Evolutionary Origin of Orphan Genes

Diethard Tautz, *Max-Planck Institute for Evolutionary Biology, Plön, Germany*
Rafik Neme, *Max-Planck Institute for Evolutionary Biology, Plön, Germany*
Tomislav Domazet-Lošo, *Ruđer Bošković Institute, Zagreb, Croatia*

Advanced article

Article Contents

Online posting date: 15th May 2013

Orphan genes are genes that occur in specific evolutionary lineages without similarity to genes outside of these lineages and have, therefore, alternatively been named taxonomically restricted genes. They were so far considered to emerge through duplication–divergence processes, but it is now becoming clear that they can also arise *de novo* out of noncoding deoxyribonucleic acid (DNA). This latter process may even occur much more frequently than previously assumed. It appears that genomes harbour many transcripts in a transition stage from nonfunctional to functional genes, also known as protogenes, which are exposed to evolutionary testing and can become fixed when they turn out to be useful. Orphan genes may have played key roles in generating lineage-specific adaptations and could be a continuous source of evolutionary novelties. Their existence suggests that functional ribonucleic acids (RNAs) and proteins can relatively easily arise out of random nucleotide sequences, although these processes still need to be experimentally explored.

Introduction

Emergence of new genes via duplication and divergence of existing genes is a well established concept in evolutionary biology (Ohno, 1970; Zhang, 2003; Demuth and Hahn, 2009; Kaessmann, 2010). However, with the advent of sequencing of full genomes, it became clear that approximately 20–40% of the identified genes could not be associated with a gene family that was known before. Such genes were originally called 'orphan' genes (Dujon, 1996), but later it was suggested to rename them 'taxonomically restricted genes' (Wilson *et al*., 2005). They occur in all domains of life, including bacteria (Wilson *et al*., 2005, 2007; Yin and Fischer, 2006) and viruses (Yin and Fischer, 2008) and methods have been developed to systematically distinguish them from spurious open reading frames (Wilson *et al*., 2007).

For some time it seemed that this class of genes would become smaller once the comparative databases would become larger and search algorithms would become more sensitive (Chothia, 1992; Fischer and Eisenberg, 1999). However, this expectation was not fulfilled. While the number of known gene families that were shared across large phylogenetic distances became saturated, the number of orphan genes kept growing (Wilson *et al*., 2005; Orengo and Thornton, 2005). **See also**: Domain Duplication and Gene Elongation; Gene Duplication: Evolution; Gene Families: Formation and Evolution

eLS subject area: Evolution & Diversity of Life

How to cite:
Tautz, Diethard; Neme, Rafik; and Domazet-Lošo, Tomislav (May 2013) Evolutionary Origin of Orphan Genes. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0024601

A Working Definition of Orphan Genes

While gene families are defined through their presence in diverse evolutionary lineages, orphans are defined through their absence in alternative lineages. This is more difficult to ascertain, because one has to include into the definition the lineages that were considered, as well as the search algorithm applied. However, with the increasing density of genome information available, it is possible to do this systematically. For any given species, one can construct an evolutionary lineage of increasingly distantly related fully sequenced genomes and for any gene of the focal species, one can use a standard search algorithm, such as BLAST (Altschul *et al*., 1997), to identify the phylogenetic level at which the gene has appeared first. This procedure is called 'phylostratigraphy' and it provides a suitable framework to define orphan genes (Domazet-Lošo *et al*., 2007; Tautz and Domazet-Lošo, 2011). Orphan genes in this context are all genes that are restricted to specific lineages in the phylostratigraphy. This implies that all genes that cannot be mapped to the first cellular ancestor of all organisms are orphan genes in some lineages. Hence, a term like 'lineage-restricted genes' would be more appropriate, but the term

