-free, in particular with respect to lumping some of the orthologous gene sets into inflated, mixed clusters of orthologs and paralogs. However, extensive work on genome annotation as well as genome-wide evolutionary studies performed with the help of these systems (50) suggest that they are sufficiently robust for extracting meaningful evolutionary and functional patterns (discussed below).

EVOLUTIONARY PATTERNS OF ORTHOLOGY AND PARALOGY

Coverage of Genomes in Clusters of Orthologs

Probably, the aspect of orthologous clusters that is of the most immediate importance to evolutionary and functional genomics is the coverage of genomes, i.e., the fraction of genes with orthologs in other species. A substantial majority of genes from each sequenced prokaryotic genome (**Figure 6*a***) and a somewhat lower fraction of eukaryotic genes (**Figure 6*b***) belong to COGs

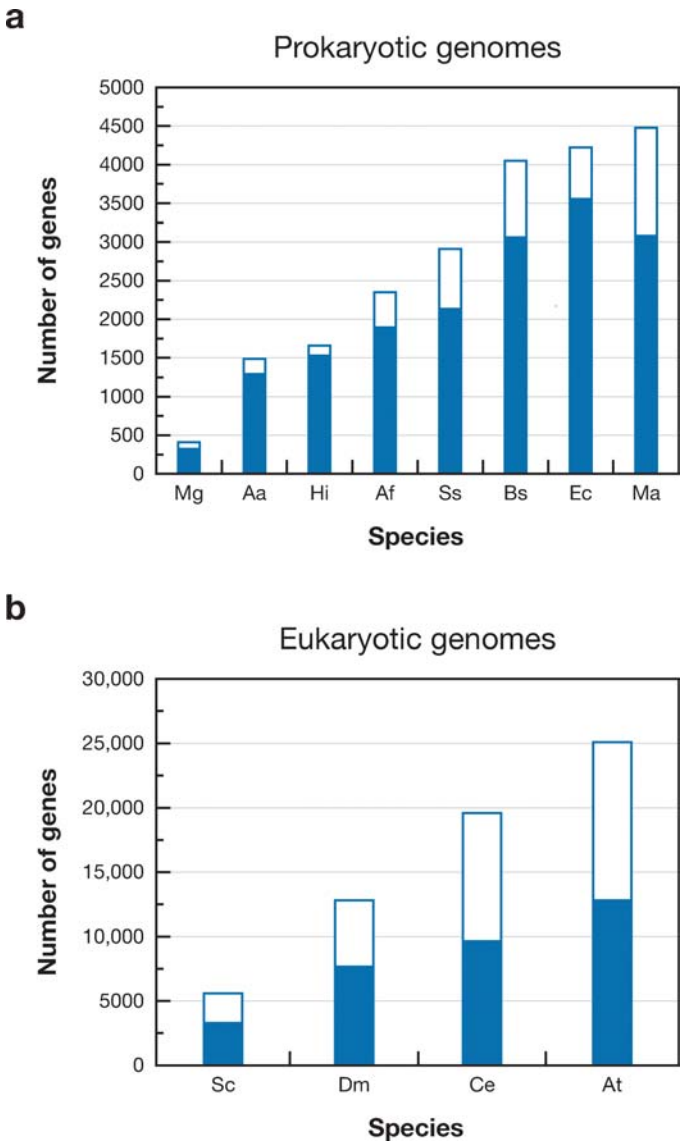


Figure 6

Coverage of selected genomes with clusters of orthologous groups of proteins (C/KOGs). (*a*) Prokaryotic genomes. (*b*) Eukaryotic genomes. The data are from (88). Filled volume, genes in C/KOGs; empty volume, genes not included in C/KOGs. Abbreviations: Bacteria: Aa, *Aquifex aeolicus*; Bs, *B. subtilis*; Ec, *E. coli*; Hi, *Haemophilus influenzae*; Mg, *Mycoplasma genitalium*. Archaea: Af, *Archaeoglobus fulgidus*; Ma, *Methanosarcina acetivorans*; Ss, *Sulfolobus solfataricus*. Eukaryotes: At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Sc, *Saccharomyces cerevisiae*. Data for *Homo sapiens* are not shown because the KOGs include an early, inflated version of the human gene set.

(or the clusters of orthologous genes from eukaryotes dubbed KOGs) (49, 88). The coverage of genomes with COGs slowly increases with the growing number of included species (91). It remains unclear whether the level of orthology tends to 100% when the number of genomes tends to infinity or there is a lower limit, and some genes are true, species-specific "orphans" evolving in a regime different from the majority of the genes (29, 68). Obviously, however, the orthology level is high and will only increase with continued genome sequencing. Therefore, the reduction of search space provided by the classification of genes into orthologous clusters is substantial and, in practical terms, should be sufficient to cope with the flood of information produced by genome sequencing.

One-to-One Orthologs and Inparalogs

As discussed above, lineage-specific duplications leading to the emergence of inparalogs complicate orthologous relationships between genes. A simple analysis of the COGs allows one to evaluate the extent of this phenomenon, with the caveat that some lumping is involved, leading to an inflated estimate of the number of inparalogs. **Figure 7** shows the distribution of the number of paralogs in COGs for four prokaryotic genomes. For all the importance of lineage-specific expansion of paralogous families, in each genome the majority of orthologous lineages (COGs) are represented by a single gene. Specifically, in *Escherichia coli*, a complex bacterium with a relatively large genome, ~71% of the COGs include a single gene, and in the case of *M. genitalium*, a bacterium with a near-minimal genome, such COGs form the overwhelming majority (~92%). This conclusion is compatible with the earlier quantitative analysis of lineage-specific expansions in prokaryotes, which detected only a few large expansions in each genome (41), and with independent analyses of the size distribution of paralogous

families (38, 53). A qualitatively similar pattern, albeit with some predictable excess of inparalogs, was observed for eukaryotic orthologous clusters (data not shown). Moreover, 1769 of the 4873 COGs (36%) contain exactly one representative from each of the included genomes. It has been proposed that one-to-one orthology could be selected for in the case of genes encoding subunits of macromolecular complexes requiring strict stoichiometry due to the deleterious effect of subunit imbalance (75, 94). Indeed, orthologous sets containing no paralogs appear to be significantly enriched in complex subunits (49, 75, 95).

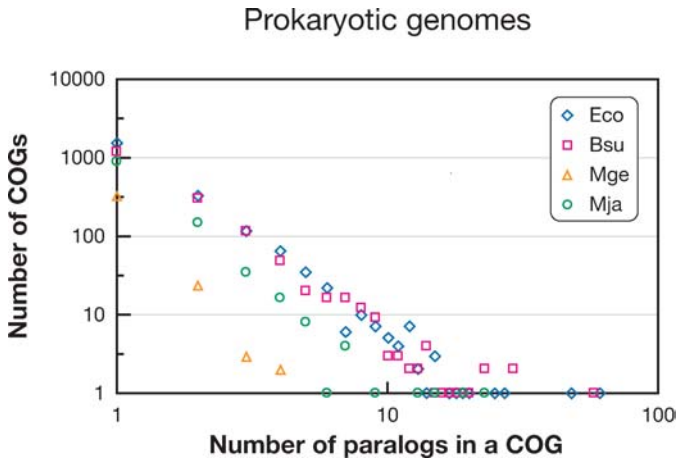


Figure 7

Distribution of the number of paralogs in COGs for selected prokaryotic genomes. The data were extracted from the current COG version (88). The plot is shown in the double-logarithmic scale.

