Review

How old is my gene?

John A. Capra[1], Maureen Stolzer[2], Dannie Durand[2,3], and Katherine S. Pollard[4]

[1] Center for Human Genetics Research and Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, USA
[2] Department of Biological Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[4] Gladstone Institutes, Institute for Human Genetics, and Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94158, USA

Gene functions, interactions, disease associations, and ecological distributions are all correlated with gene age. However, it is challenging to estimate the intricate series of evolutionary events leading to a modern-day gene and then to reduce this history to a single age estimate. Focusing on eukaryotic gene families, we introduce a framework that can be used to compare current strategies for quantifying gene age, discuss key differences between these methods, and highlight several common problems. We argue that genes with complex evolutionary histories do not have a single well-defined age. As a result, care must be taken to articulate the goals and assumptions of any analysis that uses gene age estimates. Recent algorithmic advances offer the promise of gene age estimates that are fast, accurate, and consistent across gene families. This will enable a shift to integrated genome-wide analyses of all events in gene evolutionary histories in the near future.

What is gene age?

The functions of a new gene are forged by the adaptive challenges facing the organism at the time that the gene arose. For example, genes that encode functions associated with basic cellular processes, such as transcription, are often as old as life itself, whereas many genes associated with cellular adhesion and communication arose at the dawn of multicellularity. Recent advances in computational biology and genome sequencing have made it possible to explore the connection between gene age and function across the tree of life. The resulting analyses are revolutionizing our understanding of embryonic development, molecular interactions, disease, and the interplay of environment and evolution on geological time scales.

Strictly speaking, however, a gene does not have a single age. Unlike fossils or specific evolutionary events, genes are dynamic entities with continuous histories that trace back to the origin of all life. So, how should ‘gene age’ be defined? Many previous studies have simply used the most recent common ancestor (MRCA; see Glossary) of the species containing genes with similar sequences (e.g., with a significant BLAST score). Although this relatively simple approach has produced compelling results, many genes have complex evolutionary histories that are not accurately summarized by the MRCA. Ideally, this entire history would be used in gene age analysis. In practice, gene age is frequently equated with the timing of a salient event, such as a gene duplication, horizontal transfer, or *de novo* creation of a gene [1]. However, many methodological and philosophical challenges arise when selecting the most appropriate event and estimating its age. To motivate our discussion of these problems, we first review a few striking findings that highlight the broad range of questions that can be addressed using gene age estimates.

Glossary

Character: any observable feature of an organism (e.g., a DNA sequence, a morphological phenotype, or a behavior).
Dollo parsimony: a common gain–loss phylogenetic analysis method based on parsimony and the assumption that a biological character can only be gained once, although it may experience multiple losses in different lineages.
Gain–loss models: a class of methods for reconstructing the phylogenetic history of a biological character (i.e., its state at ancestral nodes in a species tree) that considers only gain and loss events along a species tree.
Homologous family: a collection of genes with significant evidence of homology.
Homology: the relationship of DNA sequences (or other biological characters) related by vertical descent; that is, sequences derived from a common ancestor via speciation or gene duplication. Note that homology indicates only shared ancestry and does not imply conserved function.
Incomplete lineage sorting: the presence of multiple gene genealogies across a genome, some of which may not match the species tree.
Most recent common ancestor (MRCA): the most recent ancestral organism from which all genes (or other characters) of interest are derived.
Neofunctionalization: when one of the two genes created by gene duplication takes on a novel function not carried out by its progenitor.
Orthology: the relation of homologous DNA sequences (or other biological characters) created by a speciation event at their MRCA. Sequences with this relation are called orthologs and are said to be orthologous.
Paralogy: the relation of homologous DNA sequences (or other biological characters) created by a duplication of their MRCA. Sequences with this relation are called paralogs and are said to be paralogous.
Parsimony: the principle that the simplest explanation should be preferred (e.g., when applied to phylogenetics, parsimony prioritizes the tree with the smallest number of evolutionary events that fits the observations).
Phylogenetic reconciliation: a method for reconstructing the phylogenetic history of a biological character that finds a correspondence between a character tree (usually a gene tree) and a species tree in terms of a set of allowable ancestral evolutionary events.
Phylogenetic tree: a diagram that illustrates inferred evolutionary relationships between biological entities. For example, a species tree relates the history of speciation events that produced observed species. A gene tree gives the series of evolutionary events that relate genes observed across one or many species.
Subfunctionalization: when two genes created by gene duplication each take on a subset of the functions of their progenitor.
Wagner parsimony: a gain–loss phylogenetic analysis method based on parsimony that allows multiple gain and loss events, potentially with different likelihoods.

Corresponding authors: Durand, D. (durand@cmu.edu);
Pollard, K.S. (kpollard@gladstone.ucsf.edu).

Keywords: phylogenetics; gene age; molecular clock; eukaryotes.

0168-9525/$ – see front matter
http://dx.doi.org/10.1016/j.tig.2013.07.001

The evolutionary history of a gene is informative about its function

Gene age has been used productively in studies ranging from genome-scale statistical analyses to studies of specific gene families. The link between the age of a gene and when it is expressed during embryonic development is a powerful example. Species in many phyla progress through a 'phylotypic' stage, in which species with highly divergent adult morphologies display dramatic phenotypic similarities. This relation between ontogeny and phylogeny has been known for decades, but its molecular basis is still not fully understood. A recent analysis of the phylogenetic age of the genes expressed across development in zebrafish, flies, and nematodes demonstrated that genes expressed during the phylotypic stage are significantly 'older' than those expressed during earlier and later developmental stages that show species-specific characteristics [2].

Gene origin analysis has also demonstrated that many functional attributes of eukaryotic genes are associated with their time of origin. For example, younger genes in fungi, insects, and mammals have higher rates of evolution [3–5] and experience more variable selection patterns [6,7] compared with older genes. In several yeast species, young genes have fewer physical interactions and are enriched for different functions compared with old genes [8–10]. Young genes are expressed in fewer tissues [11,12] and are regulated by fewer genes [13] in humans.

The specific mechanism that gave rise to a new gene also influences its functional fate (reviewed in [1,14]). It was long thought that duplicated genes are less likely to be essential compared with their singleton counterparts due to the potential for functional overlap and compensation. This pattern was observed among duplicate genes in yeast [15], but conflicting results were obtained in mouse [16–18]. By stratifying mouse genes by age, it was demonstrated that essentiality is in fact lower among duplicate genes compared with singletons of similar age [19]. This consistent pattern is masked among all genes because older mouse genes are more likely to be essential, and duplicates are often derived from older genes.

