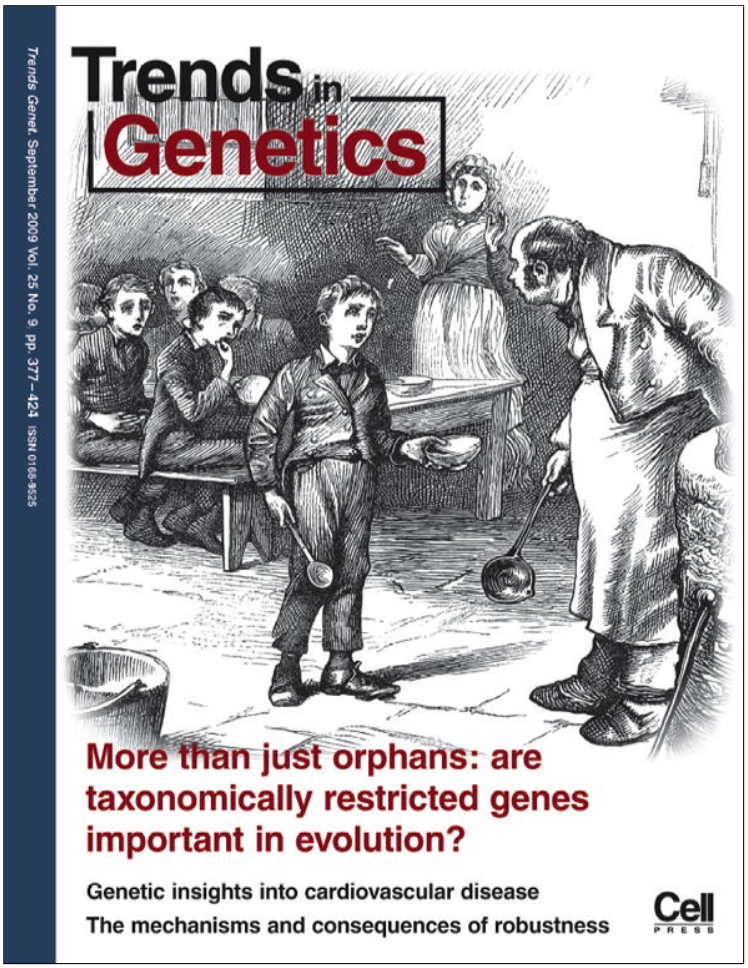

Review

Cell PRESS

More than just orphans: are taxonomically-restricted genes important in evolution?

Konstantin Khalturin, Georg Hemmrich, Sebastian Fraune, René Augustin and Thomas C.G. Bosch

Zoological Institute, Christian-Albrechts-University Kiel, Olshausenstrasse 40, 24098 Kiel, Germany

Comparative genome analyses indicate that every taxonomic group so far studied contains 10–20% of genes that lack recognizable homologs in other species. Do such 'orphan' or 'taxonomically-restricted' genes comprise spurious, non-functional ORFs, or does their presence reflect important evolutionary processes? Recent studies in basal metazoans such as *Nematostella*, *Acropora* and *Hydra* have shed light on the function of these genes, and now indicate that they are involved in important species-specific adaptive processes. Here we focus on evidence from *Hydra* suggesting that taxonomically-restricted genes play a role in the creation of phylum-specific novelties such as cnidocytes, in the generation of morphological diversity, and in the innate defence system. We propose that taxon-specific genes drive morphological specification, enabling organisms to adapt to changing conditions.

Resolving an evolutionary paradox

Advances in genome and transcriptome sequencing have demonstrated that the molecular basis of development has deep evolutionary roots [1,2]. All eumetazoan animals, from cnidarians to humans, share highly conserved signal transduction pathways, which, together with hundreds of conserved transcription factors, comprise a molecular toolkit common to all living beings [3–6]. The high degree of conservation, however, causes an evolutionary paradox [7]: if the key developmental control genes are the same, and if conserved gene families serve similar functions in different organisms, how is the enormous morphological and physiological diversity within the animal kingdom generated? What are the genetic changes that underlie the origins of new species and their adaptation to perpetually varying environments? An elegant and widely accepted solution to this evolutionary puzzle is the concept of regulatory evolution [8]. Regulatory genes are shared and highly conserved throughout the animal kingdom; therefore the differential use of similar components might underlie morphological differences among species. Such 'rewiring' of gene regulatory networks enables changes at any scale – from subtle intra-specific morphological variations to the creation of novel phylum-specific features and structures [2,8].

Do changes in *cis*-regulatory elements [8] and gene duplications [9] represent the only driving forces of morphological evolution? There is at least one additional source of diversity that is often overlooked. Every eukaryotic genome contains 10–20% of genes without any significant sequence similarity to genes of other species; these are classified as 'orphans' or 'taxonomically-restricted genes' (TRGs) [10,11]. Although such genes have arisen in the genomes of every group of organisms studied so far, [12–17], they have received comparatively little attention and their functions remain largely unknown [18,19]. It has even been argued that such 'orphans' comprise spurious, non-functional open reading frames (ORFs) [20].

In this review we discuss how the presence of TRGs within the genomes of diverse species might contribute to evolutionary processes. We consider the implications of the exclusive presence of some genes in one species or animal group. Because these are still early days for functional studies on TRGs, we focus on examples in basal metazoans (Figure 1) where at least some functional data are already available. Specifically, we discuss recent evidence from *Hydra* which points to roles for taxonomically-restricted genes in the creation of phylum-specific novelties such as cnidocytes, the generation of morphological diversity, and in the innate defence system. We propose that taxon-specific genes, in combination with rewiring of the genetic networks of conserved regulatory genes, drive morphological specification and allow organisms to adapt to constantly changing ecological conditions. We first turn our attention to early work performed in yeast and bacteria.