Orthologous Clusters and the Molecular Clock

A central tenet of Kimura's neutral theory is that the rate of evolution of a gene remains the same (with some dispersion, obviously) as long as the biological function does not change (44). Kimura never used the term ortholog (or paralog), but this generalization obviously applies to orthologous gene sets, primarily those that include no inparalogs. The subsequent evolution of the molecular clock concept involved considerable dispute, with numerous studies demonstrating substantial overdispersion of the clock (5, 6, 31). Genome-wide tests of the molecular clock have been conducted only very recently. One approach involved comparing the evolutionary distances within a COG containing no inparalogs to a standard intergenomic distance, which was defined as the median of the distribution of the distances between all one-to-one orthologs in the respective genomes (66). Under the molecular clock model, the points on a plot of intergenic distances for the given COG versus intergenomic distances will scatter around a straight line. A statistical method was developed to identify significant deviations from the clock-like behavior. Among several hundred COGs representing three well-characterized bacterial lineages, α-Proteobacteria, γ-Proteobacteria, and the *Bacillus-Clostridium* group, the clock hypothesis could not be rejected for ~70%, whereas the rest showed substantial deviations. These anomalies could be explained either by lineage-specific acceleration of evolution or by XGD (see below). The general conclusion from these analyses seems to be that the majority of orthologous genes evolve in the clock-like mode as long as there was no duplication, although the frequency of exceptions was by no means negligible.

Molecular clock: a central concept of molecular evolution, which posits that a gene evolves at a constant rate as long as its function does not change

The connections between gene duplication and evolutionary rates have been explored in considerable detail and appear to be quite complex. Ohno's original idea was that, immediately after duplication, one of the newborn paralogs is freed from purifying selection and would evolve rapidly such that it either perishes or, on relatively rare occasions, evolves a new function (neofunctionalization), after which evolution slows down again (69). Subsequent theoretical and empirical analyses have shown that this path of evolution, while apparently realized on some occasions, is probably not the most common outcome of a gene duplication (61). What happens more often seems to be relaxation of purifying selection immediately after duplication, resulting in accelerated evolution in both paralogs. This is thought to reflect subfunctionalization, i.e.,

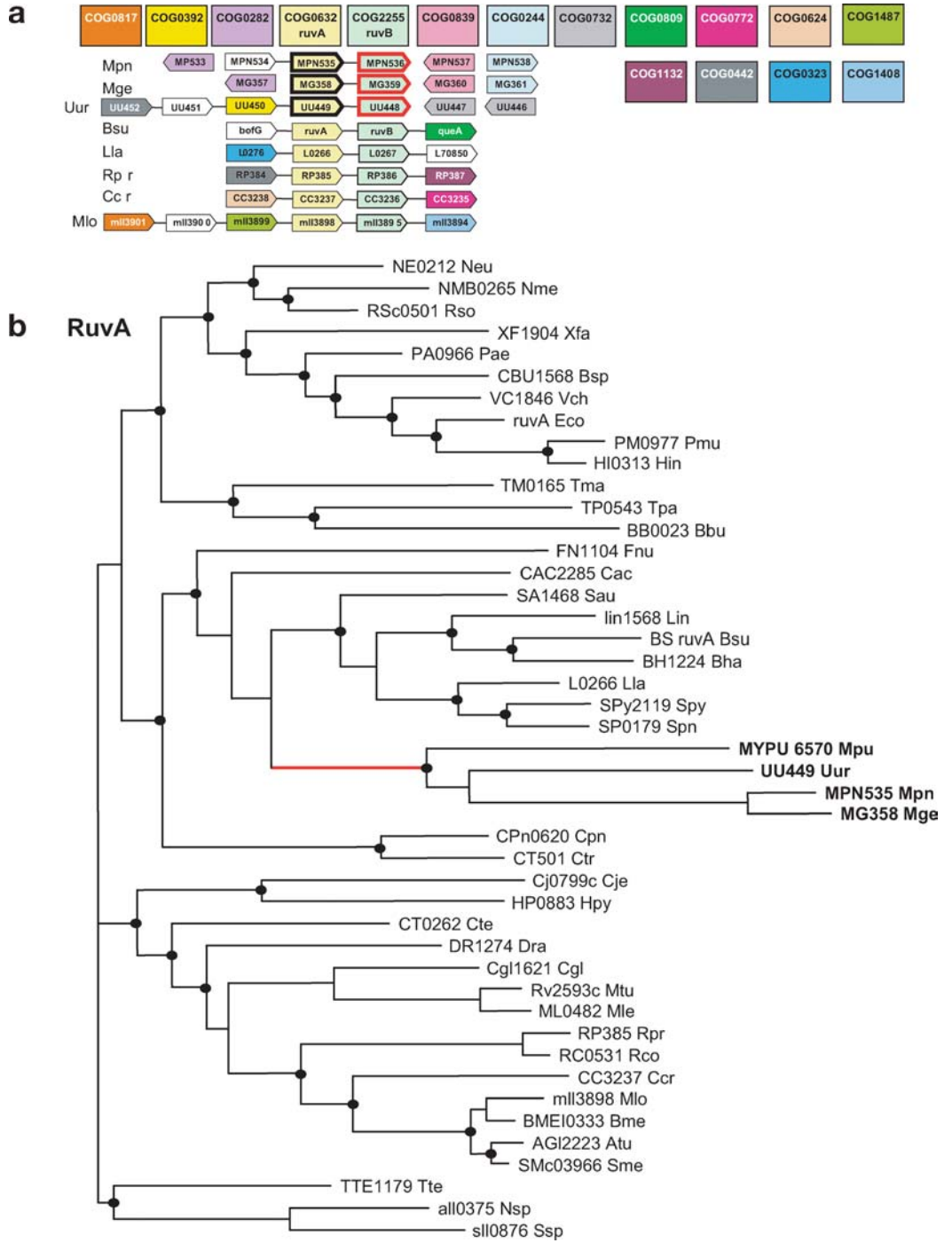
partitioning of the different functions of the multifunctional ancestral gene between paralogs (45, 59, 60). Somewhat paradoxically, however, two independent recent studies have shown that despite this acceleration, genes that have paralogs on average evolve slower than those that do not (9, 42). This difference may be due to a greater likelihood of fixation of emergent paralogs among slowly evolving (more "important") genes.