As suggested by the relation between gene age and essentiality, the evolutionary history of a gene is also connected to its disease associations. For example, Mendelian disease genes are, on average, older than nondisease genes, whereas genes associated with complex diseases are 'middle aged' [20]. Among genes associated with cancer, there is strong enrichment for origins during two evolutionary periods: the origin of all cellular life and the emergence of multicellular animals [21]. More strikingly, this distinction based on age largely recapitulates a

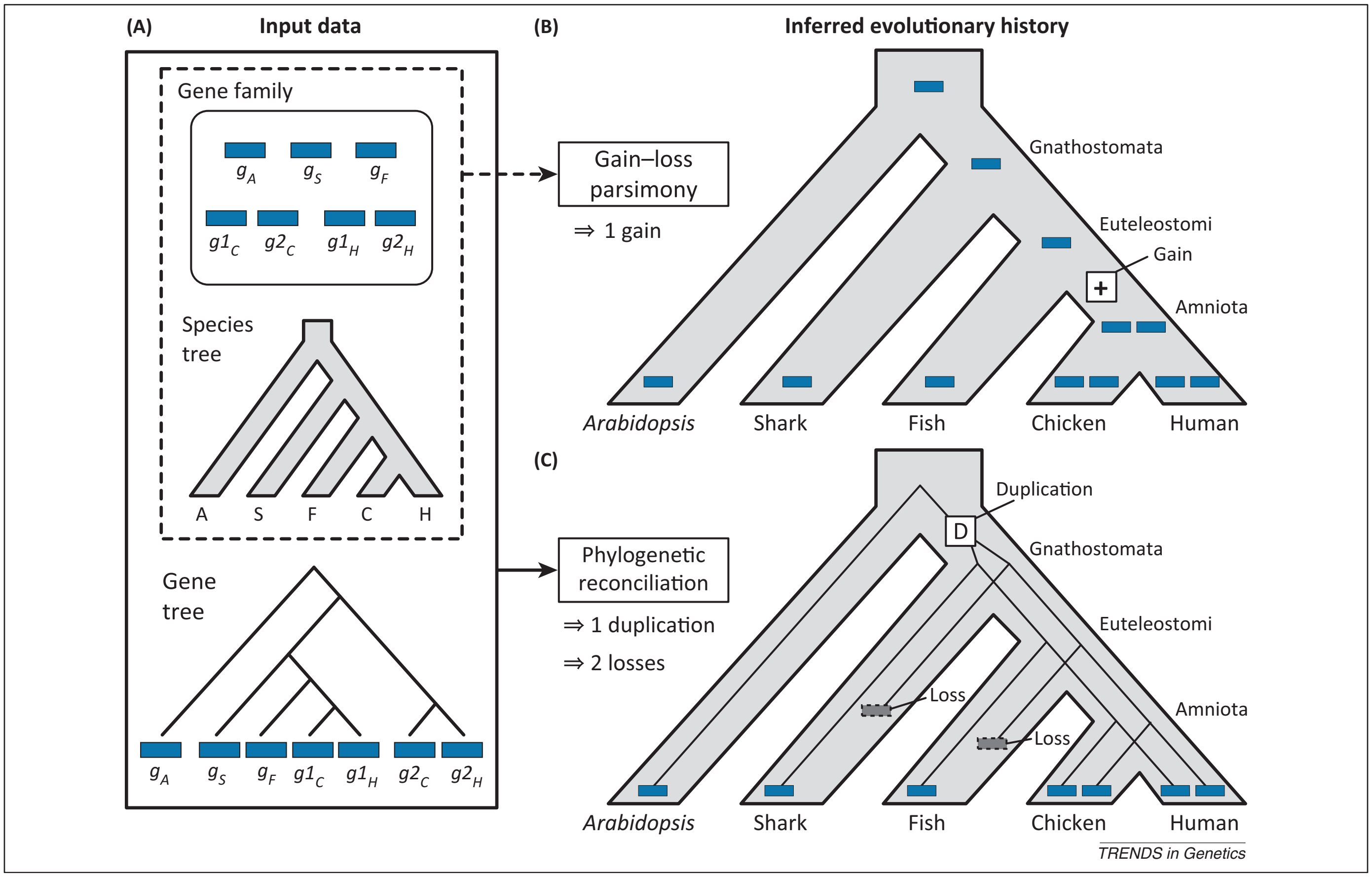


Figure 1. A typical error made by gain–loss methods is avoided with reconciliation. A gene family with a history of parallel losses illustrates the increased accuracy associated with explicit use of a gene tree by phylogenetic reconciliation. **(A)** A hypothetical gene family, based on the real enzyme family in Figure 2 (main text), has one gene in *Arabidopsis*, sharks, amphibians, and fishes, and two genes in each amniote species. **(B)** A gain–loss method, Wagner parsimony, incorrectly infers a single gene family member in the common ancestral species and a recent gain on the lineage leading to chicken and human. This scenario implies that all chicken and human genes are equally related to g_S and g_F, an inference that is not supported by the true gene tree. **(C)** Gene tree–species tree reconciliation correctly infers an earlier duplication, followed by parallel losses in the shark and fish lineages, and shows that $g1_H$ and $g1_C$ are more closely related to g_S in shark and g_F in fish, than to $g2_H$ and $g2_C$.

division of cancer genes based on the functional disruptions that they produce. The older genes correspond to the 'caretakers', which support genome integrity and the fidelity of basic cellular processes, whereas many of the genes that arose around the origin of multicellular animals are the 'gatekeepers', in which mutations disrupt the regulation of cell signaling and growth.

Gene age analyses have also been used to relate the emergence of novel gene functions with ecological, geological, and evolutionary events. For example, an analysis showing that the metazoan ancestor likely had nearly all components of the postsynaptic scaffold was used to elucidate the origin of the animal nervous system [22]. Another study demonstrated that a gene duplication, which resulted in a novel fetal globin gene, coincided with the emergence of placental mammals [23]. This new gene has a crucial role in mammalian development; fetal globin binds oxygen with a higher affinity, enabling efficient transfer of oxygen from the bloodstream of the mother to the fetus. This links a novel gene to a novel developmental process.

Taken together, these studies highlight the diversity of questions that can be addressed by evolutionary analysis of gene age and the inferential power of gene age for understanding gene function. Gene age analysis helps to explain questions in comparative development, gene regulation, processes and rates of evolution, and disease genes. Consideration of gene age in the context of organismal evolution can shed light on the genetic forces driving morphological innovation. Analyses of gene age in the broader context of the history of the Earth reveal connections between new metabolic capabilities and ecological change.

Estimating gene age in three steps

The above studies all analyze gene age, but closer scrutiny reveals that they vary substantially in how gene age is defined and the methods used to estimate this quantity. Although such differences are rarely acknowledged, they can affect the conclusions of a study. To highlight these issues, we examine the process of estimating gene age step by step and propose guidelines for best practices.

Step 1: time scales for gene age analyses

Several measures are used to quantify gene age. Here, we define the timing of an evolutionary event with respect to the lineage (i.e., the branch in the species tree) in which it occurred. For example, the duplication in Figure 1C occurred after the divergence of *Arabidopsis* and gnathostomes and before the separation of sharks and euteleostomes. Expressing gene age in terms of the species tree makes it easy to correlate new genes with functional innovations at the species level. Another common approach is to express gene age in terms of sequence divergence, which can be quantified in a variety of ways using sequence evolution models. Sequence divergence has been used, for example, to identify paralogs that originated from whole-genome versus smaller scale duplications [24]. Both of these approaches can be extended to obtain age estimates in years if there is sufficient fossil evidence to date species divergence times or calibrate rates of sequence evolution (Box 1).