Glossary

Orphan genes:: genes without detectable sequence similarity in the genomes of other organisms. According to the strictest definition these are genes which do not encode any previously-identified protein domains. There is no general agreement or rule, but usually proteins which do not show any sequence similarity in BLASTP searches with cut-off values $E < 10^{-5}$ or $E < 10^{-10}$ are referred to as 'orphan' or 'novel'. There is increasing evidence from genomic and transcriptomic sequencing that 'orphans' represent genes which have a narrow phylogenetic distribution (homologous 'orphans' can be present in closely related species but are not present in more distantly related species or in other genera).

Taxonomically-restricted genes (TRGs):: a more careful definition of 'orphan' or 'novel' genes that takes into consideration not only the absence of sequence similarity to genes or proteins in other organisms, but also the narrow phylogenetic distribution of these genes. This definition was introduced by Wilson [10] and is synonymous to 'lineage-specific genes'.

***Hydra vulgaris* AEP**: a laboratory strain of hydra that constantly undergoes sexual proliferation. In addition to asexual propagation by budding these animals produce large amounts of eggs and sperm all the year round. The AEP strain is used to generate transgenic polyps by embryonic microinjection [55].

Corresponding author: Bosch, T.C.G. (tbosch@zoologie.uni-kiel.de)

0168-9525/$ – see front matter doi:10.1016/j.tig.2009.07.006

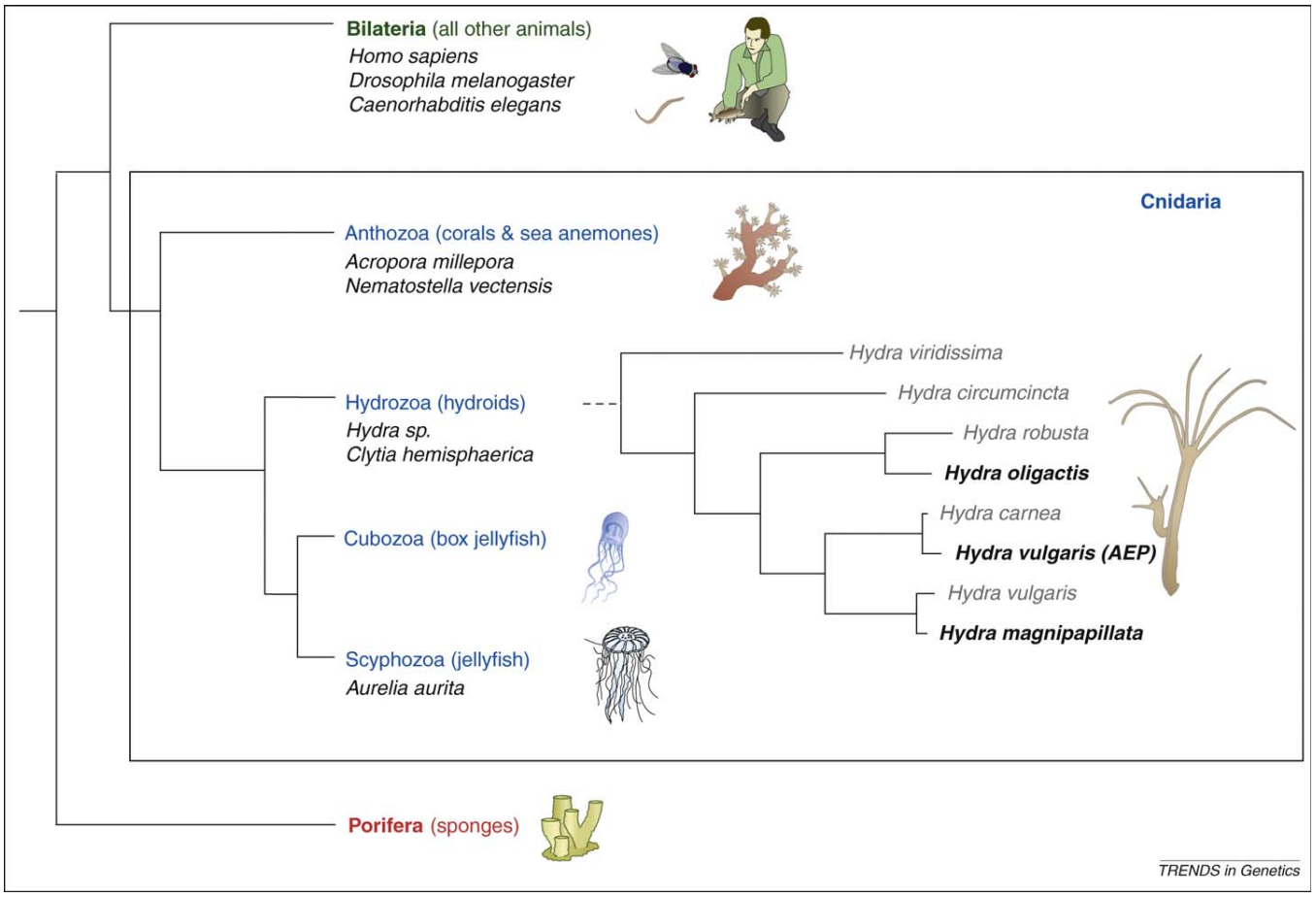

Figure 1. Phylogeny of the 'lower' Metazoa. Class Anthozoa, to which the coral *Acropora* and the sea anemone *Nematostella* belong, is basal within the phylum Cnidaria. *Hydra* is the textbook representative of the phylum and remains the best-characterized cnidarian at the cellular level, but is a derivative member of the Hydrozoa. Several hydra species serve as models for studying pattern formation, the evolution of innate immunity, and the origin of stem cell systems. Genomes of *Nematostella vectensis* [4] and *Hydra magnipapillata* [http://hydrazome.metazome.net/cgi-bin/gbrowse/hydra/] have been sequenced.