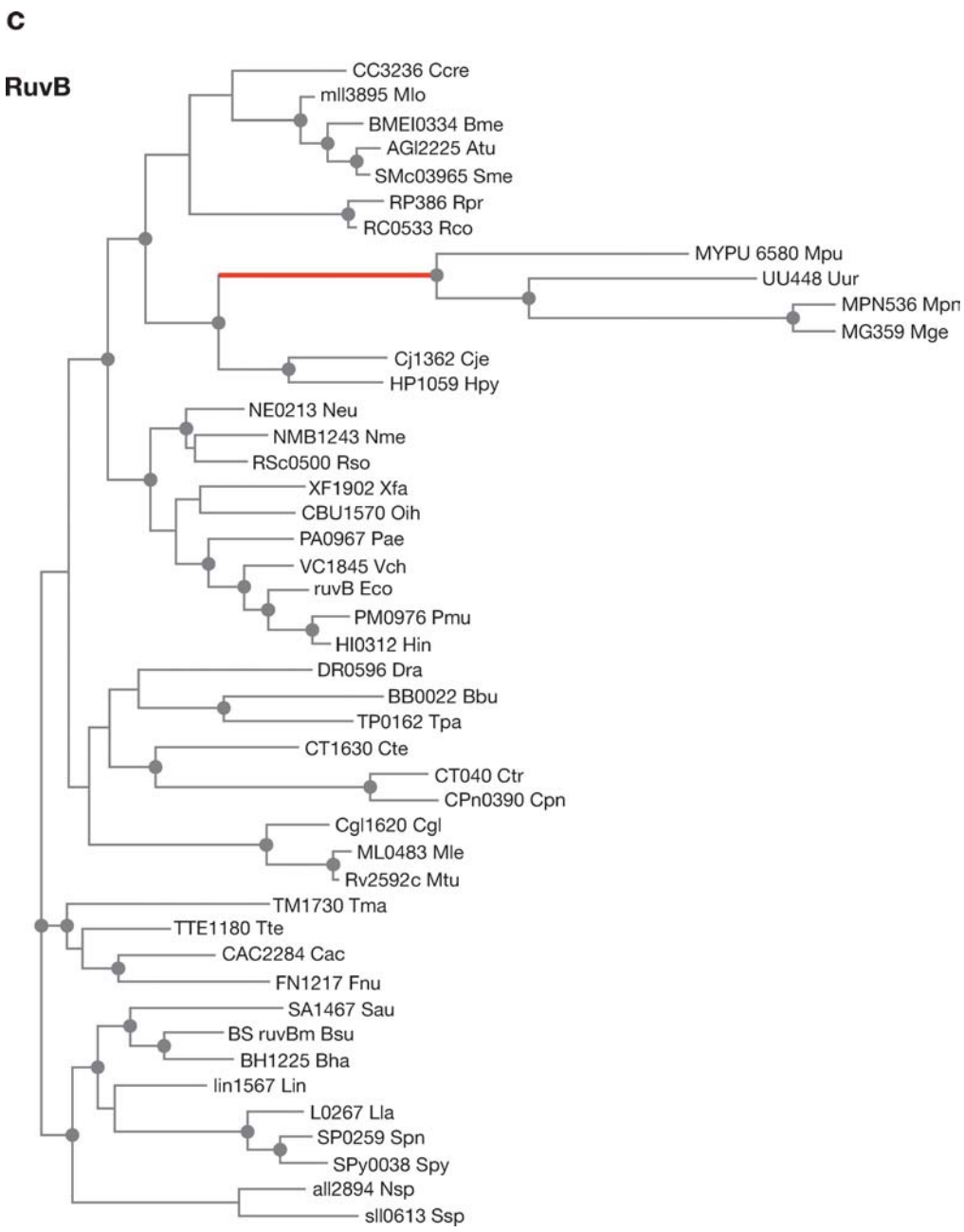
Xenologs, Pseudoorthologs, and Pseudoparalogs

As discussed above, xenologs are homologs that violate the definition of orthology due to HGT. More specifically, xenology is brought about either by XGD or by acquisition of a gene that is new for the given lineage (51). When the study of the clock-like behavior of orthologous gene sets discussed in the previous section was followed up with phylogenetic analysis of deviant cases, probable XGD was demonstrated for 10%–20% of the COGs within each of the examined bacterial taxa, establishing XGD as a major evolutionary phenomenon (66). On some occasions, XGD even takes the form of displacement in situ whereby a gene is displaced with a horizontally transferred distant ortholog without disrupting the operon structure (70). **Figure 8** illustrates one such case: the displacement of the *ruvB* gene (coding for the helicase subunit of the resolvasome) in the mycoplasmas with the ortholog from ε-Proteobacteria. In this case, the *ruvB* genes of *Mycoplasma* and those of the rest of low GC gram-positive bacteria, the taxon to which *Mycoplasma* belong, qualify as xenologs given their different phylogenetic affinities (**Figure 8*c***). A clear example of acquisition of a new gene leading to xenology is the B family DNA polymerase of γ-Proteobacteria (e.g., the *polB* gene of *E. coli*). At first sight, this gene appears to be an ortholog of the archaeal and eukaryotic B family polymerases (see COG0417). However, these

Figure 8

Xenologous displacement in situ of the *ruvB* gene in the mycoplasmas. (*A*) Organization of the Holliday junction resolvasome operon and surrounding genes in bacteria. COG0632, Holliday junction resolvasome, DNA-binding subunit; COG2255, Holliday junction resolvasome, DNA-binding subunit; COG0817; Holliday junction resolvasome, endonuclease subunit; COG0392, predicted integral membrane protein; COG0282, acetate kinase; COG0839, NADH:ubiquinone oxidoreductase subunit 6 (chain J); COG0244, ribosomal protein L10; COG0732, restriction endonuclease S subunits; COG0809, S-adenosylmethionine:tRNA-ribosyltransferase-isomerase; COG0772, bacterial cell division membrane protein; COG0624, acetylornithine deacetylase/succinyl-diaminopimelate desuccinylase and related deacylases; COG1487, predicted nucleic acid-binding protein; COG1132, ABC-type multidrug transport system, ATPase, and permease components; COG0442, prolyl-tRNA synthetase; COG0323, DNA mismatch repair enzyme; COG1408, predicted phosphohydrolases. (*B*) Unrooted phylogenetic tree for RuvA. (*C*) Unrooted phylogenetic tree for RuvB. Branches supported by bootstrap probability >70% are marked by black circles. Names of the genes from mosaic operons and the respective branches are shown in red. Species abbreviations: Atu, *Agrobacterium tumefaciens*; Bha, *Bacillus halodurans*; Bsu, *Bacillus subtilis*; Bbu, *Borrelia burgdorferi*; Bme, *Brucella melitensis*; Cje, *Campylobacter jejuni* Ccr, *Caulobacter crescentus*; Ctr, *Chlamydia trachomatis*; Cpn, *Chlamydophila pneumoniae*; Cte, *Chlorobium tepidum*; Cac, *Clostridium acetobutylicum*; Cgl, *Corynebacterium glutamicum*; Eco, *Escherichia coli*; Fnu, *Fusobacterium nucleatum*; Hin, *Haemophilus influenzae*; Hpy, *Helicobacter pylori*; Lla, *Lactococcus lactis*; Lpl, *Lactobacillus plantarum*; Lin, *Listeria innocua*; Neu, *Nitrosomonas europaea*; Mlo, *Mesorhizobium loti*; Mge, *Mycoplasma genitalium*; Mpn, *Mycoplasma pneumoniae*; Mpu, *Mycoplasma pulmonis*; Mle, *Mycobacterium leprae*; Mtu, *Mycobacterium tuberculosis*; Nme, *Neisseria meningitidis*; Nsp, *Nostoc* sp.; Oih, *Oceanobacillus ibeyensis*; Pae, *Pseudomonas aeruginosa*; Rso, *Ralstonia solanacearum*; Rpr, *Rickettsia prowazekii*; Rco, *Rickettsia conorii*; Sme, *Sinorhizobium meliloti*; Sau, *Staphylococcus aureus*; Spy, *Streptococcus pyogenes*; Ssp, *Synechocystis* PCC6803; Tma, *Thermotoga maritima*; Tte, *Thermus thermophilus*; Tpa, *Treponema pallidum*; Vch, *Vibrio cholerae*; Xfa, *Xylella fastidiosa*; Uur, *Ureaplasma urealyticum*. The figure is from (70) where the details of phylogenetic analysis are described.





genes should be classified as xenologs because the proteobacterial version clearly does not derive from the last universal common ancestor (which is also the last common ancestor of bacteria and archaea) but rather had been acquired via HGT.

As defined above, pseudoorthologs emerge via lineage-specific, differential loss of paralogous genes (**Figure 2**). A systematic search for pseudoorthologous genes requires detailed, genome-wide phylogenetic analysis, which to my knowledge has not yet been

conducted. Nevertheless, some likely cases of pseudoorthology can be gleaned by examination of COGs. Consider COG0114 (fumarase) and COG1027 (aspartate ammonia-lyase), which consist of paralogous enzymes with a high level of sequence similarity to each other. Both enzymes are widespread in bacteria and, most likely, were already present in the last common ancestor of all bacteria but apparently have been lost independently in many lineages. When comparing the genomes of two cyanobacteria, *Synechocystis* sp. and *Nostoc* sp., the genes slr0018 of the former and alr3724 of the latter, produce a SymBet and, accordingly, could be identified as orthologs by default. However, inspection of the COGs clearly shows that slr0018 is a fumarase whereas alr3724 is an aspartate ammonia-lyase. This seems to be a clear-cut case of pseudoorthology caused by lineage-specific, differential loss of paralogs, with the ensuing functional differences (see below).

