The evolutionary origin of orphan genes

Diethard Tautz and Tomislav Domazet-Lošo‡§*

Abstract | Gene evolution has long been thought to be primarily driven by duplication and rearrangement mechanisms. However, every evolutionary lineage harbours orphan genes that lack homologues in other lineages and whose evolutionary origin is only poorly understood. Orphan genes might arise from duplication and rearrangement processes followed by fast divergence; however, *de novo* evolution out of non-coding genomic regions is emerging as an important additional mechanism. This process appears to provide raw material continuously for the evolution of new gene functions, which can become relevant for lineage-specific adaptations.

Orphan genes
Genes that lack homologues in other lineages — that is, they cannot be linked by overall similarity or shared domains to genes or gene families known from other organisms.

**Max-Planck Institut für Evolutionsbiologie, August-Thienemannstrasse 2, 24306 Plön, Germany.*
‡Zoological Institute, Christian Albrechts University Kiel, 24105 Kiel, Germany.
§Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, P.P. 180, 10002 Zagreb, Croatia.
Correspondence to D.T.
e-mail: tautz@evolbio.mpg.de
doi:10.1038/nrg3053
Published online
31 August 2011

Susumu Ohno's[1] passionate support for gene (and genome) duplication as the major mechanism for creating evolutionary novelty has, for a long time, set the stage for thinking about gene emergence. The model was further elaborated by François Jacob[2], who portrayed evolution as a 'tinkerer' and specifically concluded that the "creation of entirely new nucleotide sequences could not be of any importance in the production of new information". In another seminal paper, King and Wilson[3] pointed out that the stark contrast between the rate of organismal and molecular evolution hints at the prevalence of regulatory changes, rather than the evolution of new gene functions, in diversifying lineages. Many of these early predictions are well supported by data. Gene duplications and regulatory evolution have an overwhelming function in the diversification of genomes and species (reviewed in REFS 4–7). The role of tinkering is also well supported in that fragments of existing genes can be transferred to other genes to create new functions[8]. But is this all? Was there only one time in evolution in which all building blocks of genes originated, and were these subsequently shuffled and mixed to create novelties? We argue here that the existence of orphan genes — so called because they lack homologues in other lineages — tells another story. The evolutionary origin of such genes is still unclear, even though they represent up to one-third of the genes in all genomes, including those of bacteria, archaea and phages. Orphan genes are thought to be particularly important for taxon-specific developmental adaptations and interactions with the environment[9]. Although they were already detected in the wake of the first genome projects, it is only now becoming possible to trace their origins within the framework of many available genome sequences covering a broad phylogenetic distribution.

This Review first discusses the systematic identification of orphan genes and puts this into the context of what is generally known about gene emergence through duplication and rearrangement processes. For more detail on these topics, we refer the reader to some excellent reviews[7,9,10]. We then evaluate the mechanisms of how orphan genes could emerge and what is known about the evolutionary dynamics of genes in general and orphan genes in particular. We describe accumulating evidence that *de novo* evolution of genes from non-coding sequences could have an important role in creating orphan genes and new gene functions. This leads to the general question of how new functional protein domains might evolve out of random sequences. We propose that orphan genes continuously arise in any genome and that, at least in eukaryotes, they arise mostly through *de novo* evolution; however, we also propose that only a fraction of them assumes a long-term role in their respective evolutionary lineage, mainly in the context of evolutionary radiations.

Identifying orphan genes

Orphan genes were first discussed in the context of the yeast genome-sequencing project, which suggested that approximately one-third of the identified genes fell into this category[11,12]. They were then also found in the first bacterial genomes to be sequenced. Fischer and Eisenberg[13] introduced the synonymous term ORFans for them, which is preferentially used in the microbial literature. Comparative genomics (BOX 1) has now shown that orphans are a universal feature of any genome[14],

Box 1 | Comparative genomics for the identification of orphan genes

In comparative genomics, one usually starts with a well-annotated genome (the focal species) and retrieves homologues of all genes from other genomes using the Basic Local Alignment Search Tool (BLAST)[16]. The BLAST algorithm provides a statistical measure (the expected value (*E*)) for the probability that a match could be generated by chance alone, which implies that one has to set a cutoff value beyond which a match is considered to be nonrandom. A cutoff of $E < 10^{-3}$–10^{-4} for protein similarity searches was found to satisfy the best balance between not missing true homologues and including random matches[55]. Many orphans contain low-complexity regions or fragments of transposable elements, which could yield spurious matches elsewhere in the genome or in related genomes. This can be avoided by the use of a repeat masker during BLAST.

Albà and Castresana[19] have studied whether BLAST is a suitable procedure to detect all true homologues in other genomes. They find that a problem would only occur for genes that evolve quickly and that diverge homogeneously across their whole sequence length. Reassuringly, short stretches of higher conservation, such as small functional domains, are readily recovered. As one can expect that most proteins will show such heterogeneity in evolutionary rates across their sequence, it seems safe to conclude that most remote homologues would be picked up by BLAST if they exist.

When one deals with recently evolved orphan genes, it may be of interest to confirm the functional status by investigating the intactness of the reading frame in closely related species. This also allows evolutionary rates to be determined by calculating the ratio of replacement substitutions (Ka) to silent substitutions (Ks). A Ka/Ks ratio of 1 implies neutral evolution, a ratio below 1 implies purifying selection, and a ratio above 1 implies positive selection.

For non-coding RNA genes, it is necessary to obtain EST evidence or other functional evidence to ascertain that the presumed gene is not only an annotation artefact.

including phage genomes[15]. This implies that every major taxonomic lineage includes a fraction of genes that is restricted to this group — hence their alternative name 'taxonomically restricted genes' (TRGs)[14]. It also implies that all genes that have no homologue in the common unicellular ancestor are orphan genes in the lineage from which they have arisen. As orphan genes represent a substantial fraction of every extant genome, the total number of orphans across all evolutionary lineages by far exceeds the number of known gene families. It is therefore timely to ask how such completely new genes can originate.

The working definition of an orphan gene as sharing no similarity with genes or protein domains in other evolutionary lineages is conceptually simple but operationally complex. The identification of orphans depends both on the detection method and the reference set of genomes considered, as this defines the evolutionary lineage to be investigated. The detection method is crucial for distinguishing between possible homologues and spurious matches. The reference set is important, because different results are obtained if one compares sequences between closely related species or between major phylogenetic lineages. Sequence similarity therefore needs to be assessed in parallel to the phylogenetic context.

Similarity searches. The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences[16]. It is the preferred method for detecting homologues of genes in other species. Because the classification of a gene as an orphan requires that it is not present in other evolutionary lineages, it is crucial to know whether a gene is truly absent in species belonging to other lineages or whether this absence is due to the technical limitations of the approach used to detect it. This question was intensively discussed[17–19], and it was concluded that most remote homologues would be picked up by BLAST if they existed[19] (BOX 1). Several further studies have re-assessed this conclusion[20,21] and asserted that, indeed, BLAST performs well in this respect.

Although BLAST performs well on average, it is possible that some genes have diverged beyond the point at which they can be found by BLAST. A more sensitive tool for detecting even such remote homologues is Position-Specific Iterated BLAST (PSI-BLAST)[22]. This procedure uses alignment information from more closely related homologues to develop a profile of the most conserved residues, which is then used to retrieve additional candidate homologous genes or domains to further improve the profile. However, this procedure is not well suited to large-scale analysis, as it requires careful manual supervision and carries a risk that one eventually traces convergently evolved gene families (see below). It can, however, identify connections between gene families that would otherwise have been missed.