'Orphans', 'novel' or 'taxonomically-restricted' genes? – reflections on yeast and bacteria

Early discussions about the existence and significance of 'novel' or 'orphan' genes date back to the sequencing of the yeast chromosome III in the early 1990s, when 182 genes were identified within 315 Kb [21], and only 37 of these were known to scientists at that time [21]. This surprising result clearly demonstrated that, even in a genome as extensively studied as *Saccharomyces cerevisiae*, only a small fraction of protein coding genes had been identified by mutagenesis screening. A large proportion of genes had somehow escaped classical genetic approaches because disruption of their function had not shown any obvious phenotype. These genes were referred to as 'orphans' and a debate about their function was initiated [21].

On completion of the yeast genome in 1996 the issue of 'orphans' was systematically addressed by Dujon [22]. A major part of the yeast genome was found to comprise a large set of previously undiscovered genes (on average 30–35% of the genome). Originally the term 'orphan' had a double meaning. In the initial analysis of the yeast genome it meant both 'coding region without known function' and 'coding region without matches to other genes in the databases'. Nowadays the latter meaning is mostly used (see Glossary) and, according to this 'sequence-oriented' definition, the set of orphans in the yeast genome at that time was ~26% [23]. Comparative genomics was in an infant state in 1996; such a large proportion of 'orphan' genes in the yeast genome clearly reflected sequence undersampling in available databases. It was thus thought that the number of 'orphans' would reduce quickly (eventually, down to zero) as soon as the molecular databases incorporated more sequences from different organisms [23]. However, the early view that 'orphans' would vanish has not come true.

How did the cumulative number of orphan genes change with the sequencing of new genomes? Starting with the sequencing of the yeast genome, the presence of genes without any significant similarity in databases was reported in every genome project that followed (for more information see Box 1 and Table S1), and the number is still growing. Although the existence of 'orphan' genes is not a trivial issue, given their ubiquitous presence, there have been few systematic attempts to identify their evolutionary origin and significance [18,24,25].

That 10–20% of every genome is composed of orphan genes only becomes obvious in the context of comparative genomics which, by definition, is strongly dependent on the availability of reliable datasets. The ideal dataset should

Box 1. Orphan genes constitute a large portion of every sequenced genome

One major step in every genome sequencing project is to compile a reference set of genes for a species. This procedure integrates information on previously analyzed and published genes, output from automatic gene prediction algorithms, and manual gene annotations. The outcome usually is neither perfect nor precise. Moreover, there is no common standard on how to predict genes. Even for our own species we do not know the exact number. Nevertheless, despite inherent uncertainties every genome project reveals the presence of genes that encode proteins without significant BLASTP hits in the databases. The precise number in the genome of a given species depends on the approach used to identify orphans. However, independently of the approach and the phylogenetic position of a species, orphan genes comprise about 10–20% of all genes in a genome. The percentages of orphan genes in 30 published genomes are summarized in Figure I.

It is surprising that orphans have been completely ignored by most comparative genomics studies. Comparative genomics aims to classify genes, a task quite similar to that of the classical taxonomist, but where species are substituted by genes. Taxonomy is based on the identification of morphological features which differ between closely related species. These minor differences enable the classification of species. By contrast, animals are grouped into larger systematic units (genera and higher systematic categories) based on the identification of shared traits and commonalities. Thus, similarities and differences can be viewed as two sides of an evolutionary process. Balanced and careful consideration of both commonalities and differences is vital for any proper comparative approach. We believe that (as in case of taxonomy) comparative genomics requires balanced consideration of both commonalities and differences. Oddly enough, most attention at present has been paid to genes which are shared and highly conserved throughout evolution, and not to those which are unique or lineage-specific. Taxonomists are fascinated when they manage to identify a new species; molecular biologists, on the contrary, seem to be rather bemused when stumbling on 'novel' genes.

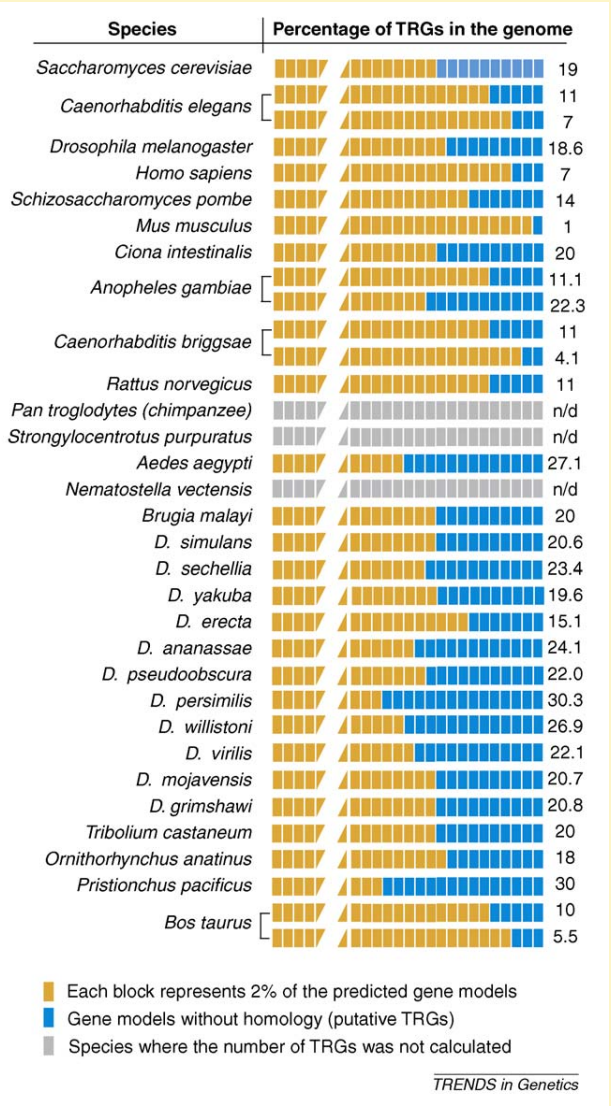

Figure I. Percentages of orphan or taxonomically-restricted genes (TRGs) in 30 animal genomes. Orphan genes comprise about 10–20% of every genome independently of the phylogenetic position of a species. The percentage of orphan genes may vary depending on selected BLAST search criteria (note alternative calculations for *C. elegans*; *Anopheles gambia*; *C. briggsae* and *Bos taurus*). For example, 11% of proteins in *C.briggsae* have no significant BLASTP hits with $E < 10^{-10}$, but the percentage of non-hits decreases to 4.1% if a cut-off value of $E < 10^{-5}$ is used [36]. Species are listed according to the date of publication of the corresponding genome sequence paper. References and information regarding BLAST settings are summarized in Supplementary Table 1.