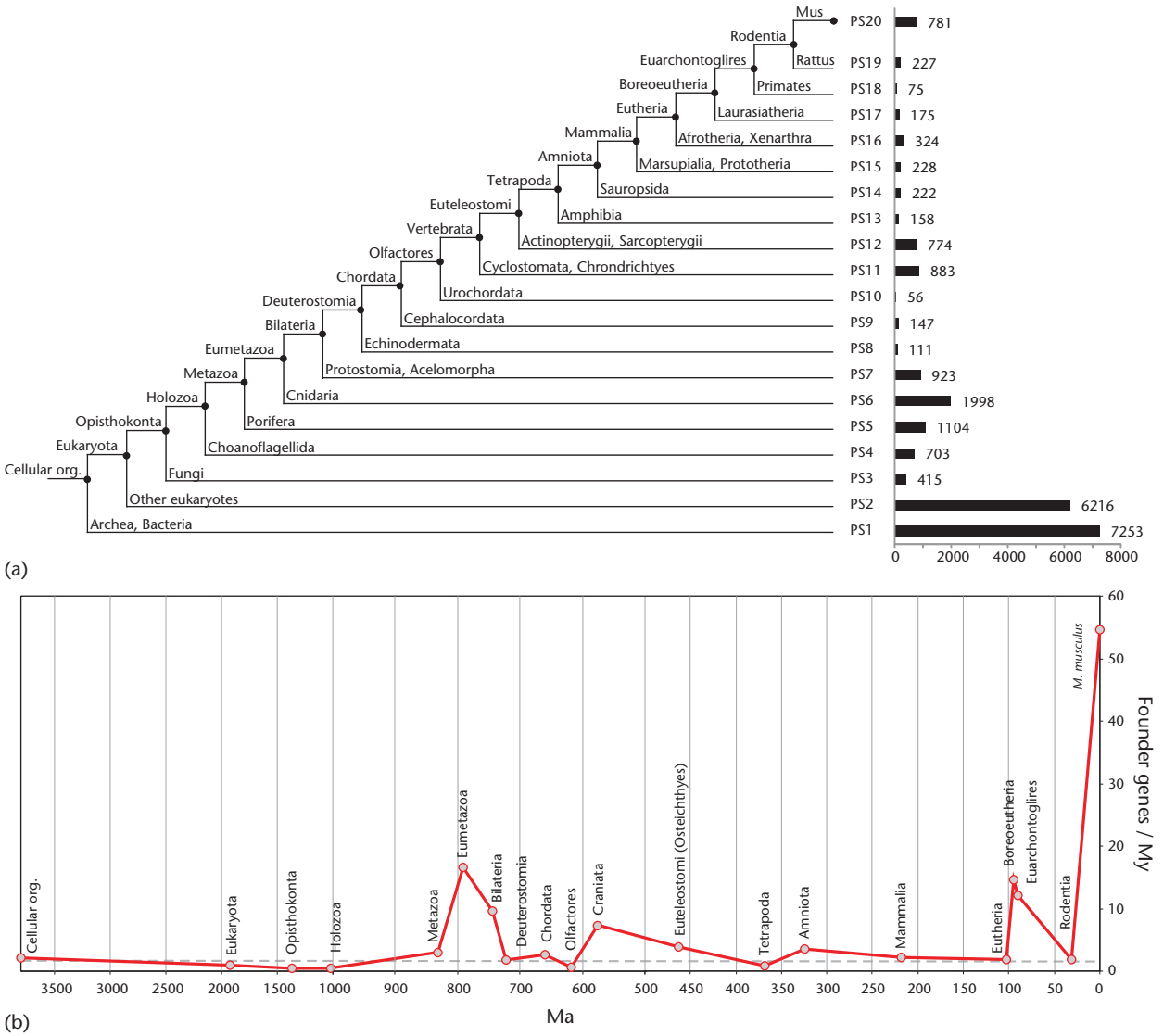

Figure 1 Examples of phylostratigraphic analyses of the mouse genome. (a) Depiction of 20 phylostrata (PS) ranging from the cellular origin to the extant house mouse (*Mus musculus domesticus*) across the whole phylogeny. Each node is represented by several fully sequenced genomes (or at least extensive EST data), representing the respective phylogenetic split. All annotated protein-coding genes of the mouse were subjected to BLAST analysis to find the oldest homologue within this phylogeny. The bar graphs to the right depict the numbers of genes found at the respective levels. The procedure of finding the oldest homologue is necessarily somewhat dependent on the BLAST cutoff chosen (see discussion on this topic in Tautz and Domazet-Lošo, 2011), but the general pattern would not change much at different cutoffs or with different search algorithms. (b) Same analysis as above, but including the time frame for the separation of the nodes and gene numbers scaled to the respective time intervals (note the nonlinear time scale to allow an optimal resolution of the nodes). This depiction allows to infer rates of emergence of genes and shows that the rate is highest in the youngest lineage leading to the extant species.

'orphan genes' was used first in the literature (Dujon, 1996 versus Wilson *et al.*, 2005) and this is why the term is retained here.