Phylogenetic context. The origin and age of a gene can only be accurately determined within an appropriate phylogenetic framework. A simple approach would be to use different phyla; for example, the presence of a gene could be compared between a representative of protostomes (such as *Drosophila melanogaster*) and deuterostomes (such as humans). However, this approach would not allow determination of whether a gene only evolved within one lineage — for example, whether it evolved in the *Drosophila* lineage only and not in other protostomes (such as beetles or nematodes) — or whether it was lost in certain lineages (FIG. 1a). Given that we now have a large number of fully sequenced genomes from various taxa, it has become possible to do such assignments in an extended phylogenetic context. This allows 'founder genes' — that is, those orphan genes that give rise to new gene families in descendant lineages[23] — to be allocated to certain time points in evolution. We have developed a general framework, the so-called 'phylostratigraphy' (REF. 23) (FIG. 1b), within which this can be combined with a statistical evaluation of macroevolutionary trends[23–25]. Phylostratigraphy can be used to systematically identify all orphan genes within the evolutionary lineages that have led to a particular extant genome.

Emergence of new genes

There are many mechanisms by which new genes could emerge, most of which are related in some form to duplication and/or transposition mechanisms. The products of such duplications will usually be traceable as paralogues or as members of gene families and would thus not be called orphan genes. With respect to orphan gene emergence, it is important to consider how founder genes emerge. For these, one can envisage two main models: first, duplication followed by divergence beyond the threshold of similarity searches and, second, *de novo* evolution from non-coding regions. In addition, one has to consider how new genes acquire regulatory sequences. Below we discuss these topics in turn.

Purifying selection
The removal of deleterious mutations through natural selection.

Basic Local Alignment Search Tool
(BLAST). A program that compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.

Protostomes
The animal superphylum that includes nematodes (for example, *Caenorhabditis elegans*) and arthropods (for example, *Drosophila melanogaster*).

Deuterostomes
The animal superphylum that includes vertebrates (for example, zebrafish) and mammals (for example, humans).

Founder genes
The phylogenetically oldest genes forming the basis of a new gene lineage, new protein domain or new gene family. The origin of founder genes is expected to correlate with evolution of functional novelty.

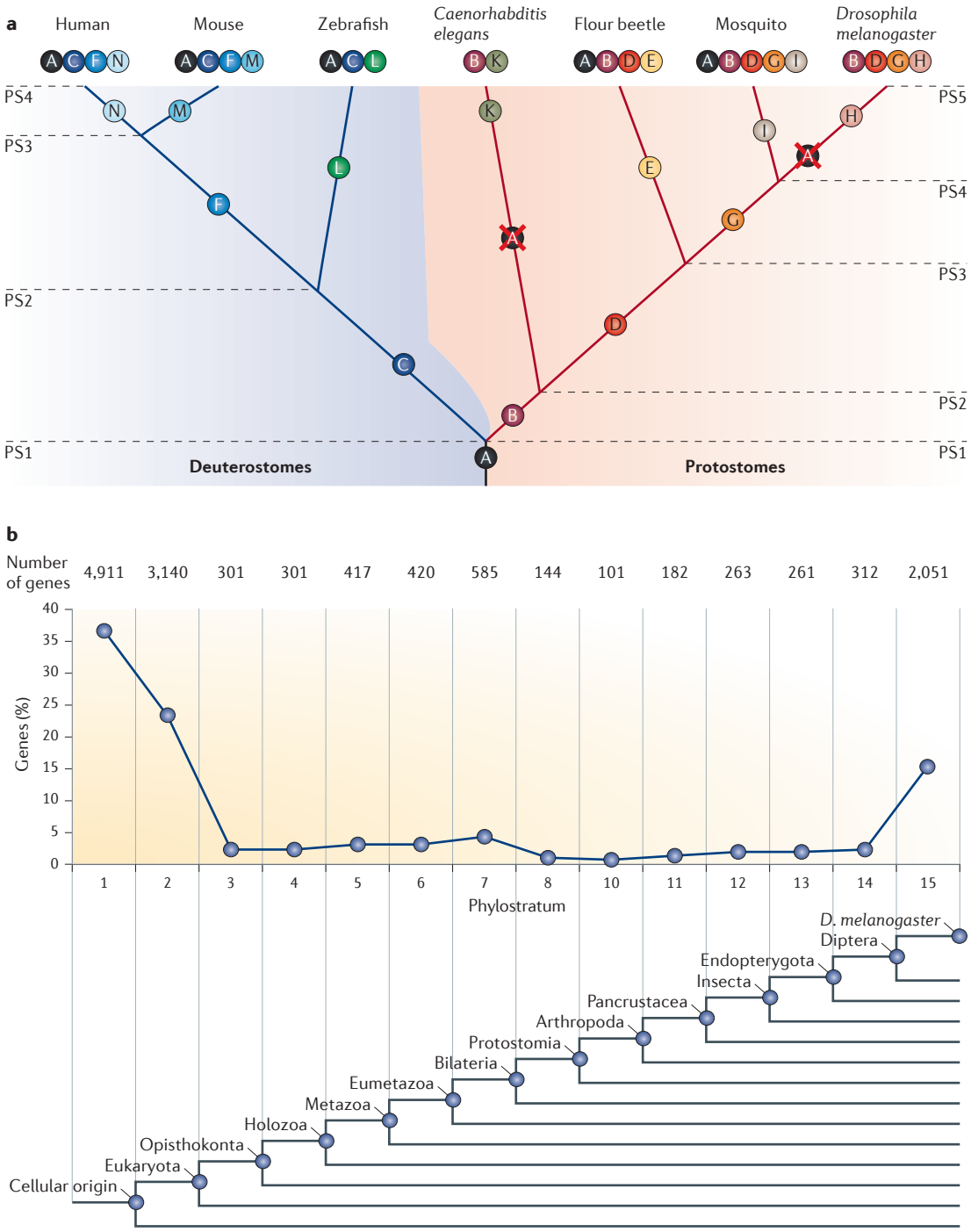


Figure 1 | **Identification of orphan genes in a phylogenetic context. a** | Example tree including taxa from deuterostomes and protostomes. Symbolic gene emergences are indicated by circles with letters, and gene losses are indicated by crossed circles. Nodes leading to humans (left) and to *Drosophila melanogaster* (right) are labelled with PS1 to PS4 and PS1 to PS5, respectively — where PS stands for phylostratum — to symbolize that the number of nodes determines the number of phylostrata that are available for a given taxon[23]. Note that only phylostratum 1 is shared between protostomes and deuterostomes, whereas other phylostrata are specific to each lineage. The genes that would be found for each taxon in this scheme are listed below the species names. In the case of gene 'A', one would conclude that it is specific to deuterostomes if one had only *Caenorhabditis elegans* and *D. melanogaster* available for comparison. However, if one includes the flour beetle and mosquito, one finds that the gene is present in protostomes as well — that is, that A is an ancestral gene. **b** | Example of a complete phylostratigraphy with *D. melanogaster* as the focal species, including the number of genes that were found in each phylostratum (listed at the top). Note that in absolute terms, the largest number of genes has arisen in the lowest phylostrata, but this is an artefact of the low phylogenetic resolution that is available at this level. Gene emergences scaled to time are plotted in FIG. 4. Panel **b** is modified, with permission, from REF. 23 © Elsevier Science.

Phylostratigraphy
A systematic procedure to identify the origin of genes within a comparative framework of fully sequenced genomes at multiple levels of the phylogenetic hierarchy (the phylostrata).