The most obvious case of pseudoparalogy is the presence of numerous pairs of homologous genes of ancestral and endosymbiotic (mitochondrial or, in plants, chloroplastic) origin in eukaryotes (34, 43, 56). These pseudoparalogs are particularly abundant among the components of the translation machinery, such as ribosomal proteins or aminoacyl-tRNA synthetases. Many additional pseudoparalogs appear to have emerged through other routes of HGT. One of the most conspicuous is the transfer of archaeal genes to bacteria, particularly hyperthermophiles. **Figure 9** shows the phylogenetic tree for the peroxiredoxin AhpC (COG0450). This tree includes two paralogous proteins from the hyperthermophilic bacterium *Aquifex aeolicus*, one of which clusters with archaeal and the other with bacterial homologs. The respective genes are pseudoparalogs because they apparently ended up in the *A. aeolicus* genome as a result not of gene duplication at any stage of evolution but of horizontal transfer of one of the peroxiredoxin genes from an archaeal source. Conversely, the tree includes three peroxiredoxins from the archaea *Thermoplasma acidophilum* and *T. volcanium*, at least two of which (the one nested within the archaeal subtree and the one with a clear bacterial affinity) are pseudoparalogs. Notably, the only peroxiredoxin of another hyperthermophilic bacterium, *Thermotoga maritima*, shows an archaeal affinity, suggesting that in this lineage, the original bacterial gene had been lost, probably after the acquisition of the archaeal version.

Protein Domain Rearrangements, Gene Fusions/Fissions, and Orthology

Orthologous protein in eukaryotes sometimes differ in their domain architectures. There seems to be a trend toward an increase in the complexity of domain architecture in parallel with the increase of organismal complexity, a phenomenon dubbed domain accretion (48, 49). Apparently, additional domains acquired by proteins from more complex organisms provide additional interactions leading, in particular, to increased complexity of signal transduction and various regulatory processes. Differences in domain architectures can also be detected between orthologs from major prokaryotic taxa; one such case is illustrated in **Figure 10*a***. The DnaG-like primases of bacteria and archaea share a highly conserved catalytic domain and appear to be orthologous, especially given that they are represented by a single protein in all bacterial and archaeal genomes (62). However, the bacterial and archaeal orthologs have different accessory domains, a Zn-finger and a distinct module of a helicase domain, respectively (**Figure 10*a***), which may reflect substantial functional differences (see next section).

As mentioned above, gene fusions and fissions, which are common in genome evolution, affect the very notion of orthology: In this case, a single gene in some species coding for a multidomain protein is orthologous to two or more distinct genes coding for the respective individual domains in another set

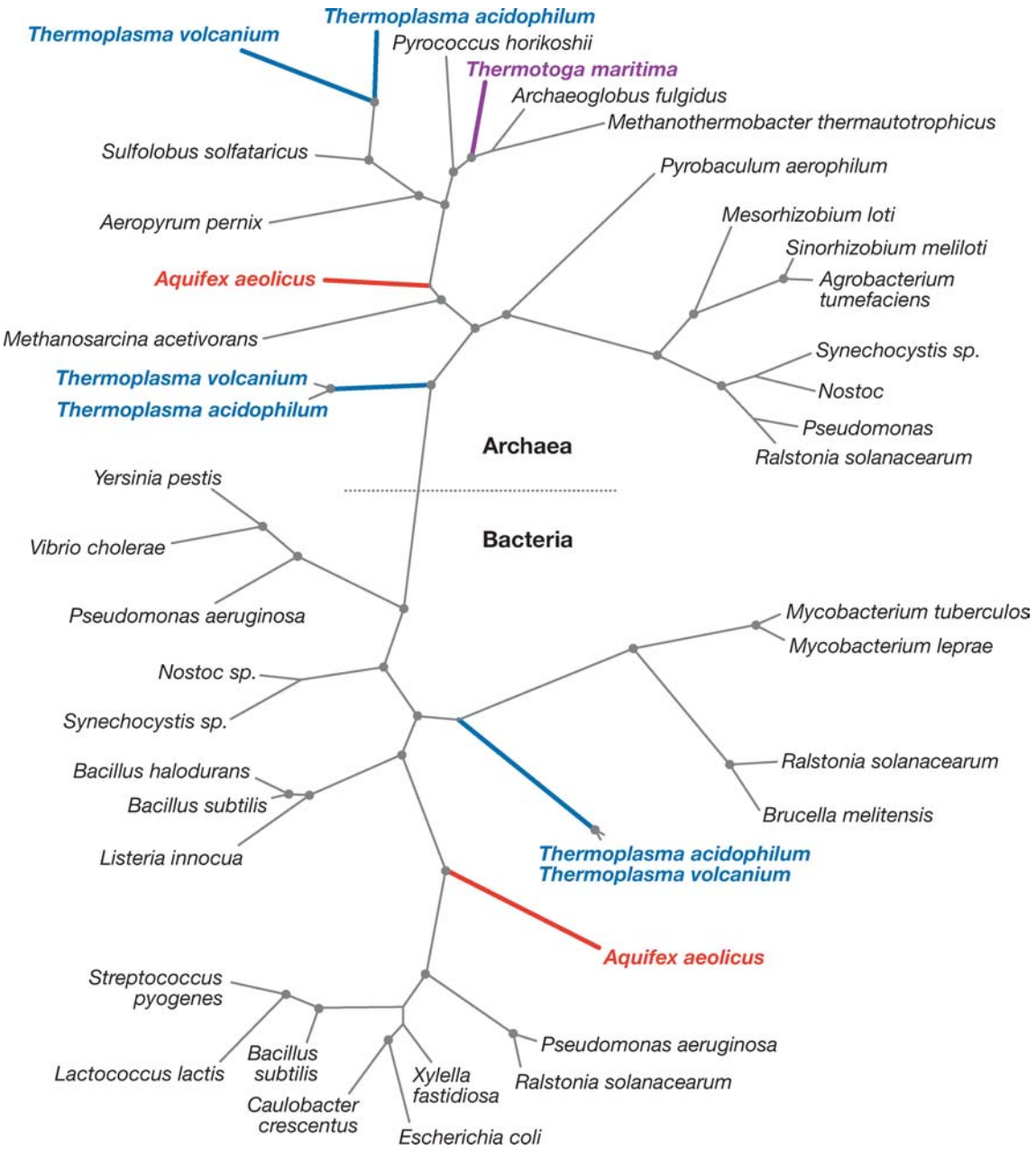


Figure 9

Horizontal gene transfer leading to pseudoparalogy. The two pseudoparalogous peroxiredoxins from *Aquifex aeolicus* are shown in red, the three pseudoparalogs from the Thermoplasmas in blue, and the only peroxiredoxin of *Thermotoga maritima* in purple. The genes are identified by full species names. The maximum likelihood, unrooted phylogenetic tree was constructed as previously described (70).