Figure 1a shows the phylostratigraphic map of the mouse genome as a typical example. Approximately 32% of the genes can be mapped to the first cellular ancestor (ps1). These include mostly genes relevant for the basal metabolism of cells. A total of 5% map to the origin of animal multicellularity (ps5) and many of them are involved in transcriptional regulation and signalling processes. In the figure we display a total of 20 phylostrata, each representing a particular major evolutionary innovation and each accompanied by a set of genes that appear first at this level of evolution.

Figure 1b shows the same phylostratigraphy, but scaled to time, that is, reflecting emergence rates of genes. This representation reveals peaks that suggest uneven rates of orphan gene emergence across time. Interestingly, these appear in conjunction with major morphological radiations and the largest peak is seen for the youngest lineage leading towards the mouse. Hence, this peak represents coding genes that are not even present in the rat, suggesting

a particularly high activity of recruitment of new genes in the youngest evolutionary lineages. This pattern is consistent across all species analysed so far and thus constitutes a framework to be considered while evaluating models of orphan gene emergence.

Origin of Orphan Genes

Apart from duplications of complete genes which can result in the formation of gene families, it is well known that fragments of existing genes can also be recombined to form new genes with new functions (Long *et al*., 2003). These could also be classified as orphan genes, but if one uses a similarity search algorithm to identify the origin of such genes, one would map their appearance to the phylostratigraphic level of the fragments or domains that have formed the gene. On the other hand, newly duplicated genes could assume new functions and go through a phase of fast sequence evolution during which the similarity to founder gene(s) is lost (Domazet-Lošo and Tautz, 2003). This is the duplication–divergence model for the origin of orphan genes, which was originally proposed because it seemed very unlikely that functional genes could emerge out of random noncoding sequences. However, as further discussed below, this picture has now changed. The *de novo* evolution of genes out of random sequences has become a realistic possibility and this develops into the major model for orphan gene evolution (Siepel, 2009; Bornberg-Bauer *et al*., 2010; Tautz and Domazet-Lošo, 2011).

Another component in the evolution of new genes can be transposons. Newly evolved genes and orphan genes frequently contain fragments of transposable elements and in some cases these were shown to have directly contributed to the formation of the respective genes (Volff, 2006; Zhou *et al*., 2008; Toll-Riera *et al*., 2009). However, it seems also possible that their presence in newly evolved genes could simply be a reflection of their ubiquitous presence in noncoding parts of the genome, that is, they might have been passively recruited to become part of the new genes.

Horizontal gene transfer could also play a role in the appearance of new genes in some lineages. There are now many well documented cases where gene functions were transferred across large phylogenetic distances (Keeling and Palmer, 2008) and such genes would appear as orphan genes in the new lineages, as long as one has no reference of the donor lineage. However, given the increasing density of fully sequenced genomes across all domains of life, possible donor lineages for horizontal gene transfers tend to become readily identified.

From Protogenes to Genes

The *de novo* evolution model constitutes the largest conceptual step forward in our understanding of gene evolution. On the basis of simple combinatorial considerations, it seemed initially very unlikely that functional proteins could arise out of essentially random sequences (Jacob, 1977). However, direct and indirect evidence for *de novo* evolution has quickly accumulated in the past years, due to the breadth of genomic comparisons that are now available; based on this, one can formulate a general model of the dynamics of gene emergence and gene turnover (**Figure 2**). This model posits that a new gene starts from an initially spuriously transcribed genome region that acquires some additional mutations to convert the resulting ribonucleic acid (RNA) into a stable heritable transcript, for example, through additional promotor mutations, or by creation of a functional polyadenylation signal. This new transcript could initially be nonfunctional, subject to neutral evolutionary processes, which could lead to random loss or fixation in given populations. However, such transcripts can also be considered to be protogenes