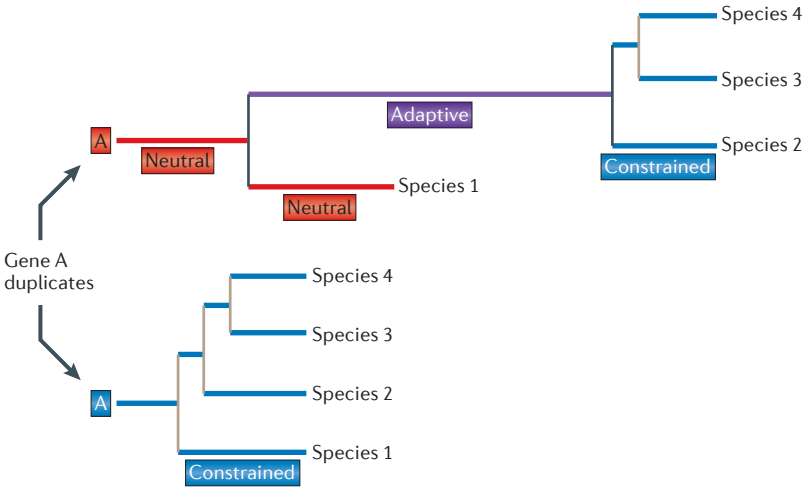


Figure 2 | **Duplication–divergence model for orphan gene evolution.** The model[55] assumes that a gene is duplicated and that one copy is then free to diverge (upper part, red), including undergoing further speciation events, whereas the other copy is not (lower part, blue). The different branch lengths symbolize the different evolutionary rates. The non-constrained copy would initially diverge with a neutral rate (it would accumulate random mutations). The freely diverging copy could assume a new function after a speciation event and would go through an adaptive phase in the respective lineage (species 2, 3 and 4). It would diverge beyond a point at which it can still be recognized by the Basic Local Alignment Search Tool (BLAST) and would thus become an orphan gene. After the adaptive phase, it would probably enter a new constrained phase. Many variations of this model are possible. The duplication could be linked to a structural rearrangement, or it might be accompanied by the insertion of transposable elements, which could form part of the gene. Alternatively, the original reading frame may be lost during the phase of neutral evolution, and a completely new reading frame may become the target during the selection phase.

General duplication mechanisms. Two main processes exist to explain how genes can become fully or partially duplicated within a genome. One is recombination, and the other is related to virus or transposon activity, whereby fragments of the viruses or transposons may contribute to the new genes by themselves. In addition, genes may become duplicated between genomes through horizontal gene transfer (HGT).

Recombination-mediated mechanisms can duplicate whole genes, including all or part of their regulatory sequences. These events can be facilitated by the presence of repetitive elements, which may serve as substrates for unequal crossovers. This can lead to tandem duplications, which by themselves can induce further unequal crossovers[7,10]. It has long been thought that only a handful of gene families — such as ribosomal genes or histone gene clusters — exist in arrays of tandem duplicates; however, it is now clear that many genes are subject to copy number variation and so provide an important source of evolutionary variation[6,26].

Transposon- and virus-associated activities can also shape genes in many ways, including by the expansion of genes through the insertion of whole or partial transposable elements, as well as by copying genes and moving them around. Reverse transcriptase produced by retroviruses or retrotransposons can copy any mRNA into DNA, which can then be inserted back into the genome[27,28]. This can lead to intron-free duplicates, which lack most regulatory elements. Alternatively, parts of genes can be transferred to other positions in the genome. If this new location happens to be the intron of another gene, a whole new exon may become integrated into the recipient gene. Indeed, it was initially speculated that exons often carry functional protein domains and that such exon-shuffling mechanisms may explain the whole diversity of proteins[29]. However, comparative genomic analysis now suggests that exon-shuffling mechanisms are restricted to certain gene families and have predominantly occurred during particular evolutionary phases[30,31]. Also, it was shown that most domain gains in metazoan genes might have occurred through gene fusion between adjacent genes[32].

HGT can be interpreted as an extended duplication mechanism, whereby the respective gene derives from another, as yet unknown lineage. HGT may frequently occur in bacteria and archaea, in which it appears to be mediated by viruses, plasmids and transposons[33]. In metazoans, the role of HGT seems to be more limited, although a number of specific cases have been documented[34,35]. If HGT occurs, one would evidently misclassify the time of origin for the respective gene, as it could have been present in the donor lineage for a long time before HGT. However, knowing that HGT has occurred does not address the question of how the gene first arose. Thus, although it is likely that a number of orphan genes will turn out to be products of HGT, the challenge of understanding the origin of these genes is not solved by understanding HGT.

Duplication–divergence. In the duplication–divergence scenario for orphan gene evolution, a new gene would initially be created through a gene duplication or transposition event and would then go through a phase of fast adaptive evolution, during which time it would lose all similarity to the gene from which it was duplicated (FIG. 2). Although this model fits well with what is generally known about duplication mechanisms, it has two major limitations. First, it is difficult to formulate a process by which natural selection can sequester one member of a duplicate pair for evolutionary exploration while retaining the other for the maintenance (and/or improvement) of the ancestral function[36,37]. Second, the divergence of a protein to a point at which it can no longer be detected by BLAST requires extensive substitutions along the entire length of the protein; this would occur only rarely, as many genes contain a functional protein domain that cannot easily be changed by mutations[19]. However, modifications to the basic model have been proposed. The initial duplication could be combined with a rearrangement or transposon insertion, in which case the original ORF could be changed so much that it could directly become subject to further evolution. Transposons are indeed frequently found to form parts of new genes[7,9,10]. Alternatively, if the original ORF became inactivated by a mutation after the duplication, a new reading frame could be used that could give rise to an entirely new protein. Although such a specific case is not yet known, one fast-evolving gene has been described that uses two reading frames from the same transcript that code for two interacting proteins[38].

Horizontal gene transfer (HGT). The exchange of genes between different evolutionary lineages.

Retrotransposons
Transposons that require an RNA intermediate for their transposition.

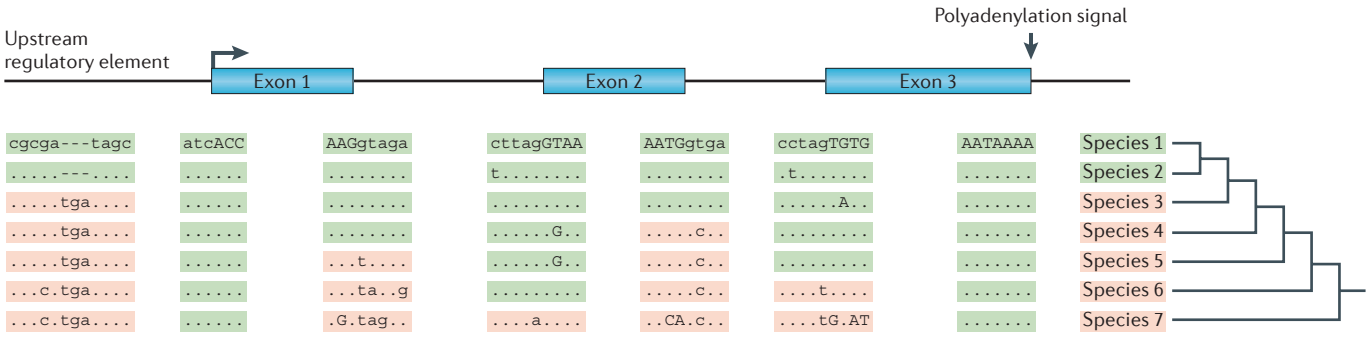


Figure 3 | ***De novo* evolution model for orphan genes.** A functional gene could evolve out of a non-coding sequence based on the cryptic presence of functional sites[41,44]. The functional sites depicted in the figure are an upstream regulatory element, the transcriptional start site (arrow), the exon junctions (exon sequences are given in capitals) and a polyadenylation signal. Some of these regions (green boxes) are also present in the outgroup species (species 3 to species 7), but these do not express the transcript, as at least one element is missing (pink boxes). Species 1 and species 2 express the transcript, as they share a modification in the upstream regulatory element, which makes it functional and thus allows specific transcription. This example is modelled according to a real case in mice[41] but is presented in an abstract form. Dots denote nucleotides that are identical to the top sequence, whereas dashes denote deletions.