Box 1. Measuring gene age in years

Methods for estimating gene age frequently define evolutionary events on a species tree. If nodes in these trees can be reliably dated, events in the history of a gene, from its origin to the most recent modification, can be interpreted in the context of geological events and the evolution of molecular and organismal traits. For example, dating lignin-degrading enzyme duplications in the common ancestor of white rot fungi (Figure 4, main text) to the same time as the origination of the boreal forest biome, demonstrated that diversification of fungal nutritional modes coincided with the diversification of forest plants [43]. Furthermore, the expansion and subsequent contraction of white rot fungi correlated with variation in the rate of carbon deposition from the Carboniferous to the Permian, highlighting the importance of lignin-degrading enzyme origination in the coal deposits of Earth [44].

When species and gene trees are built from sequence data, the length of a branch represents the expected number of substitutions per site. If the molecular clock hypothesis holds (i.e., if the rate of substitutions is the same in all lineages), then branch lengths can be translated into years using a single, accurately dated node. However, tests of the molecular clock hypothesis reveal that it is valid for relatively few data sets. In general, substitution rates vary over time, between species, and across different genes due to changes in mutation rates, selection pressures, generation times, population sizes, and many other forces. Underestimation of substitutions on long branches, taxon sampling, and extensive gene loss are additional sources of error. Finally, fossils sometimes provide good minimum ages for nodes in a species tree, but provide less information about upper bounds [52] and divergence times in gene trees. Thus, there is typically a great deal of variance in estimates of node ages in gene and species trees (Figure 5, main text), which can confound gene age estimates and downstream analyses [53,54].

Several recent approaches to gene age estimation have tackled the problems associated with calibrating phylogenetic trees in various ways. For instance, relaxed molecular clock methods (e.g., [55]) address these problems by jointly estimating substitution rates and node dates, although they rely upon modeling assumptions and often require substantial computation. Molecular clock calibration methods could also potentially leverage recent breakthroughs in sequencing ancient DNA from dated fossils [56]. Others have used simulations to suggest that evolutionary rate variation across genes does not dramatically affect gene age estimates [57].

Step 2: inferring events in the evolutionary history of a gene

Identifying members of a homologous gene family is a prerequisite for most gene age inference methods (Box 2). Decisions made during homologous family identification, which is usually guided by sequence, structural, or contextual similarity, can affect gene age estimation, as described in Step 3. Once gene families have been defined, gene age inference methods are applied. These typically fall into two categories: gain–loss approaches and phylogenetic reconciliation (Figure 1).

The most basic and commonly used gain–loss algorithm, Dollo parsimony [25], considers only the presence or absence of a gene family in each leaf of the species tree; see, for example, the 'phylostratigraphic' approach [2,21,26,27]. Because Dollo parsimony allows multiple losses, but only one gain, it infers the origin of a gene family to be the MRCA of all species that contain that family. More powerful gain–loss approaches, such as Wagner parsimony [28], consider gene family size (i.e., the number of family members) in each species and allow different weights for gains and losses. The ancestral lineages in which gene family

Box 2. Finding homologous genes families

Gene family identification is an essential prerequisite for most gene age analyses. This typically involves two steps: identifying pairs of homologous sequences and grouping those pairs into sets, each corresponding to a family that shares a common ancestor.

Several approaches to homology detection have been used in the context of gene age analysis. The most common strategy is the use of sequence comparison (e.g., BLAST [58]) to detect significant similarity between gene or protein sequences. However, as sequence similarity decreases, pairwise alignment approaches lose power to detect homology. Many additional sequence-comparison methods have been developed that are able to detect more remote homology by building statistical profiles from multiple alignments of sequences related to the query (e.g., [59–61]) or analyzing known and predicted structural similarities (e.g, SCOP [62,63]). Homology detection methods that consider structural similarity between proteins have the ability to detect remote homology. Their use is rare in gene age analysis, but see [64] for an example. In all homology searches, parameter settings (e.g., the minimum sequence similarity or minimum coverage of the sequence required) can dramatically affect the resulting homology predictions.

Once homology has been established among pairs of genes, networks are constructed from the pairwise relations, and dense subnetworks corresponding to gene families are identified (e.g., using OrthoMCL [65] or InParanoid [66]). In addition to grouping related sequences together, the clustering helps to correct false and missing homology predictions. Most clustering methods require setting one or more parameters that determine cluster sizes and, therefore, the degree of similarity among family members. A tight clustering will result in families that approximate orthologous groups, whereas more lenient parameters will yield homologous families containing paralogous subfamilies. These parameter choices influence the potential age estimates for a gene, for instance, by limiting the set of possible progenitors. Because a lenient clustering offers more flexibility in defining gene origins, it is often preferable, especially if reconciliation is used. However, large-scale resolution of orthologs and paralogs within a homologous family has proven to be a problem that requires further algorithm development [67–69].

expansions and contractions occurred are inferred by minimizing the weighted sum of gains and losses.

Phylogenetic reconciliation also starts with a rooted species tree, a set of possible events, and a weight associated with each event. Gene family information is represented as a rooted gene tree reconstructed from gene sequences in a preprocessing step. Reconciliation uses the incongruence between the gene and species trees to infer the minimum weight set of events that best explains this incongruence and the correspondence between ancestral genes and ancestral species. Over the past two decades, a rich reconciliation methodology has emerged (reviewed in [29,30]). A key distinguishing feature of different reconciliation algorithms is the set of allowable events used for inference. The most complete event model includes duplication, transfer, and loss, although many methods use only a subset of these events.

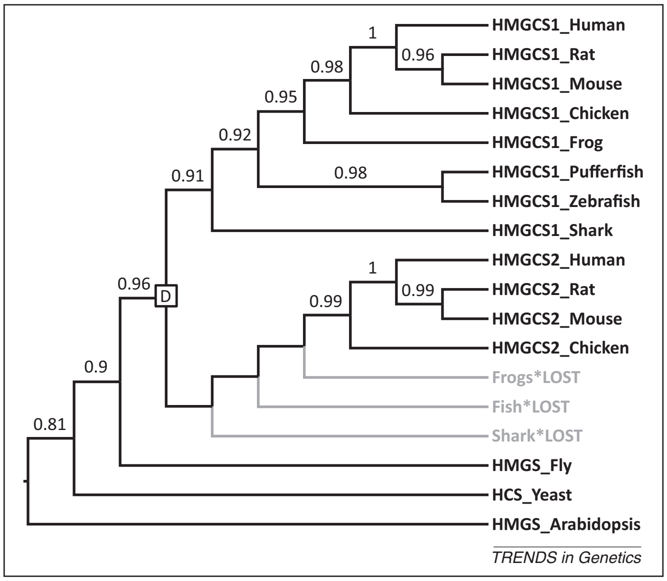
Gain–loss methods and phylogenetic reconciliation can yield different evolutionary histories for the same family. At the heart of these differences is the way that each approach uses information from gene sequences and trees. Gain–loss methods only consider gene family size in each species, but do not take sequence variation or gene family history into account. Failing to consider these relations can result in incorrect inference of gene duplications, gene transfers, and ancestral gene-to-species associations. In particular, gain–loss methods are unable to resolve families that sustained a transfer between distant species or parallel gains that occurred independently in two separate lineages. In both cases, gain–loss methods tend, incorrectly, to infer an early gain followed by later losses. Figure 1B shows another problem case in which parallel losses are erroneously interpreted as a single, recent gain. This error occurs in the 3-hydroxy-3-methylglutaryl coenzyme A synthase (HMGCS) enzyme family (Figure 2). Amniotes have two copies of this enzyme: one that acts in the cytosol (HMGCS1) and one that acts in the mitochondria (HMGCS2). By contrast, fish, frogs, and sharks have a single HMGCS gene. Gain–loss parsimony implies that the second copy arose in the amniote ancestor, but reconciliation shows that it is much older. The more accurate, reconciliation-based estimate of HMGCS age is crucial for understanding the evolution of functional specialization in human deacetylation pathways: HMGCS1 and HMGCS2 are deacetylated by sirtuins, which are themselves ancient paralogs with distinct cytosolic and mitochondrial localization [31]. Gene age analysis makes it possible to assess the relative timing of the HMGCS and sirtuin duplications.