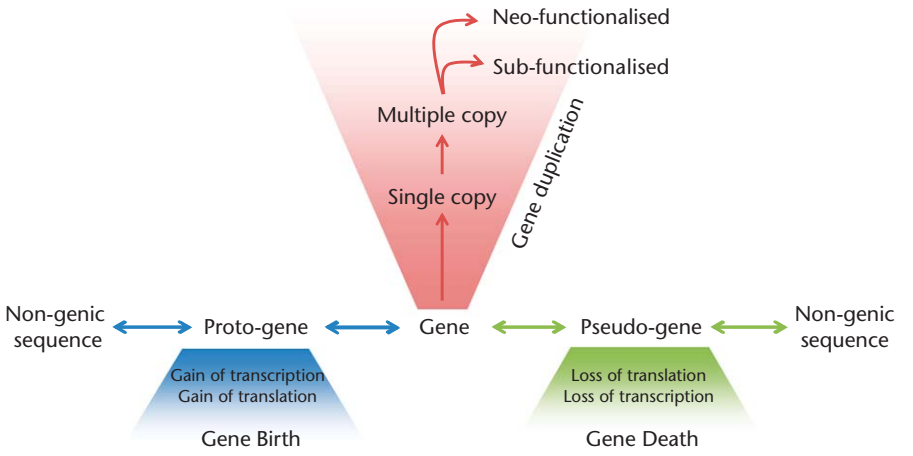

Figure 2 A general depiction of the life cycle of genes (after Carvunis *et al*., 2012). This representation assumes that genes emerge regularly out of nongenic sequences via a protogene phase. Once established as functional genes, they can expand into gene families. Alternatively, gene copies can also be lost again and become nongenic sequences.

(Siepel, 2009) that have the chance to become functional in some way, because they are exposed to evolutionary testing. They might initially act as noncoding RNA genes associated to other molecular complexes. In a second step, they could become translated and this would expose the resulting peptide to evolutionary testing. Once the peptide becomes useful for a biological process in the respective species, it would become further optimised and eventually become part of the gene repertoire of the species. Such new genes might then continue to form gene families through duplication processes and mature their functions, or could alternatively be lost again (**Figure 2**).

All these steps from a spurious transcript to a stable new gene are now well documented by concrete examples in various species. The first reports on possible cases of *de novo* evolution came from *Drosophila* (Levine *et al*., 2006; Begun *et al*., 2007), but these had still some limitations with respect to the number of closely related species that were available for comparisons. To make a convincing case for *de novo* evolution rather than lineage-specific losses, one has to show that a given genomic region is present in multiple closely related species, whereby only the focal species has developed a functional transcript in the respective region (**Figure 3**).

The emergence of a new functional RNA gene based on these standards was first documented in the mouse (Heinen *et al*., 2009). The *Pldi* gene has arisen within the house mouse species complex and the history of relevant mutations leading towards its functionality could be traced through comparisons between multiple closely related species. It is specifically expressed in postmeiotic cells of the testis and its knockout leads to a sperm motility phenotype. Microarray studies in the knockout animals suggest that it is involved in chromatin organisation (Heinen *et al*., 2009).

The first documented case of a transition from an RNA gene to a coding gene comes from yeast (Cai *et al*., 2008). The BSC4 gene in *Saccharomyces cerevisiae* codes for a 132-amino acid protein involved in the deoxyribonucleic acid (DNA) repair pathway and appears to contribute to the robustness of the cells when shifted to a nutrient-poor environment. The gene appears to be expressed as noncoding RNA in closely related species, including the sibling species *Saccharomyces paradoxus*, but the open reading frame is only found in *S. cerevisiae* (Cai *et al*., 2008).

There may be many more such cases in yeast. Analysis of data from ribosome profiling studies has shown that the cells harbour an unexpectedly large number of transcripts being apparently in the transition from an RNA gene to a protein-coding gene (Wilson and Masel, 2011; Carvunis *et al*., 2012). Although similar studies will still have to be done in other species as well, the yeast results make it likely that thousands of protogenes may be present in any given genome, ready to be tested for function and possible stable integration into the gene set.

Comparative studies in primates have also revealed many candidates for *de novo* evolved genes, including genes in the lineage towards humans. Knowles and McLysaght (2009) were the first to report three such genes, but many additional ones have been found by now (Li *et al*., 2010; Wu *et al*., 2011; Xie *et al*., 2012). For several of these genes, it was possible to show that they have indeed started out with a noncoding RNA before they acquired an apparently functional reading frame (Xie *et al*., 2012).