De novo *evolution.* In this scenario, randomly occurring sequence combinations would form cryptic functional sites (for example, transcription initiation regions, splice sites and polyadenylation sites) and would come under a regulatory control to produce a distinct processed RNA transcript (FIG. 3). This RNA could initially function as an antisense or structural RNA[39] and would eventually acquire a functional ORF from which a completely new protein could evolve. The most stringent criterion for indicating the involvement of this mechanism requires that the corresponding genomic region of the gene is present in outgroup organisms, but as a non-coding stretch that is neither transcribed nor translated. Although this possibility for the emergence of new gene functions initially seemed the least likely[2], there are now a number of fully documented cases supporting *de novo* origination by this mechanism[40–44] (BOX 2). In addition, several surveys identified many more candidates for possible *de novo* evolved genes in various species[45–49].

Regulatory element emergence. A new gene not only requires a functional transcript but also requires regulatory control sequences. Binding sites of transcription factors are composed of partially redundant stretches of a few nucleotides only and therefore often occur by chance in random sequences. However, such single binding sites are not expected to lead directly to transcription. At least in eukaryotes, an active regulatory region usually requires clusters of binding sites within an enhancer, a functional transcription start site (promoter) and an open chromatin state, which, in combination, may not frequently arise by chance alone. But as enhancers can work at long distances, they might easily become recruited to novel genes — that is, in many cases, only a functional transcription start site would be required. Indeed, there is evidence that large parts of the genome are transcribed by such spurious mechanisms[39,50]. Also, a number of *de novo* genes have been detected within introns of other genes or as antisense transcripts and are likely to share enhancer activity with the gene in which they reside.

Another possible explanation for the transition from non-coding DNA to functional transcript is given by the 'out of testis' hypothesis[28]. The chromatin of meiotic and postmeiotic cells in the testis is in a state of hyper-transcription, and rather simple promoter elements are sufficient to drive transcription[51,52]. Single mutational changes might therefore lead to an active promoter, as has been suggested for a *de novo* evolved gene in mice[41]. Also, because sexual selection mechanisms work particularly effectively in the testis, new genes expressed in this tissue might be subject to more rapid positive selection than in other tissues[52]. There is indeed a strong tendency for newly evolved genes to be expressed in the testis first[45,46], which allows them to become functionally stabilized and eventually to acquire additional regulatory elements for expression in other tissues.

Evolutionary dynamics of genes

The large-scale sequencing of whole genomes has permitted increasingly refined analyses of general trends in protein evolution. For example, attention has focused on relating rates of protein divergence to patterns of selection and constraint. Other studies have concentrated on correlating these divergence rates with other patterns, such as gene expression level, dispensability, protein abundance, gene length or network connections[53]. These analyses have not usually made an explicit distinction between orphans and non-orphans but have classified genes into age classes, which is the equivalent of stating that the members of the younger class lack homologues in other lineages — that is, they are orphans. We first discuss what is known about general trends in gene evolution, and then we address the rates of gene gain and loss over time. Finally, we ask what is known about rates of *de novo* evolution of genes and why these may have been underestimated so far.

Sexual selection
A form of selection that arises from the interaction between the sexes and their gametes rather than from interactions with the environment.

Positive selection
The increase in frequency and fixation of alleles that contributes to the fitness of an organism.

Box 2 | **Examples of unequivocally identified *de novo* evolved genes**

The most stringent criteria for a *de novo* evolved gene require evidence that a genomic region that is non-coding in multiple outgroup species is transcribed into a distinct RNA in the ingroup or focal species and that this RNA is translated or can be shown to be functional. Ideally, one should show that a non-coding RNA in the outgroup species has assumed a functional ORF in the ingroup species. The following examples fulfil these criteria.

BSC4* in *Saccharomyces cerevisiae
The gene is expressed as a non-coding RNA in closely related species. It evolved an ORF encoding a peptide of 132 amino acids in length in *S. cerevisiae*. The new protein appears to be involved in DNA repair. Its functionality is supported by population genetics, expression, proteomics and synthetic lethal data[40].

Pldi* in *Mus musculus
The *Pldi* gene has three exons; it is specifically expressed in the testis, and it evolved about 3 million years ago. The gene region was subject to selective sweeps in some populations. Although short ORFs are present, the gene is most likely to act as a non-coding RNA that is involved in chromatin organization. Knocking out this gene leads to a lowered mobility of sperm cells[41].

***CLLU1*, *C22ORF45* and *DNAH10OS* in humans**
These genes were identified as human-specific genes that have syntenic, non-transcribed regions in other primates. Functional evidence for *CLLU1*, *C22ORF45* and *DNAH10OS* comes from polymorphism data, expression data and proteomics. *CLLU1* was first identified as an upregulated gene in chronic lymphocytic leukaemia[42].

MDF1* in *Saccharomyces cerevisiae
MDF1 originated *de novo* from a previously non-coding sequence. It functions as a suppressor of mating efficiency in a rich medium by binding MATα2 and thus promotes vegetative growth. It is regulated by the antisense gene *ADF1*, which acts as a transcriptional suppressor[43].

***FLJ33706* in humans**
FLJ33706 is a six-exon gene with a human-specific ORF of 194 amino acids that evolved from a non-coding region that is generally present in eutherian mammals. The first exon and some splice junctions were partially created by the insertion of *Alu* elements. The RNA and protein are expressed in the brain, and elevated expression is observed in Alzheimer's brain samples. Polymorphism data support functionality of the reading frame[44].

General trends. An assessment of protein-coding genes in major taxa revealed an approximately log-normal distribution of divergence rates across three orders of magnitude[20]. On average, younger genes have higher divergence rates[17,20,21,54] but retain the log-normal distribution[21] — that is, recently evolved genes can show the whole range of divergence rates, including low rates of evolution[55]. Another consistent trend is that slowly evolving genes are more highly expressed than fast-evolving ones[20,21,56–59], implying that younger genes show lower expression on average[20,21,60]. Old genes also tend to encode longer proteins than younger ones do[21,61]. Interestingly, however, it is more difficult to show a consistent correlation between the relative functional importance of old versus new genes[20,53,62–64], although, on average, older genes are indeed implicated in more general functions[53]. One factor that could explain many of these general trends is the cost associated with translation errors and protein misfolding[65]. This could indeed be an important evolutionary constraint that limits the number of genes that can be maintained in an organism, but it does not make predictions about how individual gene classes evolve and contribute to adaptations.