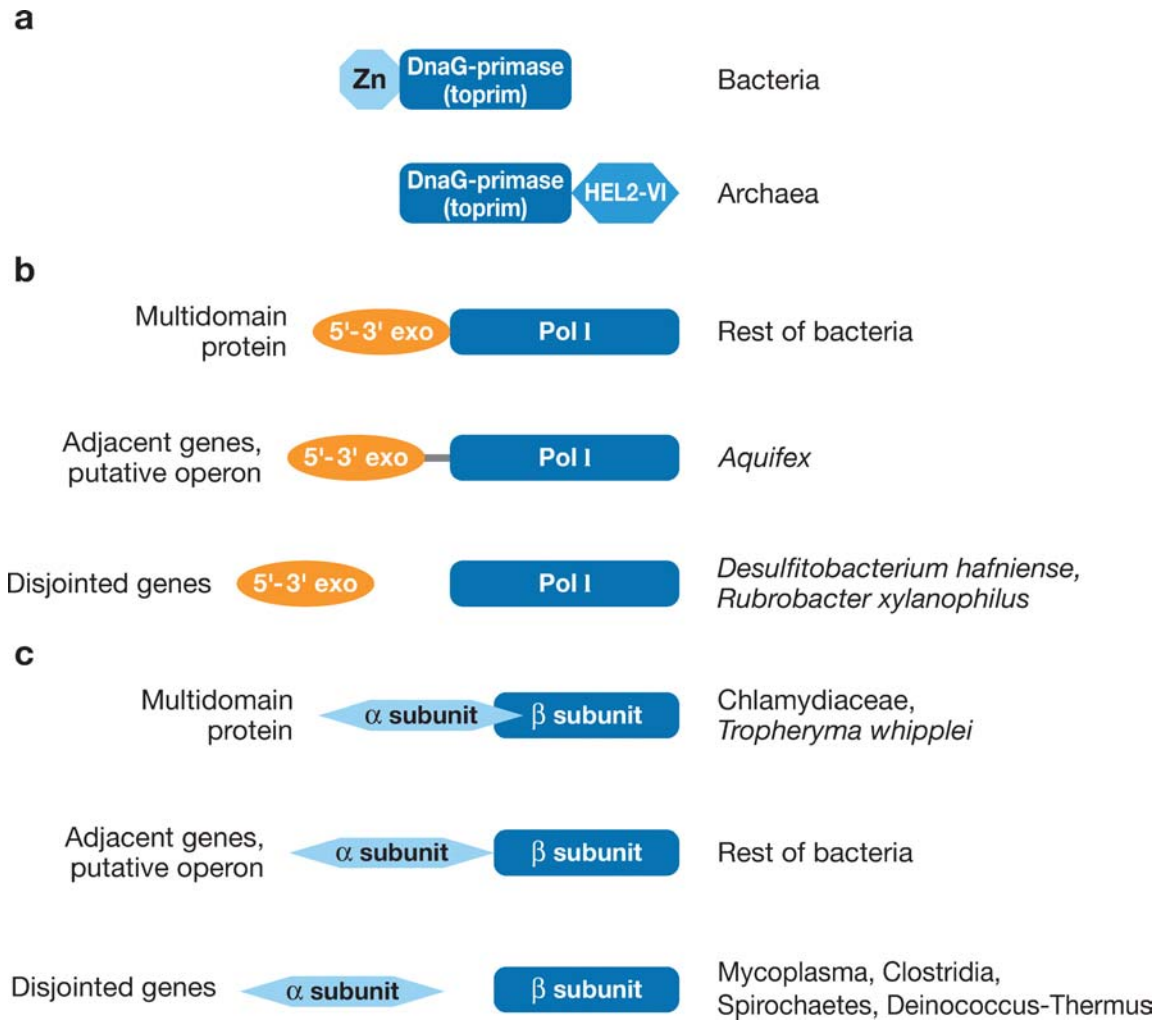


Figure 10

Rearrangements of gene structure and orthology. (*a*) Domain architectures of bacterial and archaeal DnaG-like primases. (*b*) Independent fission of the DNA polymerase I gene in multiple bacterial lineages. (*c*) Fusion of the genes for bacterial glycyl-tRNA synthetase subunits.

of species. **Figure 10 (*b* and *c*)** shows two cases of such relationships. In the example in **Figure 10*b***, the bacteria *A. aeolicus*, *A. pyrophilus*, *Desulfitobacterium hafniense*, and *Rubrobacter xylanophilus* encode the polymerase and 5′–3′ exonuclease domains of DNA polymerase I in two distinct genes, unlike all other bacteria in which these enzymes are domains of a single, multidomain protein. The bacteria in which the nuclease and polymerase activities reside in different proteins belong to three distinct lineages, suggesting three independent fissions of the *polA* gene. Whereas the genes for the nuclease and the polymerase are adjacent in the two *Aquifex* species, in the other two bacteria they are not, implying genome arrangement subsequent to gene fission. By contrast, the example in **Figure 10*c*** shows the fusion of the genes for the α and β subunits of glycyl-tRNA synthetases in the parasitic bacteria of the *Chlamydia*e branch and the pathogenic

actinobacterium *Tropheryma whipplei*. In this case, it appears likely that gene fusion occurred only once, with subsequent horizontal dissemination of the fused gene.

FUNCTIONAL CORRELATES OF ORTHOLOGY AND PARALOGY

The validity of the conjecture on functional equivalency of orthologs is crucial for reliable annotation of newly sequenced genomes and, more generally, for the progress of functional genomics. The huge majority of genes in the sequenced genomes will never be studied experimentally, so for most genomes transfer of functional information between orthologs is the only means of detailed functional characterization. To what extent is such transfer legitimate? A rough estimate can be obtained by comparing the available functional information for experimentally characterized one-to-one orthologs from model organisms. Inspection of the 1330 COGs that contain one-to-one orthologs from the well-studied bacteria *E. coli* and *B. subtilis* failed to reveal a single clear-cut case of different functions, although subtle differences, e.g., in enzyme or transporter specificities are common (E.V.K., unpublished observations). Thus, in general, the notion that one-to-one orthologs are functionally equivalent seems to hold well.

However, at greater evolutionary distances, particularly across the primary kingdom divides, there are prominent cases of apparent major differences in the functions of orthologs. Thus, the bacterial and archaeal DnaG-like primases, which figure in the previous section in connection with a difference in domain architecture, seem to function in fundamentally different processes. Bacterial DnaG is an essential component of the replication machinery, namely the polymerase responsible for the synthesis of RNA primers used to initiate replication (4). Although the function of the archaeal ortholog has not been studied in detail, there is no evidence of its involvement in replication; furthermore, it has been shown to associate with the exosome, the RNA degradation complex, suggesting a role in RNA processing (21). A converse situation seems to exist with the archaeo-eukaryotic-type primase which is an essential replication component in archaea and eukaryotes, but is involved in a distinct repair pathway in those bacteria that have this gene (11). In this case, the bacterial versions actually are likely to be xenologs of the archaeal and eukaryotic ones (2), and they apparently went through a period of rapid evolution associated with the functional change.

Acceleration of evolution accompanying a radical functional switch seems to have been a major aspect of the emergence of the eukaryotic cell. Thus, to the best of our understanding, eukaryotic tubulins are co-orthologs of the prokaryotic protein FtsZ, which is the key component of the prokaryotic cell division machinery mediating septum formation (1, 58). The well-characterized functions of tubulins are completely different: They are the principal constituents of the eukaryotic microtubules, cytoskeletal structures that are absent in prokaryotes [the recent discovery of tubulins in Prosthecobacteria (39) is remarkable but may be explained by HGT from eukaryotes to a specific bacterial lineage]. The drastic change of function in eukaryotes apparently had been accompanied by a burst of sequence evolution such that an unequivocal demonstration of the homology of FtsZ and tubulin became possible only through comparison of the respective protein structures (65). An analogous situation is seen with the other signature proteins of eukaryotes, actins, and ubiquitins, whose apparent prokaryotic orthologs have completely different functions and dramatically differ in sequence (1, 92, 96). Although in each of these cases, the relationship between prokaryotic and eukaryotic proteins is not one-to-one orthology, with many inparalogs present in eukaryotes, the functional differences among these paralogs are minor compared with the profound divide between prokaryotes and eukaryotes. The general message from this brief survey of the functional equivalency of orthologs and

of the functional switches within orthologous lineages is clear: The fundamental functions of orthologs do change but such changes are far from being common, tend to be associated with major evolutionary transitions, and are accompanied by a substantial acceleration of evolution.