Reconciliation, unlike gain–loss methods, can correctly diagnose distant transfers and handle independent expansions and contractions. This is because reconciliation methods explicitly account for gene family history by incorporating information from a gene tree. In general, reconciliation tends to be more accurate than does gain–loss, provided that the gene tree is correct. If the gene family history is consistent with the assumptions of the gain–loss method, both gain–loss and reconciliation will infer the correct history. If the family violates those assumptions, gain–loss is likely to find an incorrect scenario that requires fewer events. Reconciliation is not sensitive to this type of error because it is constrained by the structure of the gene tree. Reconciliation assumes that any disagreement with the species tree is due to duplications, losses, or transfers. Errors can arise due to incomplete lineage sorting, which can result in a gene tree that disagrees with the species tree, but is not evidence of duplication or transfer. These errors can be avoided by using one of several recent methods that account for such incongruence [32–34]. Both gain–loss and reconciliation are also sensitive to family definition, species tree accuracy [35], sparse taxon sampling, gene conversion, and convergent evolution.

Most gain–loss and reconciliation methods infer ancestral gene content based on a parsimony criterion. Alternative probabilistic approaches have been proposed for both gain–loss (e.g., [36,37]) and reconciliation models [34,38,39]. These methods are more suitable than is parsimony when event rates are high, but they come at the cost of significantly increased computational time. They can also learn event weights from data if a large number of gene families are analyzed together.

Step 3: assigning an age to a gene based on its evolutionary history

Each event in the evolutionary history of a gene has a uniquely defined age, even if it is difficult to estimate correctly in practice. By contrast, the age of a gene is



TRENDS in Genetics

Figure 2. Phylogeny of the 3-hydroxy-3-methylglutaryl coenzyme A synthase (HMGCS) gene family. Human, mouse, rat, and chicken have two copies of the HMGCS enzyme: one that acts in the cytosol (HMGCS1) and one that acts in the mitochondria (HMGCS2) [31]. By contrast, fishes, frogs, and sharks have a single copy of the enzyme. Based on this phylogenetic distribution, gain–loss parsimony infers a recent gain on the branch leading to amniotes. The HMGCS gene tree tells a different story: the two HMGCS subfamilies arose via an early duplication at the base of the vertebrate lineage followed by three parallel losses in sharks, fishes and amphibians. The branching order of the gene tree, with branch support values >0.9, strongly supports these conclusions and rejects the seemingly more parsimonious history with a single, more recent gain. Gene tree inferred using PhyML [70] from sequences aligned with MAFFT [71] and rooted using three invertebrate outgroup sequences. Branch support was assessed using aLRT scores [70]. Species: human, *Homo sapiens;* rat, *Rattus norvegicus;* mouse, *Mus musculus;* chicken, *Gallus gallus*; frog, *Xenopus tropicalis*; pufferfish, *Takifugu rubripes*; zebrafish, *Danio rerio*; shark, *Callorhinchus milii*; fly, *Drosophila melanogaster*, yeast, *Saccharomyces cerevisiae*; Arabidopsis, *Arabidopsis, thaliana*.

not uniquely defined; there is often more than one ancestral gene that is a reasonable choice for the progenitor. Additional ambiguity results from the hierarchical structure of multigene families, which arises through repeated cycles of duplication followed by functional differentiation. This process creates functionally related, but distinct, subfamilies. To assign gene age in a multigene family, such as the HMGCS enzymes or membrane-associated guanylate kinases (MAGUKs) (see below), one must determine where to query the gene family hierarchy for a given analysis.

If the goal of the analysis is to consider how gene age is related to gene function, then the mode of functional differentiation following gene duplication also influences how age is best defined. Duplicates may continue to perform the ancestral function (e.g., for increased dosage), acquire new functions (neofunctionalization), or the ancestral functions may be partitioned between them (subfunctionalization). In the case of neofunctionalization, for example, it is reasonable to equate the age of the genes with the time of the duplication, but in cases where the ancestral function is retained, the appropriate age may be much earlier.

Defining gene age is even more complicated in multidomain gene families. These families encode proteins with multiple functional domains that have been integrated at different times and in different lineages. Again, the best solution depends on context. If the goal is to investigate how changes in domain content relate to protein function, then focusing on the age of the individual domains is a good approach. However, in a study concerned with overall protein function, rather than specific domains, gene age should be guided by the gene family organization and not by domain organization.

The problems of inferring gene age in hierarchical families and in families encoding multidomain proteins both arise in the MAGUK family (Figure 3). MAGUKs generally act as scaffolds that organize protein complexes required for cell–cell interactions (reviewed in [40]). This family evolved through multiple rounds of duplication and functional differentiation [41], resulting in subfamilies that perform more specific functions, such as synaptic signaling and plasticity (DLG1/2/3/4) [42] or organization of epithelial tight junctions (ZO) [40]. This hierarchical structure complicates gene age assignment (e.g., MPP1 has at least three possible ages). In addition, MAGUK genes comprise complex combinations of domains with different histories. Domain architectures are not strictly conserved within subfamilies, and even orthologs sometimes differ in domain architecture (e.g., DLG3 has three PDZ domains in mouse and human, but only two in chicken). Hence, defining gene age in terms of the common ancestor of genes with exactly the same domain architecture is too restrictive. Defining the origin of the MAGUKs in terms of a single domain is also problematic, because most constituent domains pre-date, and none uniquely characterizes, this family. Thus, there are many reasonable definitions for the age of the MAGUK family; the best choice depends on the specific question that the analysis is intended to address.

Case study: the complex evolutionary history of fungal wood decay enzymes sheds light on their role in forest ecology and coal deposition

A beautiful example of the power and the challenges of gene age analysis is provided by a series of studies of expansions and contractions in wood-decay families during the evolution of white and brown rot fungi [43,44]. Both rots can break down cellulose in wood, but only white rot can break down lignin. This difference has profound ecological and economic implications because lignin acts as a repository for nonatmospheric, organic carbon and is the primary precursor of coal.