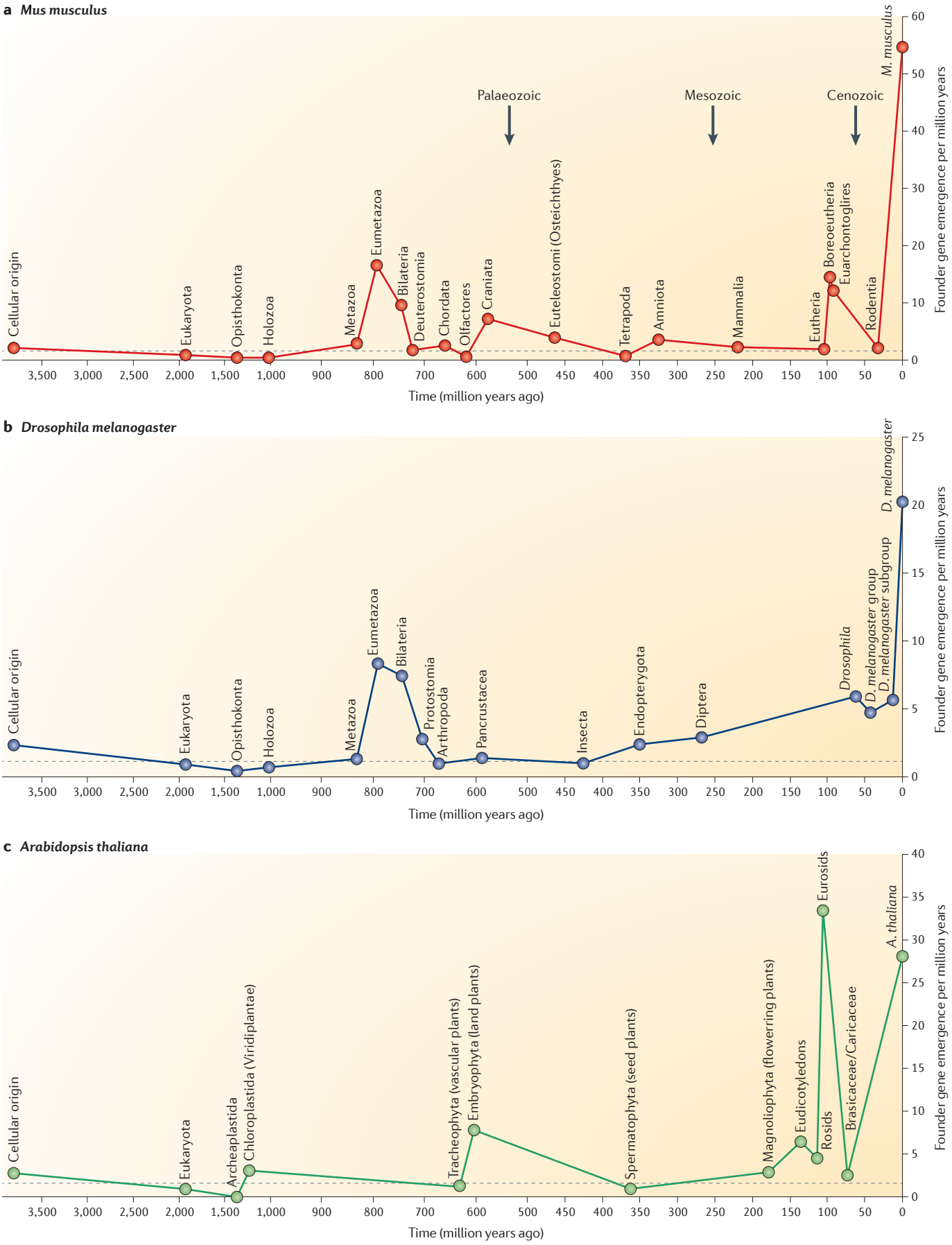
Gain and loss rates. Gene numbers in all higher organisms sequenced to date lie between 20,000 and 50,000 — that is, within a narrow range. By contrast, genome sizes can differ by several orders of magnitude. This implies that there is a balance between gene emergence and gene loss over time. Systematic studies of these gene-loss dynamics show that the propensity of gene loss is negatively correlated to gene expression levels and positively correlated to dispensability — that is, older genes are less likely to be lost than younger ones[21,66,67]. Hence, although genes may be born at a high rate, most of them are also quickly lost.

One can use phylostratigraphy to estimate the rate of gene emergence over evolutionary time by dividing the number of genes found in a given phylostratum by the respective time period for this phylostratum[23]. FIGURE 4 shows this calculation for the lineages leading to mice, *Drosophila* and *Arabidopsis*. Interestingly, one can see several peaks of gene emergence at certain time periods in these plots. In animals, the most conspicuous peak occurs long before the major radiation of animals (FIG. 4a,b), roughly corresponding to the time when the Earth went through freezing cycles[68]. In the mouse lineage, one sees additional peaks, which slightly predate the other three major phases of animal radiation (FIG. 4a). These peaks are also present in the *Drosophila* lineage, but they are less pronounced; interestingly, however, the rate of gene emergence continues to grow in younger lineages leading to *D. melanogaster* (FIG. 4b). The most conspicuous peak in the *Arabidopsis* lineage coincides with a plant-specific radiation — namely, that of the rosids[69]. Tellingly, all three species also show a large peak in the youngest phylostratum, suggesting that there is a high rate of new gene formation in the most recent evolutionary history. This is unlikely to be due to an ongoing radiation but is probably a reflection of a high steady-state rate of new gene emergence through time. Most of these genes would be lost again, but, in times of rapid radiation, larger numbers may become fixed and thus give rise to peaks in the overall pattern. This also explains why the peaks of gene emergence slightly predate the radiations, as the respective genes already existed at these times and then became the raw material for the evolutionary novelties created during the radiation.

Frequency of* de novo *evolution. The peaks of new gene emergence in the youngest phylostrata can be interpreted as being mostly due to the *de novo* formation of genes, simply because they cover time spans that are rather short for a duplication–divergence mechanism. HGT is also unlikely to occur in such large numbers in extant eukaryotic genomes (note that one candidate for an orphan created through HGT, *CG31909*, was detected in *D. melanogaster*[47]). A high rate of *de novo* evolution would appear to be in contrast to several systematic surveys of the emergence of new genes, which have suggested that the *de novo* mechanism only applies to a small percentage of genes (between 2 and 12%[47–49]). However, these surveys also partly included other classes of gene and are therefore difficult to

Selective sweeps
The reduction or elimination of nucleotide variation in the genomic region that surrounds a positively selected new mutation.

Phylostratum
A node in the phylogenetic hierarchy that is represented by one or more fully sequenced genomes and where a set of genes from an organism coalesce to founder genes.



◀ Figure 4 | **Emergence rates of founder genes in three different phyla.** Phylostratigraphy was used to obtain the number of founder genes that have arisen in each phylostratum in *Mus musculus* (**a**), *Drosophila melanogaster* (**b**) and *Arabidopsis thaliana* (**c**). Approximate divergence times were taken from REF. 100, and founder gene emergence was scaled to the time between two phylostrata, as described in REF. 23. Although the largest absolute number of genes in each taxon has arisen within the first two phylostrata (compare with FIG. 1b but note that panel **b** of this figure is from an updated version of the *Drosophila* phylostratigraphy, which includes additional fly genomes), the emergence rate over time is not particularly high for these time periods, as they cover such long time spans. Note that peaks of emergence depend on the taxonomic resolution available, and it is thus possible that earlier peaks are missed in this presentation. The dashed line denotes the average rate of founder gene emergence over the entire evolutionary history. The onset of the major radiations of animals at the beginning of the Palaeozoic era (circa 530 million years ago; Cambrian explosion), the Mesozoic era (circa 250 million years ago; vertebrate radiation) and the Cenozoic era (circa 65 million years ago; mammalian radiation) are marked by arrows. Note that, to allow a better resolution of the more recent patterns, the time axis is not displayed linearly.

evaluate with respect to frequency estimates for orphan gene emergence. The most comprehensive analysis that specifically addresses the classification according to the mechanisms described above was done for primates[48]. This study found a total of 270 primate-specific orphan genes with humans as the focal species. One-quarter (66) of these genes at least partially matched a paralogous gene within humans, and they are thus likely to have evolved according to the duplication–divergence model. Only 6% (15) of these genes could be unequivocally classified as *de novo* evolved genes based on the stringent criterion of the presence of a non-coding syntenic region in outgroup species. The rest of the cases were equivocal, mostly because they resided in rearranged genomic regions, which did not allow clear synteny to be established.

Interestingly, most (142) of these unclassified genes included a fragment of a transposable element in the transcript, suggesting that these elements were involved in the rearrangements of the respective regions. We note that many *de novo* genes described for different *Drosophila* species are also from similarly rearranged genomic regions, as this was, in fact, the search criterion used to identify them[45,46]. Hence, if these primate genes were also to be classified as *de novo*, one would conclude that three-quarters of the detected orphans have evolved by *de novo* processes versus one-quarter having evolved according to the duplication–divergence model.