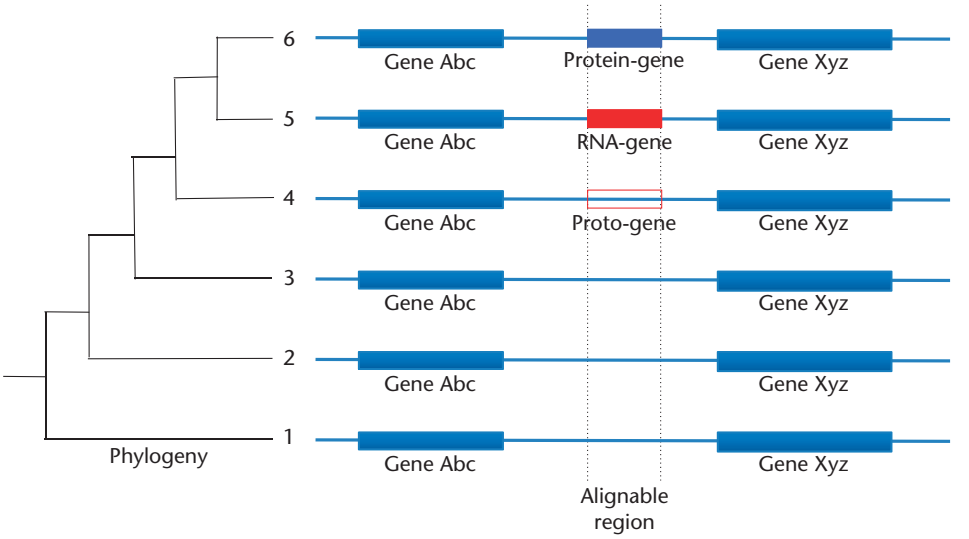

Figure 3 Depiction of the inference scheme for *de novo* evolution out of nongenic DNA. This scheme depicts a phylogeny of six related species, of which only species 6 is the focal species, where the hypothesis of a *de novo* gene evolution is tested. To make a solid case, one should show that the corresponding DNA region is present in the related species and synthenic in these species (indicated here by the depiction of the flanking genes Abc and Xyz). The region should also still be alignable, that is, the species that are compared should be sufficiently close to each other to ensure that even neutrally diverging sequences have not yet acquired too many mutations. Finally, all outgroups should not have a sign of a gene in the respective position, possibly apart of the most closely related ones, which could have a protogene or an RNA gene in the position.

Hence, contrary to the assumption that *de novo* gene evolution should be rare, it appears that the transition from protogenes to genes may be a continuously acting process, in which thousands of transcripts are involved. This also provides the explanation as to why we see in the phylostratigraphic patterns always peaks within the youngest lineages, because these represent the genes that are in the process of being evolutionary tested (Tautz and Domazet-Lošo, 2011).

Overprinting of Reading Frames

The *de novo* evolution of new proteins can potentially occur also within the existing genes through the use of an alternative reading frame. The first report of *de novo* evolution of a new gene active in the degradation of nylon oligomers in *Flavobacterium* through the use of an alternative reading frame of an existing gene came from Ohno (1984). The term 'overprinting' was later proposed by Keese and Gibbs (1992), who described cases of *de novo* evolution of genes in various viral proteins, but suggested it also as a general mechanism for *de novo* evolution. There were then indeed several reports on specific eukaryotic genes with overlapping reading frames coding for two proteins (Klemke *et al*., 2001; Nekrutenko *et al*., 2005). A particularly well studied case are the $p16^{INK4a}$ and $p19^{ARF}$ genes, two tumour suppressor genes that are initiated from different promotors, but are translated from overlapping reading frames (Quelle *et al*., 1995; Sherr, 2006; **Figure 4**). Chung *et al*. (2007) published a first attempt on systematically identifying cases of overprinting in eukaryotes and there is now abundant evidence for overprinting occurring regularly in viruses (Sabath *et al*., 2012).

Orphan Gene Functions

It has long been assumed that orphan genes have a specific role in lineage-specific adaptations (Domazet-Lošo and Tautz, 2003; Wilson *et al*., 2005; Khalturin *et al*., 2009), but there are still only few well studied cases of individual genes that support this notion. One of the most extensive studies of this question was done in hydra, where it was shown that a gene family specific to the hydra lineage is functionally involved in species-specific differences in tentacle formation (Khalturin *et al*., 2008). A systematic functional study of genes that had evolved within the *Drosophila melanogaster* lineage in the past 35 My showed that approximately one-third of them had become essential for the species, that is, knockouts had lethal effects (Chen *et al*., 2010). However, most of these were rearranged genes, that is, they would be classified under the duplication–divergence model. Among the 16 true *de novo* evolved genes in this study, two showed a lethal and one a semilethal phenotype, but none of the phenotypes was sufficiently distinctive to allow a conclusion about lineage-specific adaptations.

Given that new gene functions have appeared throughout the phylogeny, it is also of interest to assess their association with older evolutionary innovations. This can be done via the phylostratigraphic approach, where gene lists can be associated with major evolutionary steps, such as the origin of germ layers (Domazet-Lošo *et al*., 2007), or the origin of multicellularity (Domazet-Lošo and Tautz, 2010a). Interestingly, this approach showed also that younger genes tend to be increasingly more developmentally regulated compared with evolutionary older genes (Tautz and Domazet-Lošo, 2011).