consist of a large number of complete and high-quality genome sequences representing species evenly distributed across various branches of the phylogenetic tree. Unfortunately, coverage is still very poor and patchy in the animal domain of life. Currently-available complete genome sequences do not fully represent the real evolutionary diversity of the animal kingdom. By contrast, hundreds of microbial genomes are available for comparative analysis. As of June 2009, 897 complete microbial genomes had been deposited in NCBI while a further 1793 are in the process of sequencing (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). It is therefore unsurprising that several important concepts concerning 'orphan' genes originate from microbiology. For example, the idea that the gene contents of bacterial genomes are variable between species is well accepted and this is no longer considered controversial [26,27]. Moreover, comparative analyses of 122 bacterial genomes clearly indicates that the number of bacterial orphan genes is increasing linearly and is likely to continue to increase in the future [10].

Analysis of bacterial genomes demonstrates that closely related species often share 'orphan' genes, or families of 'orphan' genes, that are not present elsewhere. This gave rise to the idea that 'orphans' are, in reality, taxonomically-restricted genes. Groups of homologous 'orphan' genes can be present in several closely related organisms, but not in any others (Box 2). This prompted a proposal to use the more appropriate term 'taxonomically-restricted' genes (TRGs) instead of 'orphans' [10,11].

What is true for genome evolution in bacteria is most likely also true for eukaryotic organisms. Indeed, the rate at which new protein families have been identified during genome sequencing [28] supports the view that TRGs are important. The results of Kunin et al. contradict the idea that there is a limited number of protein families and are instead consistent with reports of unique proteins (encoded by TRGs) in every newly published genome [28]. Although the number of sequences deposited in GenBank is increasing at exponential rate, the proportion of genes that show no similarity to previously sequenced genes remains at

Box 2. Orphan or taxonomically-restricted?

At present the animal domain of life is represented by only a small number of completely sequenced genomes. Although single representatives for most phylogenetic groups have been already sequenced, coverage within different groups of animals remains poor with rare exceptions in insects, round worms and higher vertebrates. Each sequenced genome harbors 'orphan' or 'novel' genes. How will this category of genes change in the course of further genome sequencing when more high quality genome sequences become available from closely-related species?

Let us imagine that we are interested in the comparative genome analysis of nine species representing three genera (see Figure I). Initial sequencing of three species belonging to three genera identifies three orphan genes (A, B and D). Genome sequencing of all representatives of these three genera reveals that some of the genes are restricted to one species (genes B and E), which are orphans in the strict sense, while others show different patterns of distribution among the species. Some genes may be also represented by paralogs (D1, D2 and D3). This scenario predicts that large proportion of genes nowadays referred to as 'orphan' will turn out to be shared by closely-related species within a genus, but will still not be detected in more distantly related organisms. Such a mode of distribution would reflect the evolutionary origin of non-overlapping sets of 'novel' genes in response to different selection pressures imposed onto species that occupy different ecological niches. Our example demonstrates the advantage and flexibility of the TRG concept over the previously used 'orphan gene' definition.

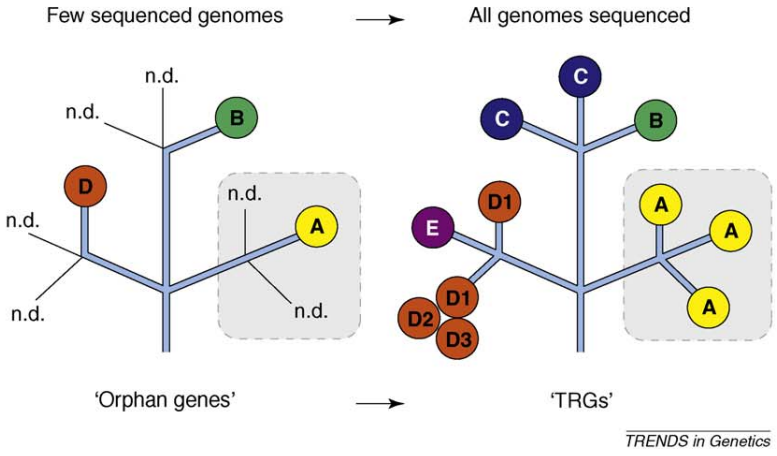

Figure I. A hypothetical tree representing nine species belonging to three genera. The 'orphan' genes are labelled A, B, C, D and E. Genomes of three species are sequenced in (a) and three orphan genes (A, B and D) are identified. Genomes of all nine species sequenced are shown in (b). As a result, additional orphan genes (C, D1-D3 and E) are identified and the true distribution of genes across the three genera is revealed. Grey box, three species of one genus; A, B, C, D, D1, D2, D3, E, orphan genes; D1, D2, D3, paralogous genes; n.d., species where the genome has not been sequenced; TRGs, taxonomically-restricted genes.

10–20%, depending on the cut-off threshold used in BLASTP similarity searches. It is a reasonable assumption that the number of TRGs will not level off and would only approach saturation once sequence data become available for all extant species.

Uncovering TRGs: a historical perspective

The publications of the first few genomes (*Saccharomyces cerevisiae* [29]; *Caenorhabditis elegans* [30]; *Drosophila melanogaster* [31], *Arabidopsis thaliana* [32], and *Homo sapiens* [33,34]) from 1997 to 2002 represented an enormous step forward and allowed comparisons of complete gene sets from several eukaryotic organisms. The obvious increase in genome complexity and total gene number from fly to human was apparent. New genes responsible for advanced functions and structures were shown to have emerged during evolution. For example, genes encoding T-cell receptors, major histocompatibility complex (MHC) molecules, and antibodies appeared to represent TRGs restricted to jawed vertebrates. However, the number of sequenced genomes at that time was few, the phylogenetic distances between sequenced organisms were large, and therefore their genomic and morphological complexity were not easily comparable.