Another reason for the downward bias against *de novo* genes in current surveys is that these surveys were intentionally conservative with respect to the possible discovery of *de novo* genes to avoid spurious results caused by ambiguous annotations. In fact, there may even be an *a priori* bias against annotating *de novo* genes, as they do not follow an expected annotation model[70]. Furthermore, all studies so far have only focused on protein-coding genes, although *de novo* emergence implies that a non-coding phase could be included in the gene evolution process. Hence, the first step in the *de novo* formation of genes is largely unexplored at present.

Synteny
Conserved genomic arrangements of genes in a linear order.

Emergence of protein structure

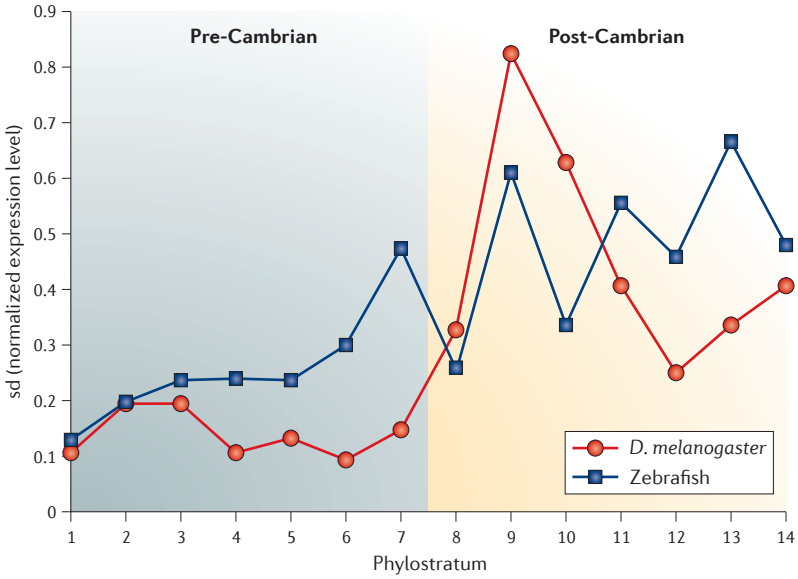
If *de novo* emergence does indeed have a large role in orphan evolution, one has to explain how a new functional protein can emerge out of a previously non-coding sequence. This would seem highly unlikely *a priori*, particularly when one considers our current knowledge of protein evolution. The comparative analysis of protein structures suggests that there may be fewer than 2,000 basic protein folds that make up the majority of characterized proteins[71]. Most of these folds were already present in the ancestral set of genes, and it has been suggested that they are derived from ancestral peptides that were active at the transition from the RNA to the protein world[72–75]. This scenario leaves little room for the *de novo* evolution of new folds and structures, as the conditions for arriving at stable folds may not exist anymore. However, there is still discussion as to whether these folds are homologous (that is, phylogenetically related) or analogous (that is, independently derived). Zhang *et al.*[76] have analysed randomly generated homopolypeptide conformations and have found that these conformations have similar folds to the library of solved structures. They suggest that the simple requirement of having a compact, hydrogen-bonded secondary structure already explains all known folds, which could imply a high likelihood of convergent evolution.

Convergence to a subset of folds would explain why structure-oriented searches suggest that the discovery of new domains is approaching saturation[77,78], whereas sequence-similarity-oriented searches suggest that new families emerge at a rate that is correlated with the addition of new sequences[79,80]. Purely structural search strategies can certainly identify similarities that go beyond sequence-similarity-based phylogenetic relationships[81]. Conversely, these seemingly different results may also be a question of perspective. Comprehensive surveys of sequenced genomes show that most proteins do indeed fall into known structural classes in every genome, but also that each new genome contains singletons that are species-specific or lineage-specific[82]. This pattern can be described with a power law[83], which suggests that the number of non-assigned proteins will keep increasing with new genome sequences, whereas the number of large, recurrent families will eventually reach saturation. In other words, the number of possible orphan proteins by far exceeds the number of known gene families, as is to be expected when completely new genes arise in every evolutionary lineage.

So far, most orphans have no structural fold assigned to them[83], although this does not exclude the fact that structures can be seen in orphan proteins as well[84–86]. However, we are still missing a systematic study that uses PSI-BLAST-based searches[87] to provide a reliable estimate of orphan gene affiliation to the known protein folds. In addition, orphans that unequivocally evolved *de novo* could be an excellent material to test the probability of the convergence of protein folds.

There is also an increasing awareness that proteins and protein regions do not have to be structured to be functional[88–90]. For a long time, studies have concentrated on proteins with defined biochemical functions, which

Box 3 | **Higher tendency of developmental regulation of orphan genes**



One way to show a distinction of old genes versus new genes with respect to stage-specific functions is to look at the variance of expression across stages for the different age classes of genes — that is, whether they are expressed throughout development or only during particular stages. Here we show for *Drosophila melanogaster* and for zebrafish that genes that emerged after the origin of the first bilaterian animal (phylostratum 7, the time of the Cambrian explosion) are much more likely to have a stage-specific expression than older genes.

The plot shows the standard deviation (sd) of gene expression across a series of ontogenetic stages and for genes derived from different phylostrata. The sd serves as a cumulative measure for the degree of ontogenetic regulation across all genes. Phylostrata 1–7 (dark shaded area) represent the time from the origin of the first cell until the origin of the first bilaterians. The plot shows that the later a gene has emerged, the higher its variance in gene expression is on average — that is, younger genes are more likely to be developmentally regulated than older genes are.

Hence, orphan genes contribute most to the ontogenetic differentiation between taxa. It needs to be emphasized that transcription factors are classified as old genes in this analysis because of the early ancestry of DNA-binding domains. Thus, although stage-specifically expressed transcription factors drive ontogenetic differentiation, there are many more orphan genes that contribute to the realization of the respective phenotypes.

could be isolated and crystallized. Intrinsically unstructured proteins, by contrast, have received less attention, although this class of proteins includes important components of transcription, translation and signalling processes[88]. Proteins with intrinsically disordered sequences tend to lack a hydrophobic core to fold into domain structures, but they can associate with other proteins or RNAs and may act as chaperones or be part of network connections[89,91]. It therefore appears that there are many routes by which an initially random protein sequence of a *de novo* evolved gene can indeed become functional.

General functions of orphan genes

If orphan genes mainly function in lineage-specific adaptations, one would expect their functions and expression to be mostly related to stages during development in which organismic lineages diverge from each other (BOX 3). It has long been known that organisms are most similar to each other during the middle stage of development, the so-called phylotypic stage: they diverge before and after this stage, giving rise to the so-called hourglass pattern of development[92]. One can use phylostratigraphy to calculate a transcriptome age index, which is an approximate measure of the cumulative phylogenetic age of the transcriptome at a given stage[25]. Plotting this index across developmental stages in zebrafish, as well as in *D. melanogaster,* indeed shows that the phylotypic stage expresses the oldest transcriptome, whereas younger transcriptomes are found in eggs and mature adults, confirming the hourglass pattern[25]. Importantly, this pattern is mainly driven by the expression of orphan genes.

Chen *et al.*[93] conducted systematic functional tests of genes that have emerged in the *Drosophila* lineage within the past 35 million years. They found that 30% of them had lethal effects when knocked down. However, most of these genes were duplicates that carried known functional domains — that is, they were not orphans. Of the 16 *de novo* evolved genes that were tested, two showed a lethal phenotype (developmental arrest), and one showed a semi-lethal phenotype[93]. Hence, even *de novo* evolved genes can quickly become essential for ontogenetic progression.