To uncover the role of lignocellulose degradation genes in wood-rot evolution, a recent study reconciled gene trees for 27 enzyme families with the associated tree of 31 fungal species using Notung [32] and DrML [39]. This analysis revealed ancestral gene family sizes and the gene duplications and losses on each branch of the species tree [44], showing that the MRCA of the white rots studied (starred node in Figure 4) had an enzymatic arsenal capable of lignin degradation and was likely a white rot. Subsequently, brown rots evolved independently, several times, through lineage-specific losses of lignin-degradation enzymes. Simultaneously, white rot lineages experienced additional expansions of those families. These results, combined with a fossil-calibrated molecular clock analysis (Box 1), predicted that the expansion in lignin-degrading

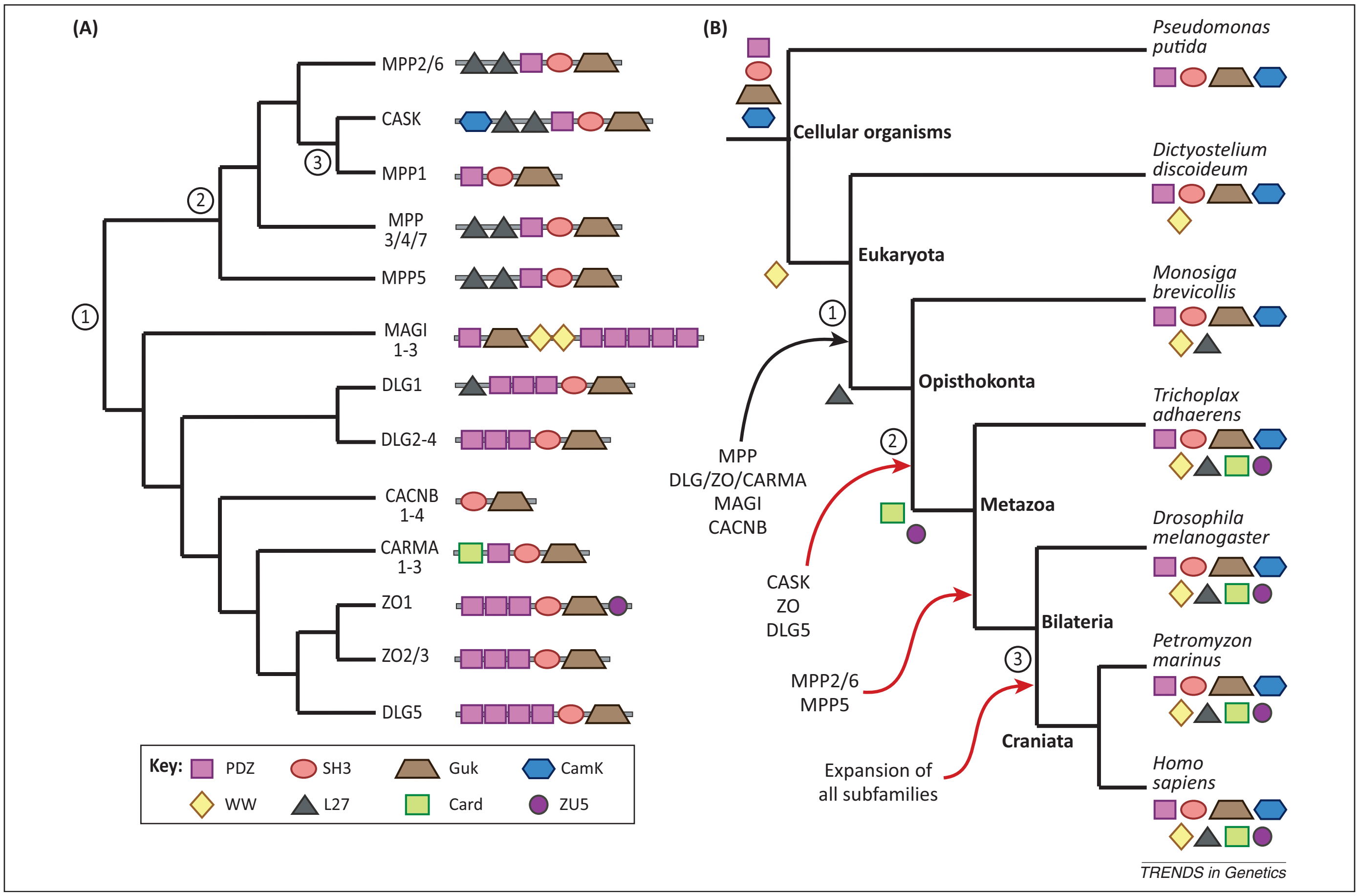


Figure 3. Gene origin and age are not uniquely defined in the human membrane-associated guanylate kinase (MAGUK) superfamily. This superfamily is a multigene family with complex substructure and domain architectures. There are many potential choices for the MAGUK progenitor, which span a broad range of times. **(A)** MAGUK gene tree inferred from the guanylate kinase domain. Nodes 1–3 show three possible origins for MPP1: (1) the origin of the entire MAGUK family (pre-Opisthokonts); (2) the duplication that gave rise to separate CASK and MPP1 genes (pre-Metazoan); and (3) the common ancestor of MPP1 and its orthologs (pre-Craniata). Domain architectures are shown on the leaves. Clades of genes with identical domain architectures are collapsed (e.g., MAGI1-3). **(B)** Species tree showing lineages when domains and specific MAGUKs first appeared. Leaves are decorated with the domains that are present in that species. Arrows indicate when the progenitor of a subfamily first arose, with the subfamily listed below that arrow. Red arrows indicate duplication events that either expanded a subfamily or gave rise to a new one. Branches 1, 2, and 3 represent three possible ages for MPP1, and correspond to nodes 1, 2, and 3, respectively, in the gene tree. Domain origins were inferred using Dollo parsimony with Count [36]. Gene origins are based on phylogenetic analysis [41] and Dollo parsimony.

enzymes occurred at the beginning of the Permian era, suggesting a possible link between the emergence of white rot and the sudden drop in coal deposition at the end of the Permian [45]. Reliable methods for inferring gene age were essential to these studies, because their biological conclusions depend crucially on when lignocellulytic gene family expansions and contractions occurred.

To investigate how method choice influences gene age estimates, we compared the reconciliation-based analysis of seven families of wood-degrading oxidoreductases [43] to our own gain–loss analysis of the same data (Figure 4B). The results showed that the gain–loss method was not able to detect the pronounced increase in oxidoreductase gene content in the white rot ancestor (16 gains versus three gains) and missed several of the independent expansions and contractions that the authors of the original paper identified by reconciliation.

These studies highlight important aspects of each of the three steps of gene age analysis. First, the species tree is used as the primary scale on which to express gene age, augmented by sequence divergence and fossil calibration to relate events in gene and species evolution to geological and ecological trends. Second, two different age estimation strategies predict different ages for the same gene, resulting in different interpretations of wood rot evolution and carbon deposition. Third, this analysis sidesteps the problem of selecting a progenitor by considering gene duplications and losses throughout the species tree.

Perspectives

Gene age analysis has great power to highlight correlations between the history of a gene and its functions in the context of organismal evolution, ecological dynamics, and global biogeochemical processes. At the same time, this field faces major challenges.