The functional connotations of paralogy are distinct from and, in a sense, opposite to those of orthology. Although in some cases, paralogs may retain the same, ancestral function, being fixed due to the gene dosage effects (amplification of rRNA genes is an obvious example), the general themes associated with paralogy are functional diversification and specialization. The subfunctionalization mode of evolution of paralogs that has been explored theoretically in great detail by Lynch and coworkers (27, 60, 61) seems best compatible with the demonstration that selective constraints affect paralogs even immediately after duplication (45). Examples of subfunctionalization are plentiful. A classic one is the distribution of the universal transcriptional function of the archaeal RNA polymerase between the three RNA polymerases of eukaryotes, with RNA polymerase I becoming responsible for the transcription of rRNA genes, RNA polymerase II transcribing protein-coding genes, and RNA polymerase III transcribing tRNA genes and those for other noncoding RNAs (101). The original version of Ohno's neofunctionalization model, whereby one of the emerging paralogs initially evolves free of constraints (like a pseudogene) but then accidentally hits on a new function, might be unrealistic or, at least, rare. More generally, however, evolution of paralogs, particularly in the context of lineage-specific expansions of paralogous families, may involve both subfunctionalization and neofunctionalization. Indeed, it seems inevitable that, among the enormous repertoire of signal-transduction systems that evolved via multiple duplications, such as protein kinases, receptors, and ubiquitin systems components (to mention just a few of the most conspicuous cases), there should be specificities that were not present in the ancestral gene. Very recently, He & Zhang explored the possibility of combination of subfunctionalization and neofunctionalization by examining protein-protein interactions of paralogous gene products in yeast (36). Their results suggested a more complex subneofunctionalization model under which the evolution of paralogs starts with rapid subfunctionalization but subsequently often switches to the neofunctionalization mode.

GENERAL DISCUSSION

Orthology and Paralogy as Evolutionary Inferences and the Homology Debates

The preceding discussion aimed to show how the notions of orthology and paralogy permeate modern genomics and provide the crucial link between genomics and evolutionary biology. To a large extent, these concepts form the foundation of evolutionary genomics and are also of major importance for functional genomics or, in more practical terms, for functional annotation of sequenced genomes. It is useful, however, to explicitly define the epistemological status of these concepts. Orthology and paralogy as well as the generalized notion of homology imply specific statements on the course of evolution of the respective genes. In other words, these statements are inferences from one or another form of phylogenetic analysis rather than observables. The principal observables in comparative genomics are sequence similarity between genes and the proteins they encode and, increasingly, structural similarity between proteins. These observations are employed, directly or indirectly (e.g., through phylogenetic analysis), to infer orthology and paralogy (or generic homology). Failure to distinguish between observables and inferences resulted in a persistent terminological morass. Strong homology, percentage homology, and similar oxymorons have survived in the biological literature for decades despite strongly

worded refutations, and even guidelines regulating the usage of "homology" that have been adopted by some journals (81, 97). Because of these abuses but, more importantly, because biologists often consider evolutionary inferences to be inherently unreliable, suggestions have been made to dispense with the inferential terms (and, presumably, the underlying concepts) altogether. In a recent provocative article, Varshavsky proposed using the new, inference-free terms "sequelog" and "spalog" to designate, respectively, proteins that show sequence or structural similarity to each other (93). In an earlier eloquent comment, Petsko argued against using the terms ortholog and paralog on the simpler grounds that these terms unnecessarily complicate the narrative in research articles without clarifying anything (80). The desire for simplicity and use of neutral terms is understandable and would have been justified if orthology and paralogy (and all the derivatives thereof) were just words. However, as I argued here and elsewhere (46), this does not seem to be the case. Instead, orthology and paralogy appear to be concepts that carry substantive meaning and, even apart from the (perhaps, debatable to some) intrinsic interest of evolutionary relationships, have major functional connotations. Although the terms orthologs and paralogs may complicate the language of genomics, in my opinion, the clarification they bring to our understanding of the evolutionary and functional relationships among genes and genomes by far outweighs any inconvenience.

Generalized Concepts of Orthology and Paralogy

Throughout most of this review and other treatises, the concepts of orthology and paralogy are applied to genes as units (i.e., whenever we speak of orthologs, we mean orthologous genes). However, I also discussed complications to this approach stemming from gene fusion/fission and, less trivially, from lineage-specific changes in the domain architecture of orthologous proteins occurring during evolution. The latter situation challenges the gene-centric definition of orthology inasmuch as certain parts of genes appear to be orthologous whereas others are not. In principle, it seems possible to extend the notion of orthology to individual domains and, ultimately, to any stretch of nucleotide sequence down to a single base (97). The fundamental definition always remains the same: Genomic elements in the compared species that descend from the same ancestral element in the genome of their last common ancestor should be considered orthologous. From a maximalist standpoint, one could argue that evolutionary relationships within a set of genomes should be considered resolved only after the status of each base pair in each genome (both in coding and noncoding regions) is established with respect to orthology and paralogy. This could be an achievable goal for closely related genomes (e.g., human and chimpanzee) but seems to be unrealistic for distant species.

On the opposite, genome-wide scale, the notions of orthology and paralogy naturally apply not only to genes but to strings of genes that retain the ancestral order (conserved synteny blocks). In relatively closely related genomes (e.g., primates and rodents or different enterobacterial species), a conserved synteny block may include hundreds or even thousands of genes, whereas in distantly related genomes, there is very little conservation of gene order (7, 99). Thus, orthology and paralogy are manifest throughout all levels of genome comparison. Nevertheless, the gene-centric perspective adopted in the preceding sections appears to be most relevant for dissecting the results of comparison of multiple genomes separated by a wide range of evolutionary distances.

A different aspect of generalization of the concepts of orthology and paralogy pertains to the complex structure of orthologous gene clusters caused by the spread of duplication events over the phylogenetic tree. A single orthologous cluster defined at the deepest

branching point of a tree is often resolved into several clusters within subtrees. The cases of tubulins and actins briefly discussed above in a different context clearly illustrate this point. All eukaryotic tubulins are co-orthologous to the single prokaryotic FtsZ proteins, just as actins are orthologous to the prokaryotic MreB. Within eukaryotes, however, several orthologous sets can be readily identified for each of these proteins.

CONCLUSIONS

Orthology and paralogy are not only key terms but are an integral part of the conceptual foundation of evolutionary and functional genomics. Only consistent usage of these and some derivative definitions, such as in- and outparalogs, provides for construction of a robust evolutionary classification of genes and reliable functional annotation of newly sequenced genomes. Further improvement of clustering and phylogenetic methods for identification of orthologs and paralogs is required for the progress of genomics as the number of sequenced genomes rapidly increases. Orthology and paralogy appear to be rich and flexible concepts that allow further development and are well suited to describe the complexity of genome evolution.

SUMMARY POINTS

1. Orthologs and paralogs are two types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication.
2. Distinguishing between orthologs and paralogs is crucial for successful functional annotation of genomes and for reconstruction of genome evolution.
3. A finer classification of orthologs and paralogs has been developed to reflect the interplay between duplication and speciation events, and effects of gene loss and horizontal gene transfer on the observed homologous relationship.
4. Methods for identification of sets of orthologous and paralogous genes involve phylogenetic analysis and various procedures for sequence similarity–based clustering.
5. Analysis of clusters of orthologous and paralogous genes is instrumental in genome annotation and in delineation of trends in genome evolution.
6. Rearrangements of gene structure confound orthologous and paralogous relationships.
7. The gene-centered concepts of orthology and paralogy can be generalized downward, to the level of strings of nucleotides and even single base pairs, and upward, to multigene arrays.

ACKNOWLEDGMENTS

I thank Yuri Wolf, Marina Omelchenko, and Kira Makarova for invaluable help with data analysis; Yuri Wolf for critical reading of the manuscript; and Walter Fitch, Roy Jensen, Pavel Pevzner, Erik Sonnhammer, Alexander Varshavsky, and Emile Zuckerkandl for instructive discussions. Due to space constraints, it was impossible to cite all relevant publications in this review; my sincere apologies and appreciation to all colleagues whose important work is not cited.