A particularly interesting pattern emerges when one combines phylogenetic age of genes with ontogenetic expression. It has long been known that embryos go through a phase of development where they are more similar between diverse species than before or after this stage. This pattern has become known as the developmental hour-glass and there has been much speculation about how it could be explained. Intriguingly, when combining expression information and age of genes in a transcriptome age index, one can also observe an hour-glass pattern, with the relatively oldest transcriptome being expressed at the stage where the organisms belonging to a phylum differ least from each other (the so-called phylotypic stage) (Domazet-Lošo and Tautz, 2010b; Quint *et al*., 2012). On the other hand, the divergence is highest among adults (Domazet-Lošo and Tautz, 2010b), that is, at the stage where all the genes required for the species-specific adaptations are expressed. Hence, this overall pattern gives also good credence to the idea that younger genes are more likely to be crucial for lineage-specific adaptations.

Open Questions

The orphan status of a given gene is usually defined via BLAST searches, but this may not always result in the recovery of highly diverged homologues. There have been specific studies to assess the appropriateness of BLAST for

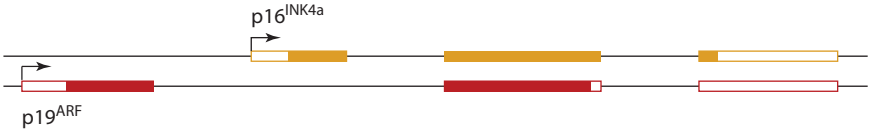

Figure 4 Example for a well studied overprinted locus. Both genes are tumour suppressor genes, but p16INK4a is the older one. p19ARF (ARF, alternative reading frame) originated through a new exon that splices to the central exon of p16INK4a but is translated from a different frame. Both proteins were shown to be functional (Quelle *et al*., 1995). Boxes indicate exons, filled boxes indicate protein-coding regions.

such analysis (Alba and Castresana, 2007; see also discussion in Tautz and Domazet-Lošo, 2011), but it is nonetheless clear that more refined search algorithms can place the origin of genes at deeper levels in some cases. These search algorithms are based on profile searches (Altschul *et al*., 1997; Remmert *et al*., 2012), which benefit from a growing database, although an increase in sensitivity can also increase the false positive rate. Hence, more studies on the true origin of gene sets will have to be conducted in the future.

On the other hand, there is now little doubt that new genes have arisen throughout the phylogenetic history and the general model of *de novo* evolution of genes appears to be well supported by now. However, this also raises several new questions. The foremost one is the question of how new promotors with a defined regulation can arise. Although it is well known that most of the genome is transcribed at some point (Berretta and Morillon, 2009; Carninci, 2010; Clark *et al*., 2011), much of this may be spurious transcription. A good protogene transcript would require to be heritable and available for evolutionary testing for many generations. Furthermore, nonspecific transcriptional low level noise, as a source to spurious transcripts, might require the recruitment and coevolution of repressors and activators upon successful detection of functionally relevant transcripts (Polev, 2012).

It has been proposed that many new genes might first be expressed in the testis (Levine *et al*., 2006; Begun *et al*., 2007; Kaessmann *et al*., 2009), because the promotor requirements for postmeiotic expression appear to be particularly relaxed (Kleene, 2005). This would make it easy to generate a stable heritable transcript by single or few mutations, as it appears to have happened for the *Pldi* gene (Heinen *et al*., 2009). Although this out-of-testis hypothesis is compelling and is supported by some evidence, a detailed analysis of the history of the emergence of new genes does not always support it (Xie *et al*., 2012). Moreover, given the universal appearance of *de novo* genes even in single-celled organisms and prokaryotes, a broader understanding of how stable transcription can emerge through simple mutations is required. Hence, the observation that yeast harbours thousands of protogenes (Carvunis *et al*., 2012) is of specific significance and it will become of special interest to trace how such transcripts have acquired active promotors.

Overprinting of existing genes would be an alternative for new proteins to emerge, without the need to have a new transcript to emerge first, because this is already available. Hence, this mechanism should be explored further, in particular in view of the fact that multiple translational start sites in eukaryotic genes are not rare (Michel *et al*., 2012) and that even short peptides can be functional (Tautz, 2009).

Another major question is how a random peptide can become functional. Almost all well defined peptide folds in functional proteins appear to have arisen already at the time of origin of life and there is an ongoing debate as to whether new folds could arise at all (Soding and Lupas, 2003; Zhang *et al*., 2006). On the other hand it is well known that many proteins do not appear to require a specific fold to be functional. It has been suggested that a significant fraction of proteins are intrinsically disordered, including important genes involved in transcription, translation and signalling processes (Dyson and Wright, 2005; Schlessinger *et al*., 2011). Nevertheless, a systematic analysis of the fraction of randomly generated peptides that could assume a specific function is still missing.