Orphan gene functions are also expected to interact specifically with the environment as a consequence of lineage-specific adaptations to this environment. The gut constitutes such an environment for many bacterial species, and Ellrot *et al.*[94] indeed identified protein families in the gut microbiome of humans that have not been found anywhere else. Host–parasite interactions are another example of 'environmental' interactions. In protozoa, surface antigens that are involved in host–parasite interactions are encoded by families of orphan genes[95]. Species-specific traits that are possibly related to feeding are controlled by an orphan gene family occurring in *Hydra* spp.[96]. A function of orphan gene families in responding to environmental stimuli was suggested for the water flea *Daphnia pulex*, a model organism for ecological research[97]. The *D. pulex* genome harbours a large number of orphan genes and orphan gene families, some of which become specifically activated in response to environmental changes[97]. Therefore, the prediction of a general function of orphan genes in lineage-specific adaptations[14,55] is now supported by a fast-growing number of convincing examples.

Conclusions

Although few orphan genes have been studied in much functional detail so far, it is becoming increasingly clear that they have contributed substantially to the evolution of organisms and evolutionary innovations. The phylostratigraphic analysis suggests that orphan genes have emerged throughout evolution and that this is an ongoing process. Many orphan genes may have arisen *de novo* from non-coding regions, a process that may be driven by the broad-scale promiscuous transcription of genomic regions into non-coding RNAs[50], which can assume functional ORFs over time. The fact that this did not result in a continuous growth of gene numbers suggests that genomes are constrained with respect to

maintaining functional regions. There is indeed a limit to what can be maintained by selection that is inversely related to the population size of the species[98]. But gene numbers may also be limited by the cost of translation in the form of selection against the toxicity of misfolded proteins[65]. Either way, gene gains that occur during radiations would be expected to be compensated by gene losses in order to return the system to the balance that is set by these constraints. Lineage-specific loss of genes might therefore be as revealing as lineage-specific gains for understanding the evolutionary history of organisms.

These considerations also shed some new light on the relative importance of regulatory evolution versus gene evolution. The 'predominantly regulatory' model is based on the assumption that the basic building blocks — that is, protein domains — are highly conserved and that it is the regulatory context that determines their function in an organism. There is indeed plenty of evidence that regulatory changes are associated with major shifts in evolution[5], but it is difficult to judge the extent to which these are isolated examples or part of a general trend[99]. The inclusion of orphan genes as a source of novelty can provide a new perspective in this discussion. It refutes the idea that all proteins are mainly composed of conserved modules that only have to be tinkered with and be placed under new regulatory control. However, *de novo* gene emergence requires the concomitant emergence of an associated regulatory element, and this suggests an additional role for regulatory evolution in shaping evolutionary novelties.

1. Ohno, S. *Evolution by Gene Duplication* (Springer, New York, 1970).
2. Jacob, F. Evolution and tinkering. *Science* **196**, 1161–1166 (1977).
3. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
4. Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298 (2003).
5. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
6. Demuth, J. P. & Hahn, M. W. The life and death of gene families. *Bioessays* **31**, 29–39 (2009).
7. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).
This is a comprehensive review of all mechanisms of formation of new genes, in particular duplication and rearrangement processes.
8. Bornberg-Bauer, E., Huylmans, A.-K. & Sikosek, T. How do new proteins arise? *Curr. Opin. Struct. Biol.* **20**, 390–396 (2010).
9. Long, M., Betran, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nature Rev. Genet.* **4**, 865–875 (2003).
10. Zhou, Q. & Wang, W. On the origin and evolution of new genes — a genomic and experimental perspective. *J. Genet. Genomics* **35**, 639–648 (2008).
11. Dujon, B. The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270 (1996).
12. Casari, G., De Daruvar, A., Sander, C. & Schneider, R. Bioinformatics and the discovery of gene function. *Trends Genet.* **12**, 244–245 (1996).
13. Fischer, D. & Eisenberg, D. Finding families for genomic ORFans. *Bioinformatics* **15**, 759–762 (1999).
14. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413 (2009).
15. Yin, Y. & Fischer, D. Identification and investigation of ORFans in the viral world. *BMC Genomics* **9**, 24 (2008).
16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
17. Albà, M. M. & Castresana, J. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).
18. Elhaik, E., Sabath, N. & Graur, D. 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 (2006).
19. Albà, M. M. & Castresana, J. On homology searches by protein BLAST and the characterization of the age of genes. *BMC Evol. Biol.* **7**, 53 (2007).
This is a crucial paper for understanding the power of BLAST for retrieving homologues and the probability of assigning orphan status to genes.
20. Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl Acad. Sci. USA* **106**, 7273–7280 (2009).
This paper shows a universal log-normal distribution of evolutionary rates of proteins and develops a steady-state model of gene gain and gene loss during genome evolution.
21. Cai, J. J. & Petrov, D. A. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol. Evol.* **12**, 393–409 (2010).
This study makes extensive use of comparative genomic data and polymorphism data from human populations to assess selection and adaptation processes in old versus young genes.
22. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
23. Domazet-Loso, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
24. Domazet-Loso, T. & Tautz, D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* **8**, 66 (2010).
25. Domazet-Loso, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).
A systematic study that showed a clear link between phylogenetically young (that is, orphan) genes and global morphological divergence in the developmental context.
26. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Rev. Genet.* **12**, 363–376 (2011).
27. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24**, 363–367 (2000).
28. Kaessmann. H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Rev. Genet.* **10**, 19–31 (2009).
29. Dorit, R. L., Schoenbach, L. & Gilbert, W. How big is the universe of exons? *Science* **250**, 1377–1382 (1990).
30. Patthy, L. Genome evolution and the evolution of exon-shuffling—a review. *Gene* **238**, 103–114 (1999).
31. Kaessmann, H., Zöllner, S., Nekrutenko, A. & Li, W. H. Signatures of domain shuffling in the human genome. *Genome Res.* **12**, 1642–1650 (2002).
32. Buljan, M., Frankish, A. & Bateman, A. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.* **11**, R74 (2010).
33. Cortez, D., Forterre, P. & Gribaldo, S. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. *Genome Biol.* **10**, R65 (2009).
34. Zhaxybayeva, O. & Doolittle, W. F. Lateral gene transfer. *Curr. Biol.* **21**, R242–246 (2011).
35. Keeling, P. J. & Palmer, J. F. Horizontal gene transfer in eukaryotic evolution. *Nature Rev. Genet.* **9**, 605–618 (2008).
36. Lynch. M. & Katju, V. The altered evolutionary trajectories of gene duplicates. *Trends Genet.* **20**, 544–549 (2004).
37. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: how duplicated genes find new functions. *Nature Rev. Genet.* **9**, 938–950 (2008).
38. Nekrutenko, A., Wadhawan, S., Goetting-Minesky, P. & Makova, K. D. Oscillating evolution of a mammalian locus with overlapping reading frames: an XLαs/ALEX relay. *PLoS Genet.* **1**, e18 (2005).
39. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Rev. Genet.* **10**, 155–159 (2009).
40. Cai, J. J., Zhao, R., Jiang, H. & Wang, W. *De novo* origination of a new proteincoding gene in *Saccharomyces cerevisiae*. *Genetics* **179**, 487–496 (2008).
This was the first study that provided direct functional evidence for the evolution of a completely new ORF out of a previously non-coding RNA.
41. Heinen, T. J., Staubach, F., Häming, D. & Tautz, D. Emergence of a new gene from an intergenic region. *Curr. Biol.* **19**, 1527–1531 (2009).
This was the first study that provided direct functional evidence for the *de novo* evolution of a new transcript out of a non-coding genomic region.
42. Knowles, D. G. & McLysaght, A. Recent *de novo* origin of human proteincoding genes. *Genome Res.* **19**, 1752–1759 (2009).
43. Li, D. *et al.* A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* **20**, 408–420 (2010).
44. Li, C. Y. *et al.* A human-specific *de novo* proteincoding gene associated with human brain functions. *PLoS Comput. Biol.* **6**, e1000734 (2010).
45. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from non-coding DNA in *Drosophila melanogaster* are frequently Xlinked and show testis-biased expression. *Proc. Natl Acad. Sci. USA* **103**, 9935–9939 (2006).
46. Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. *Genetics* **176**, 1131–1137 (2007).
47. Zhou, Q. *et al.* On the origin of new genes in *Drosophila*. *Genome Res.* **18**, 1446–1455 (2008).
48. Toll-Riera, M. *et al.* Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**, 603–612 (2009).
This is currently the most comprehensive systematic survey of orphan genes in primates, drawing specific reference to the modes of origin of this gene class.
49. Ekman, D. & Elofsson, A. Identifying and quantifying orphan protein sequences in fungi. *J. Mol. Biol.* **396**, 396–405 (2010).
50. Carninci, P. RNA dust: where are the genes? *DNA Res.* **17**, 51–59 (2010).
51. Sassone-Corsi, P. Unique chromatin remodeling and transcriptional regulation in spermatogenesis. *Science* **296**, 2176–2178 (2002).
52. Kleene, K. C. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. *Dev. Biol.* **277**, 16–26 (2005).
53. Pál, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Genet.* **7**, 337–348 (2006).
54. Cai, J. J., Woo, P. C., Lau, S. K., Smith, D. K. & Yuen, K. Y. Accelerated evolutionary rate may be responsible for the emergence of lineagespecific genes in ascomycota. *J. Mol. Evol.* **63**, 1–11 (2006).
55. Domazet-Loso, T. & Tautz, D. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* **13**, 2213–2219 (2003).