The sequencing of the *Schizosaccharomyces pombe* genome [35] in 2002 enabled the genomes of two species of similar morphological and physiological complexity to be compared. One interesting finding was that 14% (n = 681) of the *S. pombe* protein-encoding genes could not be found in *S. cerevisiae* and, conversely, 19% (n = 1104) of protein-encoding genes appeared to be unique to *S. cerevisiae*. Surprisingly, these genes were not referred to as 'lineage-specific' or 'taxonomically-restricted'; nor were the potential implications discussed in the original genome paper [35]. It is not unreasonable to assume that most taxonomists would conclude that species-specific genes (read 'apomorphic features') are most probably responsible for the specific adaptations of the two species to their ecological settings, and therefore might be responsible for the profound differences in morphology and physiology between *S. pombe* and *S. cerevisiae*. Molecular biologists, however, did not think that way and concentrated on the genes shared by both species.

With the sequencing of *Caenorhabditis briggsae* in 2003 [36] it became possible to explore the genetic differences and similarities between two closely related species, *C. elegans* and *C.briggsae,* which are barely distinguishable morphologically despite having diverged ~100 million years ago. Differences in gene content between the two species were carefully documented and the issue of orphan genes was assessed: 2108 genes in *C. elegans* (11%) and 2141 genes in *C. briggsae* (11%) did not have any

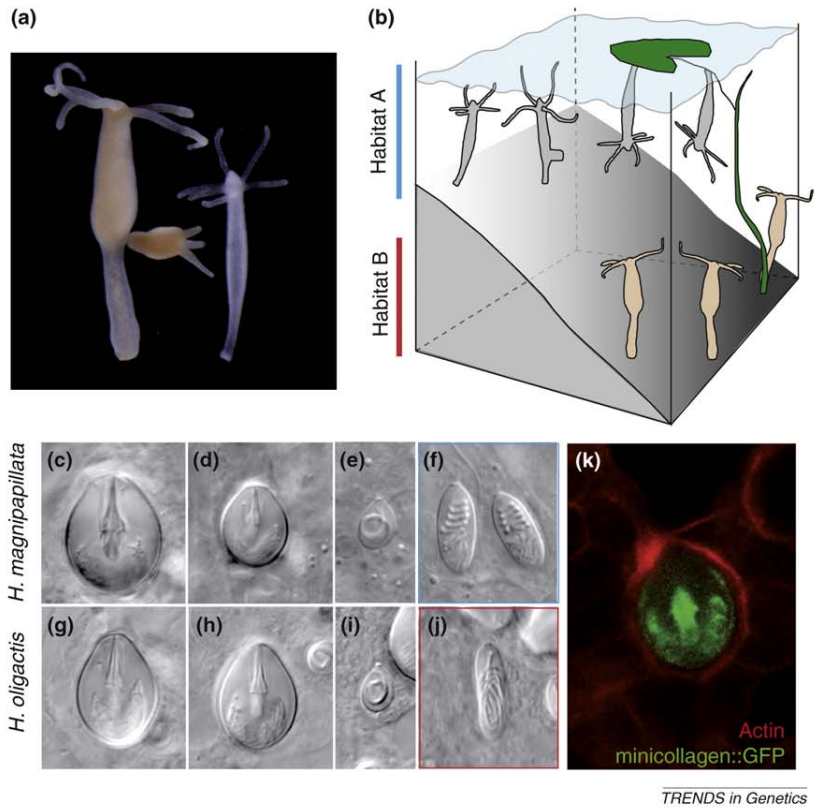

Figure 2. Representatives of the genus *Hydra* differing morphologically and physiologically. **(a)** Photographs of *H. oligactis* (left) and *H. magnipapillata* (right). **(b)** The two *Hydra* species have different environmental preferences [67,68]; *H. magnipapillata* lives close to the surface (Habitat A) whereas *H. oligactis* occupies a deeper habitat (depicted as Habitat B here). TRGs might contribute to the diversity of structures that are adapted for different food sources in a particular ecological setting. For example, over 60 TRGs in *Hydra* are expressed in cnidocytes, a cell type exclusively restricted to Cnidaria and commonly used for systematic distinction. **(c)**–**(f)**, Cnidocytes of *H. magnipapillata;* **(g)**–**(j)** cnidocytes of *H.oligactis*. Note that the two species can be distinguished on the basis of the coils in holotrichous isorhizas: in *H. magnipapillata* holotrichous isorhizas have transverse coils **(f)** while holotrichous isorhizas in *H. oligactis* **(j)** are without transverse coils. **(k)** A stenotele expressing eGFP under control of the promoter of TRG *nb001*; modified from Ref. [41]. *nb001* encodes minicollagen, a short secreted protein that is one of the major components of the cnidocyte capsule. In total 17 different minicollagen genes are present in the genome of *Hydra magnipapillata* [46].

significant BLASTP hit (E value $< 10^{-10}$) in the genome of the other species, and represent putative species-specific genes that emerged in one of the two genomes since they diverged [36]. It was proposed that these genes are quickly evolving and might reveal sites of rapid evolution.

Further advances in genome sequencing have brought new surprises and do not support initial expectations of a gradual increase in complexity and gene number from 'simple' to more 'advanced' animals. In particular, EST (expressed sequence tag) and genome data from the sea anemone *Nematostella vectensis* and the anthozoan coral *Acropora millepora* (all Cnidarians) revealed that basal metazoans possess most of the gene families found in bilaterians and have retained many ancestral genes that have been lost from *Drosophila* and *C. elegans* [4–6,37]. Cnidarians, therefore, are much more 'human-like' than flies and worm in terms of their gene content [4–6].