A fundamental problem is that a gene does not have a single, well-defined age. For genes in a complex, multigene family, there are several events that could be considered the origin of the gene. Furthermore, genes that encode multidomain proteins may contain sequence fragments with different histories. The many potential ages for MAGUK genes illustrate both of these phenomena (Figure 3). An appealing solution is to consider all possible progenitors at all levels of the gene family hierarchy simultaneously by summarizing the collective events in the history of a gene on a species tree. This comprehensive

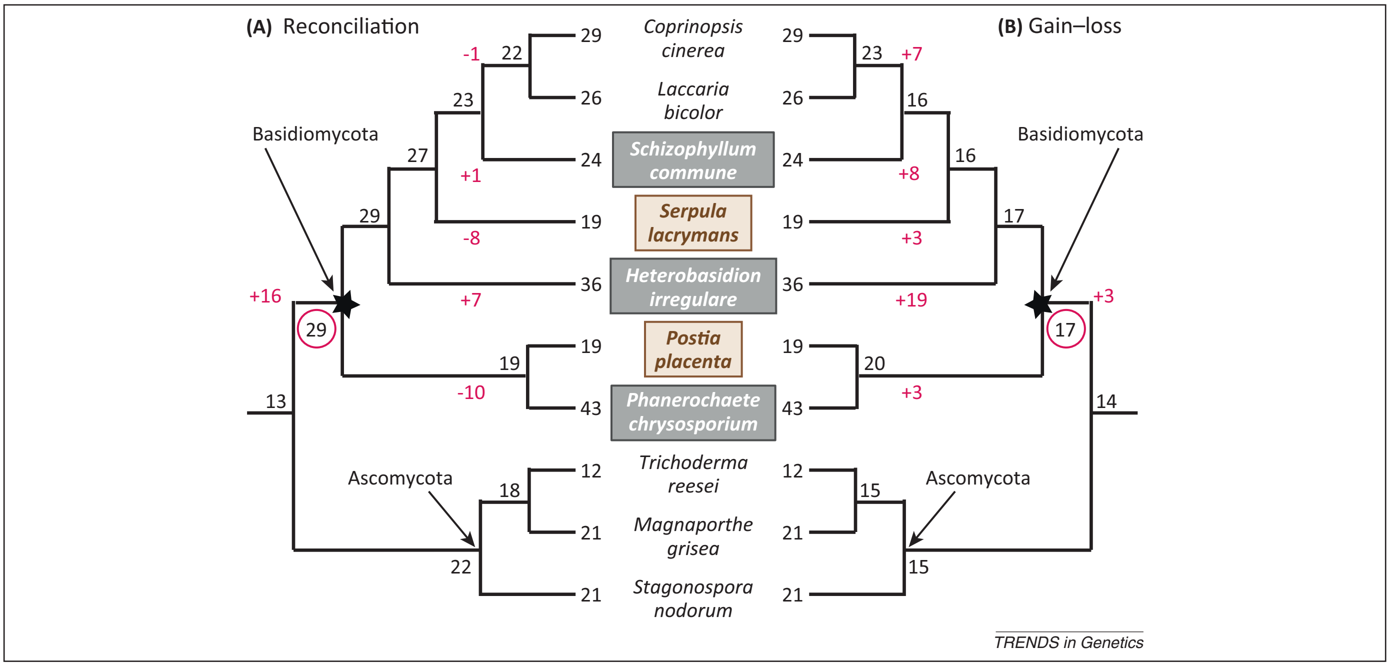


Figure 4. Reconciliation reveals the dynamic history of fungal oxidoreductase genes. Analysis of fungal oxidoreductase gene families shows that a gain–loss approach can obscure the dynamics of gene family expansion and contraction, whereas reconciliation identifies a richer set of events. Internal nodes are labeled with the number of inferred ancestral oxidoreductase genes. Branch labels show inferred gains and losses that changed dramatically between the two analyses. Brown rot fungi are indicated by brown text in a tan box; white rot by white text in a gray box. **(A)** The original, reconciliation-based analysis of seven oxidoreductase gene families in ten fungal species, adapted from Figure 1B in [43]. This analysis infers a moderate oxidoreductase complement in the white rot most recent common ancestor (MRCA; starred node), with substantial independent expansions in the Ascomycota and the Basidiomycota. The oxidoreductase gene complement in the ancestral white rot species (red circle) has the most oxidoreductase genes among ancestral nodes. Independent, lineage-specific gene duplications and losses in white and brown rots, respectively, gave rise to present-day oxidoreductase counts. **(B)** Our analysis of the same data set using a gain–loss analysis predicts smaller gene family sizes in the white rot ancestor and lineage-specific expansions, rather than contractions, in *Serpula* and the lineages leading to *Coprinopsis* and *Laccaria* and to *Postia* and *Phanerochaete*. The expansions in *Heterobasidion* and *Schizophyllum* are substantially over-estimated compared with (A). Ancestral gene family sizes were inferred using Wagner parsimony with equal weights implemented in Count [36].

characterization solves the problem of selecting a single age for each gene and enables detection of subtle patterns in the family as a whole. New techniques enable testing of alternate hypotheses concerning when a novel function first arose, for example, by experimentally probing the function of an ancestral protein by resurrecting its inferred amino acid sequence via synthesis or mutagenesis of modern proteins [46–48].

The second major challenge is that current methods can predict different ages for the same gene. Conceptual tractability and computational efficiency have made gain–loss methods popular, despite their biases [49]. However, reconciliation methods are more accurate, because they take advantage of the information encoded in the gene family tree and make less restrictive assumptions. The HMGCS and white rot studies provide concrete examples of how different gain–loss and reconciliation-based age estimates can be, to say nothing of the downstream functional predictions. Fortunately, although reconciliation requires more complex calculations, algorithmic developments have made substantial headway in improving accuracy and reducing computational time [33,50].

Unlike many other areas of genomics, inexpensive high-throughput sequencing is not a panacea for gene age analysis and introduces new problems. Accurate gene age analysis requires scalable algorithms for homology detection, multiple sequence alignment, phylogenetic inference, and reconciliation that outpace sequencing projects and simultaneously improve accuracy. Furthermore, increasing data not only changes the scale of the challenges faced, but also their fundamental nature. In phylogenetic inference, for example, problems with taxon sampling will likely decrease, but noise due to incomplete lineage sorting will increase as we sample more genomes.

Despite these challenges, gene age analysis on a multi-genome scale offers exciting opportunities. With the ever-increasing rate of genome sequencing, reliable, automated gene annotation methods are essential. Because evolutionary patterns are now the main source of information about gene functions for most genomes, the potential contribution of gene age analysis to functional annotation is significant. Genome-scale analyses of gene age in many gene families, considered simultaneously, can illuminate how genes work together in present-day organisms and how systems of interacting components evolved. Statistical analyses of broad trends across a large number of gene families are revealing underlying principles, such as the complex association between age, mechanism of origin, and essentiality among mouse genes [19]. They can also link trends in gene family expansion and contraction to phenotypic and ecological changes.

More efficient algorithms and faster hardware are putting reconciliation-based age analysis of all gene families in a collection of genomes within reach, but fully exploiting this potential will require additional methodological advances. Genome-scale phylogenetics requires better ways to assess how much noise there is in a large collection of trees. A related issue is that most nodes in the tree of life cannot be accurately dated (Figure 5, Box 1), making it challenging to link gene family origins and expansions to geologic events.