56. Pál, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).
57. Subramanian, S. & Kumar, S. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168**, 373–381 (2004).
58. Lemos, B., Bettencourt, B. R., Meiklejohn, C. D. & Hartl, D. L. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein–protein interactions. *Mol. Biol. Evol.* **22**, 1345–1354 (2005).
59. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337 (2006).
60. Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannenhalli, S. & Plotkin, J. B. Young proteins experience more variable selection pressures than old proteins. *Genome Res.* **20**, 1574–1581 (2010).
61. Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R. & Tatusova, T. A. The relationship of protein conservation and sequence length. *BMC Evol. Biol.* **2**, 20 (2002).
62. Hurst, L. D. & Smith, N. G. C. Do essential genes evolve slowly? *Curr. Biol.* **9**, 747–750 (1999).
63. Hirsh, A. E. & Fraser, H. B. Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).
64. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102**, 5483–5488 (2005).
65. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).
This paper investigates the selective pressures behind protein evolution and suggests that selection against the toxicity of misfolded proteins generated by ribosome errors is a major mechanism that limits the number of genes in a genome.
66. Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* **13**, 2229–2235 (2003).
67. Borenstein, E., Shlomi, T., Ruppin, E. & Sharan, R. Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. *Nucleic Acids Res.* **35**, e7 (2007).
68. Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. A neoproterozoic snowball earth. *Science* **281**, 1342–1346 (1998).
69. Wang, H. *et al.* Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl Acad. Sci. USA* **106**, 3853–3858 (2009).
70. Siepel, A. Darwinian alchemy: human genes from noncoding DNA. *Genome Res.* **19**, 1693–1695 (2009).
71. Orengo, C. A. & Thornton, J. M. Protein families and their evolution—a structural perspective. *Annu. Rev. Biochem.* **74**, 867–900 (2005).
72. Fetrow, J. S. & Godzik, A. Function driven protein evolution. A possible proto-protein for the RNA-binding proteins. *Pac. Symp. Biocomput.* **3**, 485–496 (1998).
73. Lupas, A. N., Ponting, C. P. & Russell, R. B. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**, 191–203 (2001).
74. Söding, J. & Lupas, A. N. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**, 837–846 (2003).
75. Alva, V., Remmert, M., Biegert, A., Lupas, A. N. & Söding, J. A galaxy of folds. *Protein Sci.* **19**, 124–130 (2010).
76. Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E. & Skolnick, J. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl Acad. Sci. USA* **103**, 2605–2610 (2006).
77. Sammut, S. J., Finn, R. D. & Bateman, A. Pfam 10 years on: 10 000 families and still growing. *Brief. Bioinform.* **9**, 210–219 (2008).
78. Levitt, M. Nature of the protein universe. *Proc. Natl Acad. Sci. USA* **106**, 11079–11084 (2009).
79. Kunin, V. *et al.* Myriads of protein families, and still counting. *Genome Biol.* **4**, 401 (2003).
80. Yooseph. D. *et al.* The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5**, e16 (2007).
81. Cheng, H., Kim, B. H. & Grishin, N. V. MALISAM: a database of structurally analogous motifs in proteins. *Nucleic Acids Res.* **36,** D211–D217 (2008).
82. Marsden, R. L., Lee, D., Maibaum, M., Yeats, C. & Orengo, C. A. Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res.* **34**, 1066–1080 (2006).
This study provides an analysis of 203 completed genomes (mostly from bacteria and archaea) and demonstrates that the number of protein families is continually expanding over time and that orphans appear to be an intrinsic part of these genomes.
83. Lee, D., Grant, A., Marsden, R. L. & Orengo, C. Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins* **59**, 603–615 (2005).
84. Siew, N. & Fischer, D. Structural biology sheds light on the puzzle of genomic ORFans. *J. Mol. Biol.* **342**, 369–373 (2004).
85. Narra, H. P., Cordes, M. H. & Ochman, H. Structural features and the persistence of acquired proteins. *Proteomics* **8**, 4772–4781 (2008).
86. Capra, J. A., Pollard, K. S. & Singh, M. Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol.* **11**, R127 (2010).
87. Biegert, A., Mayer, C., Remmert, M., Söding, J. & Lupas, A. The MPI Toolkit for protein sequence analysis. *Nucleic Acids Res.* **34**, W335–W339 (2006).
88. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nature Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
89. Mészáros, B., Tompa, P., Simon, I. & Dosztányi, Z. Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.* **372**, 549–561 (2007).
90. Schlessinger, A. *et al.* Protein disorder—a breakthrough invention of evolution? *Curr. Opin. Struct. Biol.* **21**, 412–418 (2011).
91. Tompa, P. & Kovacs, D. Intrinsically disordered chaperones in plants and animals. *Biochem. Cell Biol.* **88**, 167–174 (2010).
92. Duboule, D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev. Suppl.* **1994**, 135–142 (1994).
93. Chen, S., Zhang, Y. E. & Long, M. New genes in *Drosophila* quickly become essential. *Science* **330**, 1682–1685 (2010).
94. Ellrott, K., Jaroszewski, L., Li, W., Wooley, J. C. & Godzik, A. Expansion of the protein repertoire in newly explored environments: human gut microbiome specific protein families. *PLoS Comput. Biol.* **6**, e1000798 (2010).
95. Kuo, C. H. & Kissinger, J. C. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evol. Biol.* **8**, 108 (2008).
96. Khalturin, K. *et al.* A novel gene family controls species-specific morphological traits in *Hydra*. *PLoS Biol.* **6**, e278 (2008).
97. Colbourne, J. K. *et al.* The ecoresponsive genome of *Daphnia pulex*. *Science* **331**, 555–561 (2011).
98. Tautz, D. A genetic uncertainty problem. *Trends Genet.* **16**, 475–477 (2000).
99. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995–1016 (2007).
100. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).

Acknowledgements
We thank former and current colleagues and laboratory members, as well as three anonymous reviewers who have contributed to the ideas presented here. We thank R. Neme, M. Matejčič and M. S. Šestak for providing phylostratigraphic maps. The work of the authors is supported by institutional funds of the Max-Planck Society, the Ruđer Bošković Institute, the Zoological Institute of the Christian-Albrechts-University Kiel and the Unity Through Knowledge Fund (grant number 49).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Diethard Tautz's homepage:
http://www.evolbio.mpg.de/english/index.html
Tomislav Domazet-Lošo's homepage:
http://www.irb.hr/en/home/tdomazet

ALL LINKS ARE ACTIVE IN THE ONLINE PDF