LITERATURE CITED

1. Amos LA, van den Ent F, Lowe J. 2004. Structural/functional homology between the bacterial and eukaryotic cytoskeletons. *Curr. Opin. Cell Biol.* 16:24–31
2. Aravind L, Koonin EV. 2001. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res.* 11:1365–74
3. Arvestad L, Berglund AC, Lagergren J, Sennblad B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19(Suppl.)1:i7–15
4. Benkovic SJ, Valentine AM, Salinas F. 2001. Replisome-mediated DNA replication. *Annu. Rev. Biochem.* 70:181–208
5. Bromham L, Penny D. 2003. The modern molecular clock. *Nat. Rev. Genet.* 4:216–24
6. Cutler DJ. 2000. Understanding the overdispersed molecular clock. *Genetics* 154:1403–17
7. Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23:324–28
8. Darwin C. 1859. *On the Origin of Species*. London
9. Davis JC, Petrov DA. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol.* 2:E55
10. Davison AJ, Scott JE. 1986. The complete DNA sequence of varicella-zoster virus. *J. Gen. Virol.* 67:1759–816
11. Della M, Palmbos PL, Tseng HM, Tonkin LM, Daley JM, et al. 2004. Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. *Science* 306:683–85
12. Doerks T, von Mering C, Bork P. 2004. Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. *Nucleic Acids Res.* 32:6321–26
13. Doolittle WF. 1998. You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* 14:307–11
14. Doolittle WF. 1999. Lateral genomics. *Trends Cell Biol.* 9:M5–8
15. Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–29
16. Doolittle WF. 2000. Uprooting the tree of life. *Sci. Am.* 282:90–95
17. Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21:2596–603
18. Duret L, Mouchiroud D, Gouy M. 1994. HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22:2360–65
19. Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8:163–67
20. Eulenstein O, Mirkin B, Vingron M. 1998. Duplication-based measures of difference between gene and species trees. *J. Comput. Biol.* 5:135–48
21. Evguenieva-Hackenburg E, Walter P, Hochleitner E, Lottspeich F, Klug G. 2003. An exosome-like complex in Sulfolo*bus solfataricus*. *EMBO Rep.* 4:889–93
22. Fisher RA. 1928. The possible modification of the response of the wild type to recurrent mutations. *Am. Nat.* 62:115–26
23. **Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99–106**
24. Fitch WM. 1995. Uses for evolutionary trees. *Philos. Trans. R. Soc. London Ser. B* 349:93–102

23. The classical work defining, for the first time, orthologs and paralogs as terms and concepts.

25. Fitch WM. 2000. Homology a personal view on some of the problems. *Trends Genet.* 16:227–31
26. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512
27. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–45
28. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270:397–403
29. Fukuchi S, Nishikawa K. 2004. Estimation of the number of authentic orphan genes in bacterial genomes. *DNA Res.* 11:219–31, 311–13
30. Galperin MY, Koonin EV. 2004. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 32:5452–63
31. Gillespie JH. 1984. The molecular clock may be an episodic clock. *Proc. Natl. Acad. Sci. USA* 81:8009–13
32. Gogarten JP. 1994. Which is the most conserved group of proteins? Homology-orthology, paralogy, xenology, and the fusion of independent lineages. *J. Mol. Evol.* 39:541–43
33. Gray GS, Fitch WM. 1983. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. *Mol. Biol. Evol.* 1:57–66
34. Gray MW, Burger G, Lang BF. 2001. The origin and early evolution of mitochondria. *Genome Biol.* 2
35. Hannenhalli S, Chappey C, Koonin EV, Pevzner PA. 1995. Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics* 30:299–311
36. He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157–64
37. Huxley THH. 1860. 'The Origin of Species'. *Westminst. Rev.* 17:541–70
38. Huynen MA, van Nimwegen E. 1998. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* 15:583–89
39. Jenkins C, Samudrala R, Anderson I, Hedlund BP, Petroni G, et al. 2002. Genes for the cytoskeletal protein tubulin in the bacterial genus Prosthecobacter. *Proc. Natl. Acad. Sci. USA* 99:17049–54
40. Jensen RA. 2001. Orthologs and paralogs—we need to get it right. *Genome Biol.* 2: INTERACTIONS1002
41. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* 11:555–65
42. Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* 4:22
43. Karlberg O, Canback B, Kurland CG, Andersson SG. 2000. The dual origin of the yeast mitochondrial proteome. *Yeast* 17:170–87
44. Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge Univ. Press
45. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biol.* 3: RESEARCH0008
46. Koonin EV. 2001. An apology for orthologs—or brave new memes. *Genome Biol.* 2: COMMENT1005

27. The idea of subfunctionalization as the mode of evolution of paralogs is introduced as an alternative to neofunctionalization.

33. This paper introduces the notion of xenology.

36. The latest study on functional diversification of paralogs integrates the previous models in the subneofunctionalization scheme whereby the subfunctionalization phase immediately after duplication is succeeded by neofunctionalization.

40. Continuation of the debate on the importance of orthologs and paralogs as concepts and terms. Emphasizes the importance of exact definitions, in particular, that the notion of paralogy applies not only to genes in the same genome.

46. Reply to the "Homologuephobia" comment of Petsko. Emphasizes that orthologs and paralogs are not just words but crucial concepts of evolutionary genomics.

47. Koonin EV. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* 1:127–36
48. Koonin EV, Aravind L, Kondrashov AS. 2000. The impact of comparative genomics on our understanding of evolution. *Cell* 101:573–76
49. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, et al. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* 5:R7
50. Koonin EV, Galperin MY. 2002. *Sequence—Evolution—Function. Computational Approaches in Comparative Genomics*. New York: Kluwer
51. Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* 55:709–42
52. Koonin EV, Mushegian AR, Bork P. 1996. Non-orthologous gene displacement. *Trends Genet.* 12:334–36
53. Koonin EV, Wolf YI, Karev GP. 2002. The structure of the protein universe and genome evolution. *Nature* 420:218–23
54. Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multidomain proteins. *Trends Genet.* 21:25–30
55. Kunin V, Ouzounis CA. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* 13:1589–94
56. Lang BF, Gray MW, Burger G. 1999. Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* 33:351–97
57. Lawrence JG, Hendrickson H. 2003. Lateral gene transfer: When will adolescence end? *Mol. Microbiol.* 50:725–27
58. Lowe J, van den Ent F, Amos LA. 2004. Molecules of the bacterial cytoskeleton. *Annu. Rev. Biophys. Biomol. Struct.* 33:177–98
59. Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–55
60. Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–73
61. Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet.* 20:544–49
62. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, et al. 1999. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* 9:608–28
63. Mirkin B, Muchnik I, Smith TF. 1995. A biologically consistent model for comparing molecular phylogenies. *J. Comput. Biol.* 2:493–507
64. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* 3:2
65. Nogales E, Downing KH, Amos LA, Lowe J. 1998. Tubulin and FtsZ form a distinct family of GTPases. *Nat. Struct. Biol.* 5:451–58
66. Novichkov PS, Omelchenko MV, Gelfand MS, Mironov AA, Wolf YI, Koonin EV. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *J. Bacteriol.* 186:6575–85
67. O'Brien KP, Remm M, Sonnhammer EL. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33 Database Issue:D476–80
68. Ochman H. 2002. Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. *Trends Genet.* 18:335–37

49. Description of the first collection of sets of probable orthologs from 7 sequenced eukaryotic genomes (KOGs). Reports analysis of various evolutionary patterns in KOGs, including lineage-specific gene loss, functional characteristics of one-to-one orthologs, and quantitative assessment of domain accretion.

63. The first method for tree reconciliation, in principle, the approach of choice for identification of orthologs.

66. An assessment of the validity of molecular clock on genome scale. Shows that the majority of clusters of one-to-one orthologs evolve in the clock-like mode but also that a significant minority experienced XGD.