Recent analyses from the 12 *Drosophila* genomes project [38] revealed that 15–30% of genes do not show any significant sequence similarity to *D. melanogaster* gene models (see Figure I in Box 1; Figure 2 and Supplementary Table 9 in Ref. [38]), suggesting that they are species-specific. Moreover, 2.5% of genes (n = 296) were not detected at the root of the *Drosophila* genus and therefore most likely arose *de novo* within this animal group. The issue of species-specific genes was barely mentioned in the corresponding paper [38] and no evolutionary implications were discussed, which is perhaps surprising given previous studies on rapidly evolving and 'orphan' genes in *Drosophila*. In 1997 an unprecedentedly high rate of gene evolution in *Drosophila* was reported by Schmid and Tautz [24] who later on, with Aquadro and Domazet-Loso, conducted the first systematic screening for rapidly evolving and 'orphan' genes in this species [18, 25]. These studies led to a model of orphan gene evolution based on gene duplications [18]. Recent studies describing *de novo* gene evolution from ancestral noncoding DNA have complemented this model by a new mechanism of gene emergence [39].

Every group of animals seems to have ~10–20% of lineage-specific genes, reflecting the early results in Hydra [40]. TRGs are ubiquitous, but what is the function of these genes? It is here that the stinging cells of cnidaria provide an important clue.

TRGs and the evolution of a cnidarian-specific structure

Cnidaria (corals, jellyfishes and polyps) represent the simplest animals at the tissue level of organization (Figure 1). However, despite of their morphological simplicity they also possess what might be one of the most sophisticated and complex of all cell types in the animal kingdom – stinging

cells. These cells, which can shoot tubule and inject toxic substances into their targets, have evolved within Cnidaria to facilitate the capture of prey. These 'stinging cells' (termed cnidocytes or nematocytes) are exclusively restricted to cnidarians and are considered to be the predominant synapomorphic feature of this phylum. Cnidrian species display a remarkable degree of diversity in cnidocyte shapes and sizes which appear to have been adapted to different food sources. Do TRGs play a role in the generation of this taxon-specific cell type?

In *Hydra* (Figure 2), three different types of nematocytes (stenoteles, desmonemes and isorhiza) are formed as derivatives of the multipotent interstitial stem cell lineage (reviewed in Ref. [41]). All known Hydra species share similar stenoteles and desmonemes whereas differences in a subtype of isorhiza (holotrichous isorhiza; Figure 2f and j) serve as distinguishing characteristics between different hydra groups [42]. What are the relative contributions of 'conserved' and 'novel' genetic components in generating a phylum-specific structure? The differentiation of nematocytes from multipotent stem cells requires the activity of several conserved transcription factors including the *zic/odd-paired* homolog *HyZic* [43] and the *achaete-scute* homolog *Cnash* [44]. However, the structural set-up of capsules seems to rely mostly on several novel cnidarian-specific proteins. Of 51 nematocyte-specific genes identified as a result of a large scale gene expression analysis in *Hydra magnipapillata,* 41 (80%) lacked detectable orthologs in other metazoan phyla [45]. The major structural constituents of the nematocyte capsule are short novel proteins, termed minicollagens, which contain a collagen-like domain flanked by two cysteine-rich domains. Covalently bound to an additional capsule protein, NOWA, they stabilize the capsule wall [46]. Conserved protein domains in NOWA such as SCP and C-type lectin are combined in a configuration not yet identified in any protein outside the Cnidaria. At least 17 minicollagen genes are present in the *Hydra magnipapillata* genome and homologous proteins have been identified in the coral *Acropora*, sea anemones *Metridium* and *Nematostella*, and in several hydrozoans – *Hydractinia*, *Podocoryne* and *Clytia* [46]. Another novel protein, spinalin, is a structural component of spines inside the capsule [47]. Both minicollagens and spinalin are found only in Cnidaria, representing TRGs with an easily inferable function.

What genetic machinery is required to build such a sophisticated cell type? A suppression subtractive hybridization (SSH) approach identified genes that are expressed during cnidocyte development [41]. In the analysis of the resulting cDNA library (enriched for genes expressed in interstitial cells and their derivatives) we identified 16 genes without homologs in other organisms. *In situ* expression analysis of these putative taxonomically-restricted genes showed that most are expressed in developing stinging cells. Although the majority of these genes could be detected in all types of cnidocytes, some were restricted to specific types, e.g. developing stenoteles or isorhiza. A comparison between two Hydra species (*H. magnipapillata* and *H. oligactis*; Figure 1 and Figure 2a) and the distantly related sea anemone *Nematostella vectensis* revealed that most of TRGs identified were restricted to *Hydra*. Homologs of two genes could be identified in the sea anemone genome but not in any other animals outside cnidaria [41], suggesting that these two genes represent phylum-specific TRGs.

Detailed analysis of the identified cnidocyte specific genes at the genome and transcriptome level has shed light on possible mechanisms of TRG evolution and regulation [41]. Most TRGs show striking genomic complexity and extensive diversification by splicing. Conventional and unconventional splice-sites are used, giving rise to a manifest nematocyte-specific transcriptome. Moreover, two related (but non-identical) copies of the nematocyte-specific gene *nb039* are located in different regions of the genome. This could indicate that novel proteins which contribute to the genomic complexity of nematocyte-specific TRGs originated as a result of gene duplication, supporting the model for TRG evolution proposed by Domazet-Loso and Tautz [18].

How is the transcription of a newly-evolved gene regulated? To address this question the 5′-flanking region of one of the genes investigated, *nb001*, was used to generate eGFP^{+} transgenic animals. A reporter construct with 1 kb of the *nb001* 5′-flanking region faithfully recapitulated the endogenous expression of *nb001* in developing cnidocytes (Figure 2k). Interestingly, the *nb001* promoter lacks any detectable conserved *cis*-regulatory elements, pointing to the existence of novel transcription factor binding sites or even novel taxonomically-restricted transcription factors. Such experiments open a new perspective on TRGs and might help to identify associated transcriptional machinery.