Finally, given the unexpectedly high number of protogenes in presumably any genome, as well as the observation that the youngest lineages in the phylostratigraphy show the highest rates of new gene emergence, one could wonder why the average gene content of organisms does not keep rising. In fact, there are good reasons to assume that there is a threshold for a maximum number of genes that could be evolutionarily stable, mostly because of biochemical limits of error control that can be attained (Drummond and Wilke, 2008). An obvious consequence is that there is not only a high rate of gene birth in the youngest lineages, but most likely an equally high rate of gene loss, because most species will already be at the limits of a long-term stable gene content. Lineage-specific losses of genes are already well documented (Krylov *et al*., 2003) and a recent report shows that this turnover of gain and loss occurs already at the level of long noncoding RNAs (Kutter *et al*., 2012). Hence, the question of the evolutionary dynamics of gene gain and loss also deserves more attention in the future.

References

Alba MM and Castresana J (2007) On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology* **7**: 53.

Altschul SF, Madden TL, Schaffer AA *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Begun DJ, Lindfors HA, Kern AD and Jones CD (2007) Evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics* **176**: 1131–1137.

Berretta J and Morillon A (2009) Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Reports* **10**: 973–982.

Bornberg-Bauer E, Huylmans AK and Sikosek T (2010) How do new proteins arise? *Current Opinion in Structural Biology* **20**: 390–396.

Cai J, Zhao RP, Jiang HF and Wang W (2008) *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179**: 487–496.

Carninci P (2010) RNA dust: where are the genes? *DNA Research* **17**: 51–59.

Carvunis AR, Rolland T, Wapinski I *et al*. (2012) Proto-genes and *de novo* gene birth. *Nature* **487**: 370–374.

Chen SD, Zhang YE and Long MY (2010) New genes in *Drosophila* quickly become essential. *Science* **330**: 1682–1685.

Chothia C (1992) Proteins – 1000 families for the molecular biologist. *Nature* **357**: 543–544.

Chung WY, Wadhawan S, Szklarczyk R, Pond SK and Nekrutenko A (2007) A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Computational Biology* **3**: 855–861.

Clark MB, Amaral PP, Schlesinger FJ *et al*. (2011) The reality of pervasive transcription. *PloS Biology* **9**.

Demuth JP and Hahn MW (2009) The life and death of gene families. *Bioessays* **31**: 29–39.

Domazet-Lošo T, Brajkovic J and Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics* **23**: 533–539.

Domazet-Lošo T and Tautz D (2003) An evolutionary analysis of orphan genes in *Drosophila*. *Genome Research* **13**: 2213–2219.

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

Domazet-Lošo T and Tautz D (2010b) Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biology* **8**: 66.

Drummond DA and Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**: 341–352.

Dujon B (1996) The yeast genome project: what did we learn? *Trends in Genetics* **12**: 263–270.

Dyson HJ and Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology* **6**: 197–208.

Fischer D and Eisenberg D (1999) Finding families for genomic ORFans. *Bioinformatics* **15**: 759–762.

Heinen T, Staubach F, Haming D and Tautz D (2009) Emergence of a new gene from an intergenic region. *Current Biology* **19**: 1527–1531.

Jacob F (1977) Evolution and tinkering. *Science* **196**: 1161–1166.

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

Kaessmann H, Vinckenbosch N and Long MY (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics* **10**: 19–31.

Keeling PJ and Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* **9**: 605–618.

Keese PK and Gibbs A (1992) Origins of genes: big bang or continuous creation? *Proceedings of the National Academy of Sciences of the United States of America* **89**: 9489–9493.

Khalturin K, Anton-Erxleben F, Sassmann S *et al*. (2008) A novel gene family controls species-specific morphological traits in hydra. *PloS Biology* **6**: 2436–2449.

Khalturin K, Hemmrich G, Fraune S, Augustin R and Bosch TCG (2009) More than just orphans: are taxonomically restricted genes important in evolution? *Trends in Genetics* **25**: 404–413.

Kleene KC (2005) Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. *Developmental Biology* **277**: 16–26.

Klemke M, Kehlenbach RH and Huttner WB (2001) Two overlapping reading frames in a single exon encode interacting proteins – a novel way of gene usage. *EMBO Journal* **20**: 3849–3860.