69. Ohno S. 1970. *Evolution by Gene Duplication*. New York: Springer-Verlag
70. Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol.* 4:R55
71. Ouzounis C. 1999. Orthology: another terminology muddle. *Trends Genet.* 15:445
72. Owen R. 1848. *On the Archetype and Homologies of the Vertebrate Skeleton*. London: Murray
73. Page RD, Charleston MA. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* 7:231–40
74. Panchen AL. 1994. Richard Owen and the concept of homology. In *Homology: The Hierarchical Basis of Comparative Biology*, ed. BK Hall, pp. 21–62. San Diego: Academic
75. Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–97
76. Patterson C. 1988. Homology in classical and molecular biology. *Mol. Biol. Evol.* 5:603–25
77. Pennisi E. 1998. Genome data shake tree of life. *Science* 280:672–74
78. Pennisi E. 2001. Microbial genomes. Sequences reveal borrowed genes. *Science* 294:1634–35
79. Perrière G, Duret L, Gouy M. 2000. HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.* 10:379–85
80. Petsko GA. 2001. Homologuephobia. *Genome Biol.* 2: COMMENT1002
81. Reeck GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, et al. 1987. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* 50:667
82. Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and inparalogs from pairwise species comparisons. *J. Mol. Biol.* 314:1041–52
83. Snel B, Bork P, Huynen M. 2000. Genome evolution: gene fusion versus gene fission. *Trends Genet.* 16:9–11
84. Snel B, Bork P, Huynen MA. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* 12:17–25
85. Sonnhammer EL, Koonin EV. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18:619–20
86. Storm CE, Sonnhammer EL. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 18:92–99
87. Storm CE, Sonnhammer EL. 2003. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res.* 13:2353–62
88. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformat.* 4:41
89. Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–37
90. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, et al. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* 6:279–91
91. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29:22–28
92. van den Ent F, Amos LA, Lowe J. 2001. Prokaryotic origin of the actin cytoskeleton. *Nature* 413:39–44

69. A seminal work, the first to present a coherent concept of gene duplication as a major formative force of evolution.

72. Introduces homology referring to "the same organ in different animals under every variety of form and function".

80. A witty comment that sparked the discussion of the meaning and importance of the terms orthologs and paralogs.

81. An early condemnation of incorrect uses of the term homology (as in "percent homology," "strong homology" etc). Emphasizes that homology should be used exclusively to refer to common origin of genes (proteins).

82. Introduces the terms in- and outparalogs.

85. Conceptualizes and explains the notions of in- and outparalogs, and coorthologs.

89. Description of the first method for identifying clusters

93. Varshavsky A. 2004. 'Spalog' and 'sequelog': neutral terms for spatial and sequence similarity. *Curr. Biol.* 14:R181–83
94. Veitia RA. 2004. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics* 168:569–74
95. Veitia RA. 2005. Gene dosage balance: deletions, duplications and dominance. *Trends Genet.* 21:33–35
96. Wang C, Xi J, Begley TP, Nicholson LK. 2001. Solution structure of ThiS and implications for the evolutionary roots of ubiquitin. *Nat. Struct. Biol.* 8:47–51
97. Webber C, Ponting CP. 2004. Genes and homology. *Curr. Biol.* 14:R332–33
98. Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. *Trends Genet.* 18:472–79
99. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. 2001. Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context. *Genome Res.* 11:356–72
100. Yanai I, Wolf YI, Koonin EV. 2002. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol.* 3:research0024
101. Zawel L, Reinberg D. 1995. Common themes in assembly and function of eukaryotic transcription complexes. *Annu. Rev. Biochem.* 64:533–61
102. Zhang L. 1997. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *J. Comput. Biol.* 4:177–87
103. Zmasek CM, Eddy SR. 2002. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformat.* 3:14
104. Zuckerkandl E, Pauling L. 1962. Molecular evolution. In *Horizons in Biochemistry*, ed. M Kasha, B Pullman, pp. 189–225. New York: Academic
105. Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence of proteins. In *Evolving Gene and Proteins*, ed. Bryson V, Vogel HJ, pp. 97–166. New York: Academic

of orthologs in multiple genomes and the first version of Clusters of Orthologous Groups of proteins (COGs).

93. The latest twist in the debate on homology, orthology and paralogy. Inference-free terms are proposed to designate sequence and structural similarity between proteins.

105. This and the preceding paper are seminal works that laid the foundation of molecular evolution and include discussion of different types of homologous relationships presaging the concepts of orthology and paralogy.

Annual Review of Genetics

Volume 39, 2005

Contents

John Maynard Smith
Richard E. Michod 1

The Genetics of Hearing and Balance in Zebrafish
Teresa Nicolson 9

Immunoglobulin Gene Diversification
Nancy Maizels 23

Complexity in Regulation of Tryptophan Biosynthesis in *Bacillus subtilis*
Paul Gollnick, Paul Babitzke, Alfred Antson, and Charles Yanofsky 47

Cell-Cycle Control of Gene Expression in Budding and Fission Yeast
Jürg Bähler 69

Comparative Developmental Genetics and the Evolution of Arthropod Body Plans
David R. Angelini and Thomas C. Kaufman 95

Concerted and Birth-and-Death Evolution of Multigene Families
Masatoshi Nei and Alejandro P. Rooney 121

Drosophila as a Model for Human Neurodegenerative Disease
Julide Bilen and Nancy M. Bonini 153

Molecular Mechanisms of Germline Stem Cell Regulation
Marco D. Wong, Zhigang Jin, and Ting Xie 173

Molecular Signatures of Natural Selection
Rasmus Nielsen 197

T-Box Genes in Vertebrate Development
L.A. Naiche, Zachary Harrelson, Robert G. Kelly, and Virginia E. Papaioannou 219

Connecting Mammalian Genome with Phenome by ENU Mouse Mutagenesis: Gene Combinations Specifying the Immune System
Peter Papathanasiou and Christopher C. Goodnow 241

Evolutionary Genetics of Reproductive Behavior in *Drosophila*: Connecting the Dots
Patrick M. O'Grady and Therese Anne Markow 263

Sex Determination in the Teleost Medaka, *Oryzias latipes*
Masura Matsuda 293

Orthologs, Paralogs, and Evolutionary Genomics
Eugene V. Koonin 309

The Moss *Physcomitrella patens*
David Cove 339

A Mitochondrial Paradigm of Metabolic and Degenerative Diseases, Aging, and Cancer: A Dawn for Evolutionary Medicine
Douglas C. Wallace 359

Switches in Bacteriophage Lambda Development
Amos B. Oppenheim, Oren Kobiler, Joel Stavans, Donald L. Court, and Sankar Adhya 409

Nonhomologous End Joining in Yeast
James M. Daley, Phillip L. Palmbos, Dongliang Wu, and Thomas E. Wilson 431

Plasmid Segregation Mechanisms
Gitte Ebersbach and Kenn Gerdes 453

Use of the Zebrafish System to Study Primitive and Definitive Hematopoiesis
Jill L.O. de Jong and Leonard I. Zon 481

Mitochondrial Morphology and Dynamics in Yeast and Multicellular Eukaryotes
Koji Okamoto and Janet M. Shaw 503

RNA-Guided DNA Deletion in Tetrahymena: An RNAi-Based Mechanism for Programmed Genome Rearrangements
Meng-Chao Yao and Ju-Lan Chao 537

Molecular Genetics of Axis Formation in Zebrafish
Alexander F. Schier and William S. Talbot 561

Chromatin Remodeling in Dosage Compensation
John C. Lucchesi, William G. Kelly, and Barbara Panning 615

INDEXES

Subject Index 653

ERRATA

An online log of corrections to *Annual Review of Genetics* chapters may be found at http://genet.annualreviews.org/errata.shtml