Taken together, the invention of a new morphological feature such as the cnidocyte seems to be tightly interlinked with the evolution of TRGs. Different *Hydra* species preferentially live in different habitats and most likely encounter different types of prey (Figure 2b). Thus, TRGs might contribute to the diversity of structures (cnidocytes) that are adapted for different food sources in a particular environment.

TRGs and the generation of morphological diversity in closely related species

The *Hydra* peptide project [48–50] together with unbiased screening approaches in *Hydra magnipapillata* [40,51] and *Hydra vulgaris* AEP [52] revealed a number of *Hydra* proteins that lack homologs in other animals. A substantial proportion of 'novel' genes detected in the *Hydra* peptide project encode neuropeptides that have no metazoan homologs. The function of several of these has been elucidated. Neuropeptide Hym-355 [AB025945], for example, enhances neuronal differentiation by inducing multipotent interstitial stem cells to enter the neuron differentiation pathway [53]. Peptides and short secreted proteins produced by epithelial cells comprise another group of TRGs. One interesting example is the head-specific secreted protein ks1 [X78596] that is expressed in ectodermal epithelia cells. RNAi interference experiments have shown that this gene is important for proper tentacle formation and differentiation of the battery cells – unique and typical cnidarian morphological structures [54].

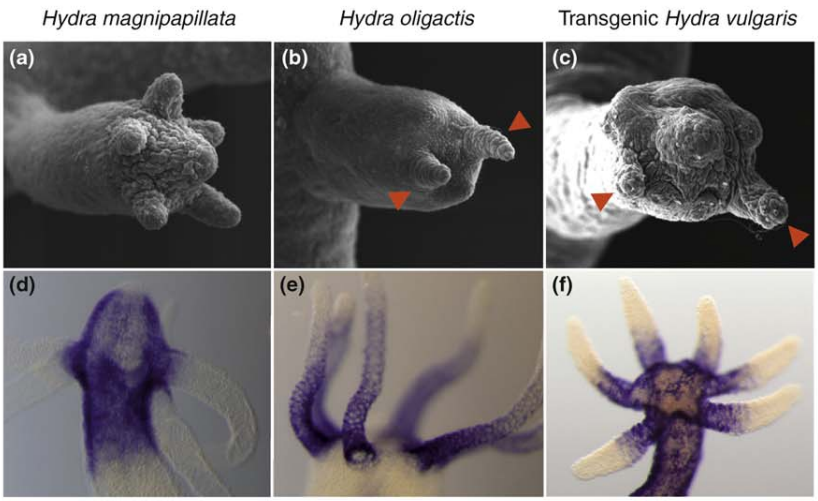

Figure 3. Two closely related species *H. oligactis* and *H. magnipapillata* that differ in their mode of tentacle formation. In *H. magnipapillata* **(a)** five or four tentacles develop symmetrically and simultaneously while in *H.oligactis* **(b)** two laterally located tentacles appear first. Development of tentacles that are needed to catch and take up food is influenced by TRG *Hym301*. The pattern of tentacle formation in *Hydra vulgaris* can be made asymmetric by overexpression of *Hym301* gene in tentacles **(c)**. Thus, TRGs play a role in generating morphological diversity in *Hydra*. Representative scanning electron micrographs of bud evagination in *H. magnipapillata, H. oligactis* and transgenic *H. vulgaris* AEP^{A14} overexpressing *mHym301A* are shown in a–c. Note the slightly staggered order in which tentacles arise. **(d)** Expression of the *mHym301A* gene in *H. magnipapillata*; **(e)** Expression of the *oHym301A* gene in *H. oligactis*. **(f)** Transgenic *H. vulgaris* AEP^{A14} overexpressing *mHym301A*. *mHym301A* transcripts are present in ectodermal epithelial cells all over the body column and in tentacle epithelial cells; modified from Ref. [41]. **(d-f)** Localization of *Hym301* transcripts by *in situ* hybridization. Orange arrowheads indicate asymmetric tentacles in *H.oligactis* **(b)** and in transgenic *H.vulgaris* AEP **(c)**.

The development of transgenic technology in *Hydra* [55] has provided an opportunity to assess the *in vivo* function of TRGs. For example, RNAi knockdown experiments of the epitheliopeptide Hym-301, initially identified by the Hydra peptide project [48], provided hints that *Hym-301* gene is involved in tentacle formation [56]. Unbiased SSH screening for genes that differ in sequence and expression between *H. magnipapillata* and *H.oligactis* led to the identification of additional *Hym301*-like genes and demonstrated a correlation between their expression patterns and the modes of tentacle formation in these two species [57] (Figure 3). The expression of *Hym301* in the tentacles of *H. oligactis* correlates with the formation of two long and functional tentacles prior to the appearance of the other tentacles (Figure 3 b and e). By contrast, *H. magnipapillata* and *H. vulgaris* express *Hym301* genes in the tentacle formation zone but not in the tentacles themselves, and four or five short tentacles develop simultaneously (Figure 3a and d). Overexpression of *Hym301* gene in tentacles of *H.vulgaris AEP* (in the pattern typical of another species – *Hydra oligactis*) induced changes in morphology that mirror the phenotypic differences observed between species – asymmetric appearance of tentacles during budding and regeneration (Figure 3c and f) [57].

What is the selective advantage of having different modes of tentacle formation? Tentacles are the only structures which allow *Hydra* polyps to catch and take up food. The minor differences in their shape, number, and arrangement might permit access to different ecological niches and therefore must be under strict selective constraints. The remarkable degree of diversity in tentacle development might therefore indicate adaptation to different food sources. TRGs thus seem to influence the development of structures to make them maximally effective for exploiting particular types of food. We propose that differences in *Hym301* regulation constitute a principal source of molecular variation that can be acted on by natural selection. The way in which natural selection fine-tunes the expression of *Hym301* genes or their gene regulators remains to be elucidated (Figure 4).

TRGs: effector molecules in the Hydra defence system?