Knowles DG and McLysaght A (2009) Recent *de novo* origin of human protein-coding genes. *Genome Research* **19**: 1752–1759.

Krylov DM, Wolf YI, Rogozin IB and Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research* **13**: 2229–2235.

Kutter C, Watt S, Stefflova K *et al*. (2012) Rapid turnover of long noncoding rnas and the evolution of gene expression. *PloS Genetics* **8**.

Levine MT, Jones CD, Kern AD, Lindfors HA and Begun DJ (2006) Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the USA* **103**: 9935–9939.

Li CY, Zhang Y, Wang ZB *et al*. (2010) A human-specific *De novo* protein-coding gene associated with human brain functions. *PLoS Computational Biology* **6**: e1000734.

Long M, Betran E, Thornton K and Wang W (2003) The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics* **4**: 865–875.

Michel AM, Choudhury KR, Firth AE *et al*. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Research* **22**: 2219–2229.

Nekrutenko A, Wadhawan S, Goetting-Minesky P and Makova KD (2005) Oscillating evolution of a mammalian locus with overlapping reading frames: An XL alpha s/ALEX relay. *PloS Genetics* **1**: 197–204.

Ohno S (1970) *Evolution by Gene Duplication*. New York: Springer.

Ohno S (1984) Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proceedings of the National Academy of Sciences of the USA* **81**: 2421–2425.

Orengo CA and Thornton JM (2005) Protein families and their evolution – a structural perspective. *Annual Review of Biochemistry* **74**: 867–900.

Polev D (2012) Transcriptional noise as a driver of gene evolution. *Journal of Theoretical Biology* **293**: 27–33.

Quelle DE, Zindy F, Ashmun RA and Sherr CJ (1995) Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell* **83**: 993–1000.

Quint M, Drost HG, Gabel A *et al*. (2012) A transcriptomic hourglass in plant embryogenesis. *Nature* **490**: 98–101.

Remmert M, Biegert A, Hauser A and Soding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**: 173–175.

Sabath N, Wagner A and Karlin D (2012) Evolution of viral proteins originated *de novo* by overprinting. *Molecular Biology and Evolution* **29**: 3767–3780.

Schlessinger A, Schaefer C, Vicedo E *et al*. (2011) Protein disorder – a breakthrough invention of evolution? *Current Opinion in Structural Biology* **21**: 412–418.

Sherr CJ (2006) Divorcing ARF and p53: an unsettled case. *Nature Reviews Cancer* **6**: 663–673.

Siepel A (2009) Darwinian alchemy: human genes from noncoding DNA. *Genome Research* **19**: 1693–1695.

Soding J and Lupas AN (2003) More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**: 837–846.

Tautz D (2009) Polycistronic peptide coding genes in eukaryotes – how widespread are they? *Briefings in Functional Genomics & Proteomics* **8**: 68–74.

Tautz D and Domazet-Lošo T (2011) The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**: 692–702.

Toll-Riera M, Bosch N, Bellora N *et al*. (2009) Origin of primate orphan genes: a comparative genomics approach. *Molecular Biology and Evolution* **26**: 603–612.

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**: 913–922.

Wilson BA and Masel J (2011) Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biology and Evolution* **3**: 1245–1252.

Wilson GA, Bertrand N, Patel Y *et al*. (2005) Orphans as taxonomically restricted and ecologically important genes. *Microbiology* **151**: 2499–2501.

Wilson GA, Feil EJ, Lilley AK and Field D (2007) Large-scale comparative genomic ranking of taxonomically restricted genes (TRGS) in bacterial and archaeal genomes. *PloS One* **2**: e324.

Wu DD, Irwin DM and Zhang YP (2011) *De novo* origin of human protein-coding genes. *PloS Genetics* **7**: e1002379.

Xie C, Zhang YE, Chen JY *et al*. (2012) Hominoid-specific *de novo* protein-coding genes originating from long non-coding RNAs. *PloS Genetics* **8**: e1002942.

Yin YB and Fischer D (2006) On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. *BMC Evolutionary Biology* **6**: 63.

Yin YB and Fischer D (2008) Identification and investigation of ORFans in the viral world. *BMC Genomics* **9**: 24.

Zhang JZ (2003) Evolution by gene duplication: an update. *Trends in Ecology and Evolution* **18**: 292–298.

Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E and Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences of the USA* **103**: 2605–2610.

Zhou Q, Zhang GJ, Zhang Y *et al*. (2008) On the origin of new genes in *Drosophila*. *Genome Research* **18**: 1446–1455.