Review

Cell PRESS

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

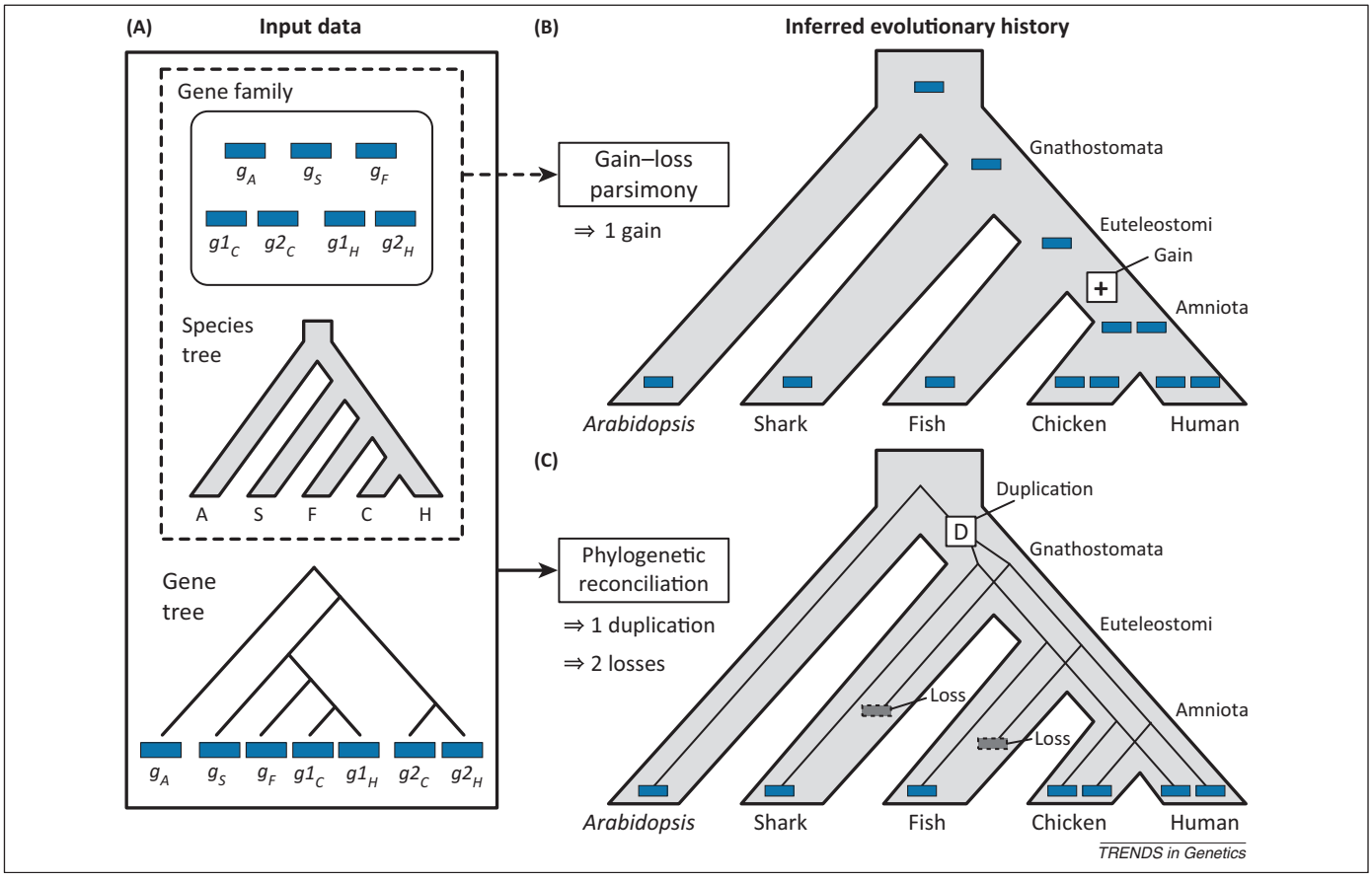

Figure 1. A typical error made by gain–loss methods is avoided with reconciliation. A gene family with a history of parallel losses illustrates the increased accuracy associated with explicit use of a gene tree by phylogenetic reconciliation. **(A)** A hypothetical gene family, based on the real enzyme 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 $g2_C$.