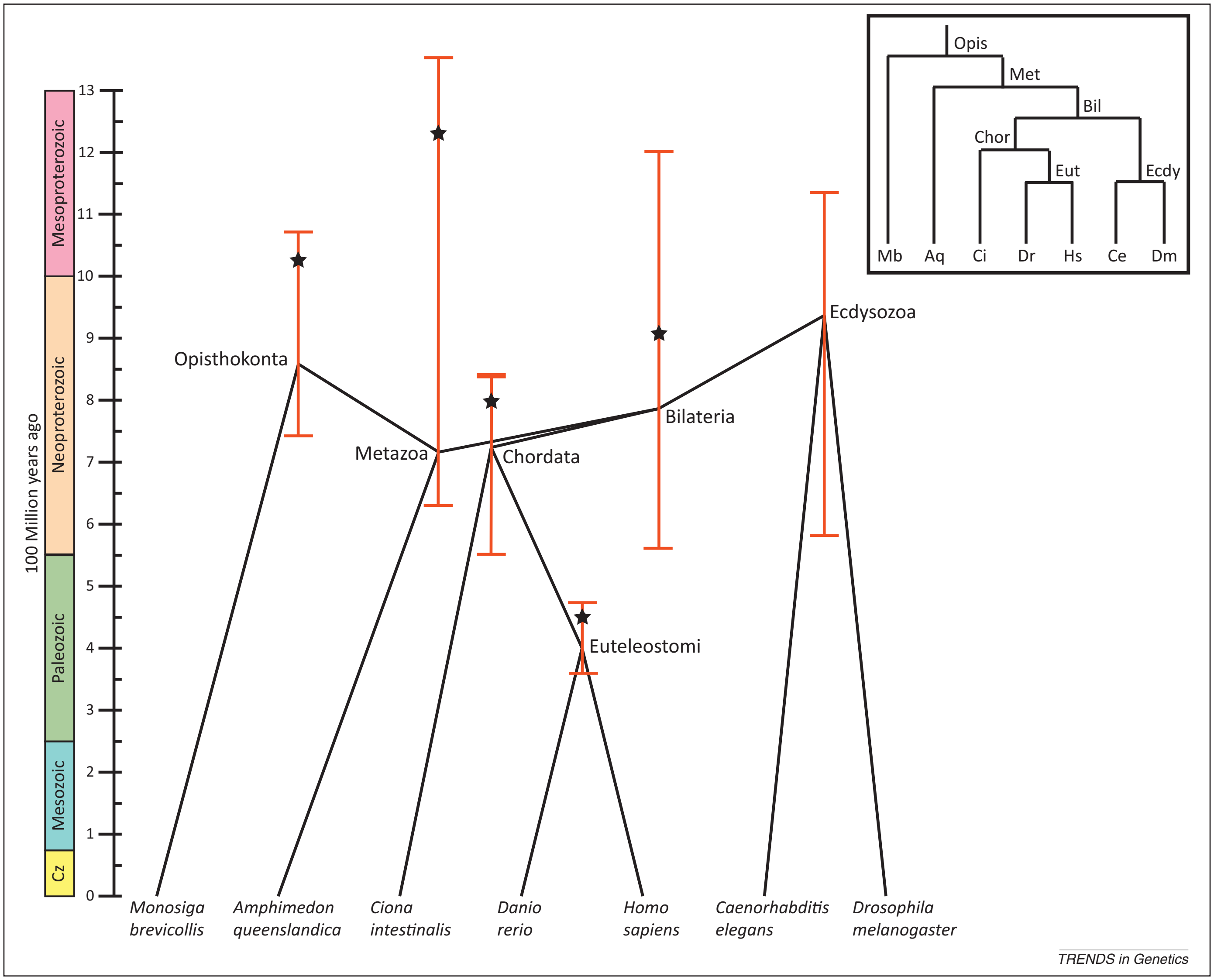


Figure 5. Estimates of species divergence times vary greatly. Metazoan species tree annotated with estimates of ancestral divergence times, obtained from the TimeTree database (http://timetree.org) using the species tree obtained from the National Center for Biotechnology Information (NCBI) taxonomy [72]. Nodes are plotted at the mean age estimate across all surveyed literature that contained that divergence [73]. The relative timing of these mean estimates violates the branching order of the commonly accepted tree (inset box), as can be seen from the distorted layout in which several branches appear to be traveling backwards in time. In the accepted tree, the Opisthokonta, Metazoa, and Bilateria nodes all pre-date Ecdysozoa and its sister node Chordata. This uncertainty is further reflected in the fact that minimum and maximum age estimates for each node (red bars) differ by hundreds of millions of years. To attempt to resolve these issues, an 'expert result' (starred) was selected for each node from a single article that was deemed to have the 'best' estimate for that divergence [45]. Across the tree, this expert result is consistently much older than the mean age estimate for the same node, indicating that there may be systematic underestimation of node ages in the literature. The expert results are more consistent with the branching order of the accepted tree, although they do not correctly place Opisthokonta earlier than Metazoa.

Statistical methods for identifying significant trends in the resulting evolutionary data are also needed. Development of these methods would enable studies of the underlying principles of gene evolution and delineation of the extent to which events in a particular gene family (e.g., an expansion) are temporally associated with events in other families.

With more genomes, the lines between phylogenetics and population genetics are increasingly blurred. In that context, gene age analysis can be extended to consider the age of alleles arising from recent mutations that are still polymorphic within a population. Allele-age analyses typically use polymorphism frequencies to estimate when mutations arose and whether they were subject to selection during a particular time period [51].

As more researchers embrace a more accurate, phylogenetic approach to gene age analysis, we must develop analytical tools that are accessible to researchers with a range of quantitative skills. Developing user-friendly software that supports every step of the process from gene tree inference, to gene age inference and interpretation of gene age history results should be a priority. Tools that will have the greatest impact are those that automate processing of large-scale data sets, visualization tools for exploratory analysis, and statistical tools for correcting systematic error, assessing significance, and extracting trends (e.g., ProteinHistorian [49]). With the availability of methods that can analyze terabytes of gene sequence data at the push of a button, it will be more important than ever to articulate clearly the strengths and weaknesses of alternative gene age analysis methods and to evaluate these in the context of each study.

Acknowledgments

D.D. and M.S. acknowledge support of the National Science Foundation (Grant #DBI1262593) and a computational facilities access grant (#MCB0000010P) from the Pittsburgh Supercomputing Center. K.S.P. and J.A.C. were supported by institutional funds from the J. David Gladstone Institutes and Vanderbilt University, respectively, and a grant from the National Heart, Lung and Blood Institute (Grant #HL098179).

