Review

CellPress

Coming of age: orphan genes in plants

Zebulun W. Arendsee, Ling Li, and Eve Syrkin Wurtele

Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA

Sizable minorities of protein-coding genes from every sequenced eukaryotic and prokaryotic genome are unique to the species. These so-called 'orphan genes' may evolve *de novo* from non-coding sequence or be derived from older coding material. They are often associated with environmental stress responses and species-specific traits or regulatory patterns. However, difficulties in studying genes where comparative analysis is impossible, and a bias towards broadly conserved genes, have resulted in underappreciation of their importance. We review here the identification, possible origins, evolutionary trends, and functions of orphans with an emphasis on their role in plant biology. We exemplify several evolutionary trends with an analysis of *Arabidopsis thaliana* and present *QQS* as a model orphan gene.

All species have a cadre of unique genes

Until the past few years the consensus was that new genes arise via combinations of processes such as duplication, fusion, fission, and transposition of existing protein-coding genes. Fischer and Eisenberg noticed that all sequenced bacteria contained genes without detectable homologs in any sequenced relative [1]. They postulated this uniqueness was a real phenomenon, rather than an artifact of poor annotation or sparse sequencing among nearby species, as some claimed [2]. Since the advent of next-generation sequencing, the analysis of a multitude of genomes has shown that such orphan genes are widespread across all domains of life [3–5] and viruses [6].

Continual genesis of novel genes in an organism where protein count is fairly constant implies equilibrium between gene origin and extinction. A reasonable hypothesis is that most of the turnover occurs in the youngest genes [7]. This has been demonstrated in *Drosophila* [8] and is reflected in the degree of conservation of existing genes in each genome. Under this model there is a vast, dynamic reservoir of novel genes. We will discuss: (i) the origin of orphans and their regulatory elements, (ii) their maturation into established genes, and (iii) the functions into which they are recruited.

Orphans originate in diverse ways

Orphans may be defined as genes with coding sequences utterly unique to the species; in other words, genes that produce previously non-existing (novel) proteins. They are a subset of taxonomically restricted (also called lineage-specific) genes that are specific to a particular taxon (e.g., malvid-specific or Brassicaceae-specific genes). Genes are generally classified as being orphans if they lack coding-sequence similarity outside their species (usually quantified by BLAST). This classification method accepts as orphans, genes that are newly born from non-genic sequence, as well as descendants of ancient genes whose coding sequences have changed beyond recognition; it rejects horizontally transferred genes and duplicated genes that may have assumed a new function but whose proteins are still recognizable (i.e., 'new' genes that are not orphans).

Analysis of the genomic contexts and sequences of orphan genes can often reveal their origins, as reviewed in [7]. Some can be traced to highly divergent products of gene duplications, overlapping or anti-sense reading frames (overprinting), domesticated transposons, resurrected pseudogenes, or early frameshift mutations [8–12]. Others may arise *de novo* from non-coding sequence. Early doubts that protein-coding genes could spontaneously arise [13] have been put to rest by a flood of papers tracing orphans to their non-genic roots [14–23]. A recent study suggests a continuum between very weakly transcribed and translated open reading frames (ORFs) and highly functional, mature genes [24]. Table 2 from [10] shows a cross-species, quantitative comparison of the origins of orphan genes. In *A. thaliana*, over half of the orphans appear to have arisen *de novo*, based on similarity to non-genic regions of *Arabidopsis lyrata* [9].

Estimates of the percentage of genes that are orphans in various species ranges wildly from <1–71% [5,9,10,24–33], with 5–15% being fairly typical [10,31,33]. A portion of this disparity is attributable to the varying evolutionary distance between each focal species and its nearest sequenced relatives [10]. Other sources of variation are the quality of the genome datasets and the methods used in orphan identification (e.g., three independent studies of *A. thaliana* report 958, 1324, and 1430 orphans, respectively [9,34,35]). However, much of the variation reported is likely due to real differences in evolutionary pressures and molecular genetic phenomena that are as yet unknown.

Phylostratigraphy classifies genes by age

Genes can be stratified by age via a technique known as phylostratigraphy that traces modern genes back to their orphan founders [36]. Figure 1 shows a phylostratigraph of the protein-coding genes of *A. thaliana*. The general approach is to select hierarchical taxonomic groups ascending from the focal species, and for each gene find the oldest

Corresponding authors: Li, L. (liling@iastate.edu); Wurtele, E.S. (mash@iastate.edu).

Keywords: orphan; phylostratigraphy; *Arabidopsis*; QQS; cysteine-rich secretory proteins.

1360-1385/

 http://dx.doi.org/10.1016/j.tplants.2014.07.003

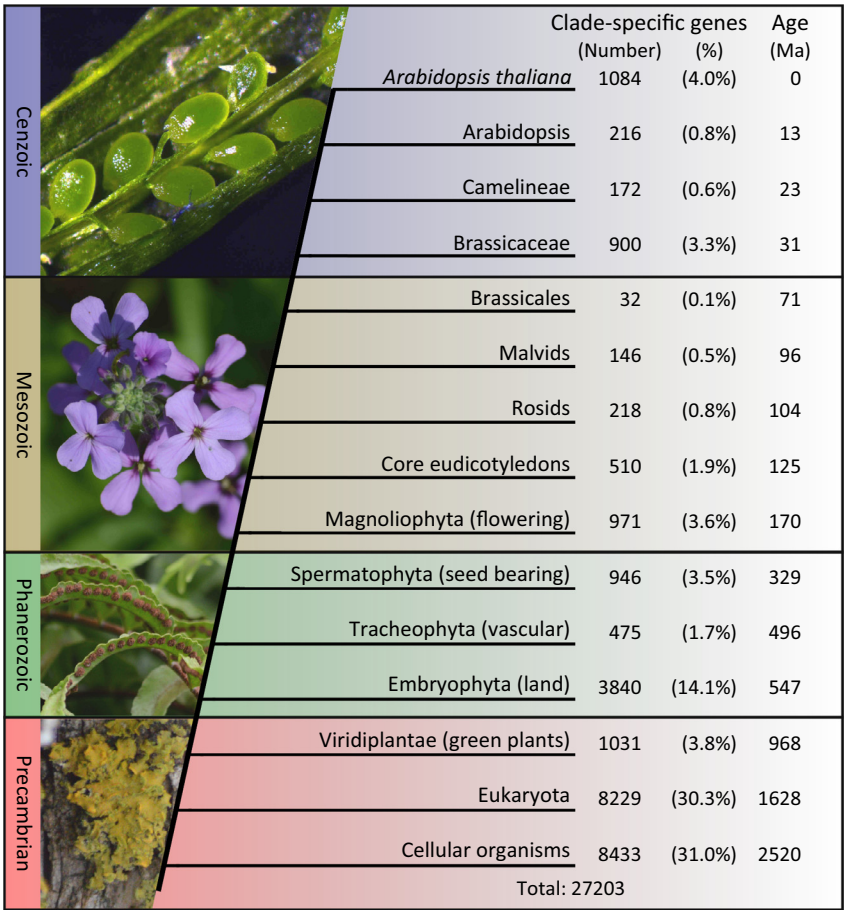

Figure 1. Age stratification of all genes in *Arabidopsis thaliana*. Each gene is assigned to the oldest clade (or 'phylostratum') that contains a homolog, as inferred by a protein-BLAST of each *A. thaliana* gene against a selected set of genomes (Table S1 in the supplementary material online) with a threshold *e* value of 10^{-5}. All *A. thaliana* genes are from the TAIR10 release (in cases where one locus has multiple gene models, we used the first). Organelle genomes were not included in the analysis. Age is in Ma (millions of years ago) and refers to the estimated time since the diversification of the clade from its most recent common ancestor. For references for the age assignments and a list of genomes searched in each phylostratum, see Table 1. For complete searchable phylostratum assignments, see AtGeneSearch (http://www.metnetdb.org/MetNet_atGeneSearch.htm).

taxon in which it has a homolog [36]. By describing the characteristics of increasingly ancient phylostrata, the path from genomic noise to mature protein is revealed.

Conventional phylostratigraphic analyses make two major assumptions. The first is that simple search algorithms (such as protein-BLAST) are adequate for the identification of distant homologs. This assumption is supported by evolutionary simulations [37]. However, a recent study of viral orphan genes that used more sensitive algorithms (PSI-BLAST, HHBlits, and HHPred) predicted homologs for about a quarter of genes that had been identified as genus-specific by protein-BLAST [38]. A second assumption is that the oldest components of genes are no older than the gene founders. This assumption is violated if an old domain or exon is incorporated into a young protein. This issue has been noted previously, but was considered not to be a serious impediment, at least not in metazoans [7]. A recent review acknowledges the successes of phylostratigraphy [36,39,40] but argues that phylogenetic reconciliation methods offer a more nuanced understanding of the events underlying gene histories [41].

What are the prospects of a young orphan gene? Phylostratigraphic analyses indicate that some orphans survive to fixation. These manifest as gene families that are taxonomically restricted to the clade descending from the species in which they arose. Genes from older phylostrata tend towards greater length, complexity, and connectivity. Figure 2 includes an overview of several cross-phylostrata traits in *A. thaliana* genes, and compares them to non-genic ORFs.

Orphans mature with time

A steady, several-fold increase in protein length from species-specific genes to universally conserved genes has been noted in several metazoans [11,42], yeast [24], and *A. thaliana* [35] (Figure 2A). This is largely due to an increase in the number of exons because the average exon length is somewhat constant, as seen in metazoa [11] and in the older *A. thaliana* phylostrata (Figure 2B). Although in some species (such as rice, zebrafish, and humans) genes from recent phylostrata have particularly long exons [11,43], in *A. thaliana* there is a significant increase in exon size (about twofold) across the first several phylostrata.

By several criteria, younger genes are more random and specialize over time. For example, amino acid composition bias increases with age in yeast [24] and many bacteria