References

1 Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 1313–1326

2 Domazet-Loso, T. and Tautz, D. (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468, 815–818

3 Alba, M.M. and Castresana, J. (2005) Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* 22, 598–606

4 Cai, J.J. *et al.* (2006) Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in Ascomycota. *J. Mol. Evol.* 63, 1–11

5 Wolf, Y.I. *et al.* (2009) The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7273–7280

6 Cai, J.J. *et al.* (2010) Broker genes in human disease. *Genome Biol. Evol.* 2, 815–825

7 Vishnoi, A. *et al.* (2010) Young proteins experience more variable selection pressures than old proteins. *Genome Res.* 20, 1574–1581

8 Qin, H. *et al.* (2003) Evolution of the yeast protein interaction network. *Proc. Natl. Acad. Sci. U.S.A.* 100, 12820–12824

9 Kim, W.K. and Marcotte, E.M. (2008) Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS Comput. Biol.* 4, e1000232

10 Capra, J.A. *et al.* (2010) Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol.* 11, R127

11 Milinkovitch, M.C. *et al.* (2010) Historical constraints on vertebrate genome evolution. *Genome Biol. Evol.* 2, 13–18

12 Nagaraj, S.H. *et al.* (2010) The interplay between evolution, regulation and tissue specificity in the human hereditary diseasome. *BMC Genomics* 11 (Suppl. 4), S23

13 Warnefors, M. and Eyre-Walker, A. (2011) The accumulation of gene regulation through time. *Genome Biol. Evol.* 3, 667–673

14 Dittmar, K. and Liberles, D.A. (2010) *Evolution after Gene Duplication*. Wiley-Blackwell

15 Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66

16 Liao, B.Y. and Zhang, J. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet.* 23, 378–381

17 Su, Z. and Gu, X. (2008) Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. *J. Mol. Evol.* 67, 705–709

18 Makino, T. *et al.* (2009) The complex relationship of gene duplication and essentiality. *Trends Genet.* 25, 152–155

19 Chen, W.H. *et al.* (2012) Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol. Biol. Evol.* 29, 1703–1706

20 Cai, J.J. *et al.* (2009) Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biol. Evol.* 1, 131–144

21 Domazet-Loso, T. and Tautz, D. (2010) Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in Metazoa. *BMC Biol.* 8, 66

22 Sakarya, O. *et al.* (2007) A post-synaptic scaffold at the origin of the Animal Kingdom. *PLoS ONE* 2, e506

23 Wheeler, D. *et al.* (2001) An orphaned mammalian beta-globin gene of ancient evolutionary origin. *Proc. Natl. Acad. Sci. U.S.A.* 98, 1101–1106

24 Durand, D. and Hoberman, R. (2006) Diagnosing duplications: can it be done? *Trends Genet.* 22, 156–164

25 Farris, J.S. (1977) Phylogenetic analysis under Dollo's Law. *Syst. Zool.* 26, 77–88

26 Domazet-Loso, T. *et al.* (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23, 533–539

27 Domazet-Loso, T. and Tautz, D. (2008) An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.* 25, 2699–2707

28 Swofford, D.L. and Maddison, W.P. (1987) Reconstructing ancestral character states under Wagner parsimony. *Math. Biosci.* 87, 199–229

29 Doyon, J.P. *et al.* (2011) Models, algorithms and programs for phylogeny reconciliation. *Brief. Bioinform.* 12, 392–400

30 Nakhleh, L. *et al.* (2009) Gene trees, species trees, and species networks. In *Meta-analysis and Combining Information in Genetics*, (Guerra, R. and Goldstein, D., eds), pp. 275–293, Chapman & Hall, CRC Press

31 Hirschey, M.D. *et al.* (2011) Sirt1 and Sirt3 deacetylate homologous substrates: Acecs1,2 and Hmgcs1,2. *Aging* 3, 635–642

32 Vernot, B. *et al.* (2008) Reconciliation with non-binary species trees. *J. Comput. Biol.* 15, 981–1006

33 Stolzer, M. *et al.* (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 28, i409–i415

34 Rasmussen, M.D. and Kellis, M. (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* 22, 755–765

35 Salichos, L. and Rokas, A. (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331

36 Csuros, M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26, 1910–1912

37 De Bie, T. *et al.* (2006) Cafe: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271

38 Akerborg, O. *et al.* (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5714–5719

39 Gorecki, P. *et al.* (2011) Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinformatics* 12 (Suppl. 1), S15

40 Funke, L. *et al.* (2005) Membrane-associated guanylate kinases regulate adhesion and plasticity at cell junctions. *Annu. Rev. Biochem.* 74, 219–245

41 de Mendoza, A. *et al.* (2010) Evolution of the Maguk protein gene family in premetazoan lineages. *BMC Evol. Biol.* 10, 93

42 Zheng, C.Y. *et al.* (2011) Maguks, synaptic development, and synaptic plasticity. *Neuroscientist* 17, 493–512

43 Eastwood, D.C. *et al.* (2011) The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. *Science* 333, 762–765

44 Floudas, D. *et al.* (2012) The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* 336, 1715–1719

45 Thomas, L., ed. (2012) Origin of coal. In *Coal Geology*, pp. 3–52, John Wiley & Sons Ltd, http://dx.doi.org/10.1002/9781118385685.

46 Harms, M.J. and Thornton, J.W. (2010) Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* 20, 360–366

47 Dean, A.M. and Thornton, J.W. (2007) Mechanistic approaches to the study of evolution: the functional synthesis. *Nat. Rev. Genet.* 8, 675–688

48 Robinson, R. (2012) Resurrecting an ancient enzyme to address gene duplication. *PLoS Biol.* 10, e1001447

49 Capra, J.A. *et al.* (2012) Proteinhistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS Comput. Biol.* 8, e1002567

50 Schrider, D.R. *et al.* (2009) All human-specific gene losses are present in the genome as pseudogenes. *J. Comput. Biol.* 16, 1419–1427

51 Kelley, J.L. (2012) Systematic underestimation of the age of selected alleles. *Front. Genet.* 3, 165

52 Yang, Z. and Rannala, B. (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* 23, 212–226

53 Elhaik, E. *et al.* (2006) The 'inverse relationship between evolutionary rate and age of mammalian genes' is an artifact of increased genetic distance with rate of evolution and time of divergence. *Mol. Biol. Evol.* 23, 1–3

54 Graur, D. and Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* 20, 80–86

55 Drummond, A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4, e88

56 Willerslev, E. and Cooper, A. (2005) Ancient DNA. *Proc. R. Soc. B* 272, 3–16

57 Alba, M.M. and Castresana, J. (2007) On homology searches by protein blast and the characterization of the age of genes. *BMC Evol. Biol.* 7, 53

58 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

59 Altschul, S.F. *et al.* (1997) Gapped Blast and Psi-Blast: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

60 Eddy, S.R. (2011) Accelerated profile hmm searches. *PLoS Comput. Biol.* 7, e1002195

61 Soding, J. *et al.* (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* 33, W244–W248

62 Murzin, A.G. *et al.* (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540

63 Sillitoe, I. *et al.* (2013) New functional families (Funfams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* 41, D490–D498

64 Winstanley, H.F. *et al.* (2005) How old is your fold? *Bioinformatics* 21 (Suppl. 1), i449–i458

65 Li, L. *et al.* (2003) Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189

66 Remm, M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052

67 Altenhoff, A.M. and Dessimoz, C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.* 5, e1000262

68 Salichos, L. and Rokas, A. (2011) Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS ONE* 6, e18755

69 Trachana, K. *et al.* (2011) Orthology prediction methods: a quality assessment using curated protein families. *Bioessays* 33, 769–780

70 Guindon, S. *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of Phyml 3.0. *Syst. Biol.* 59, 307–321

71 Katoh, K. and Standley, D.M. (2013) Mafft Multiple Sequence Alignment Software Version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780

72 Sayers, E.W. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 37, D5–D15

73 Kumar, S. and Hedges, S.B. (2011) Timetree2: species divergence times on the iPhone. *Bioinformatics* 27, 2023–2024