Hydras are soft-bodied animals that lack migratory phagocytic cells, hemolymph, and impermeable barriers (such as a cuticle or an exoskeleton), resulting in seemingly high vulnerability to pathogens. Because the animals are constantly exposed to numerous microbes the epithelium is well-equipped to prevent infectious agents from entering the body [58–60]. An inducible defence system, mediated by the epithelial cells, is activated following pathogen invasion; this activation is driven by the increased expression of genes encoding antimicrobial peptides [58]. Strikingly, most of these antimicrobial peptide genes show no sequence similarity to genes in other species [58]. Periculin-1, named for its rapid response to a wide variety of bacterial and tissue 'danger' signals [58], provides an intriguing example. The deduced Periculin-1 amino acid sequence and the charge distribution within the molecule revealed an anionic N-terminal region and a cationic C-terminal region containing 8 cysteine residues [58]; identifiable orthologs are not present in sequence databases. Periculin-1 has strong bactericidal activity and is expressed in the endodermal epithelium as well as in a subpopulation of ectodermal interstitial cells. Screening efforts across diverse taxa reveal that each animal species contains a significant number of such 'orphan' genes which encode potent antimicrobial peptides. For example,

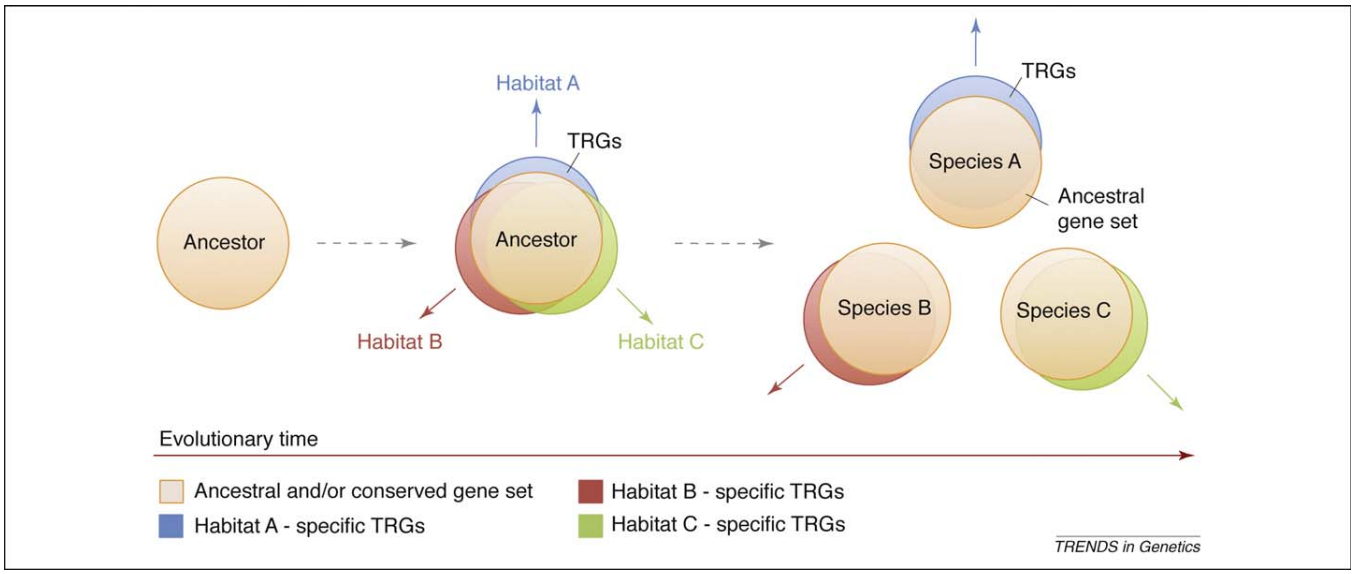

Figure 4. TRGs might help to cope with new ecological niches and changing environments. Environmental change could result in the emergence of taxonomically-restricted genes (TRGs) that over evolutionary time contribute to adaptation towards a new optimum. Ancestral species diverge into three new species that adapt to different environments (e.g. Habitats A, B and C). As a result, each of three species will accumulate a subset of their own TRGs that are not present in other species.

Aurelia aurita, a common jellyfish, produces aurelin, a 40 aa-residue novel antimicrobial peptide [61]. Similarly, the antibacterial immune response genes *diptericin* and *attacin* are restricted to the group of Diptera [18,62]; and the 11 kDa metal ion–binding S100 protein psoriasin uniquely protects mammalian epithelia from infection [63]. Taxonomically-restricted host defence molecules represent an extremely effective chemical warfare system that facilitates the disarming of taxon-specific microbial attackers and, at the same time, shapes the colonizing microbiota.

Concluding remarks

Organisms under natural conditions do not live in isolation, but have evolved, and continue to evolve, in the context of complex communities and specific environmental conditions. However, the mechanisms by which environmental factors are integrated at the molecular level remain unclear. Organisms must adjust both their morphology and the ways in which they interact with other organisms in order to cope with new niches and changing environments. As there are endless numbers of ecological niches, endless numbers of adaptations must exist; species-specific features will reflect these habitat-specific adaptations. TRGs might appear and continuously evolve to mediate these lineage-specific adaptations. When changes in ecological settings favor an individual that is different from the mean individual in a population, TRGs can have a critical effect on adaptation towards the new optimum (Figure 5). In this scenario, over evolutionary time, TRGs seem to have modified major components of the innate immune defence system, have provided important structural components of evolutionary novelties, and have been responsible for the generation of morphological diversity within closely related species.

Although our understanding of the role of TRGs as effector molecules of major physiological and morphological changes continues to expand, and our technological abilities to uncover them in a wide range of closely related species living in different habitats has improved, there are several unresolved issues that need to be addressed in the future.

An issue of great complexity and almost completely uncharted territory concerns the evolution of TRGs. Mechanisms that create TRGs still remain obscure. There are at least two possible scenarios. First, TRGs might be created

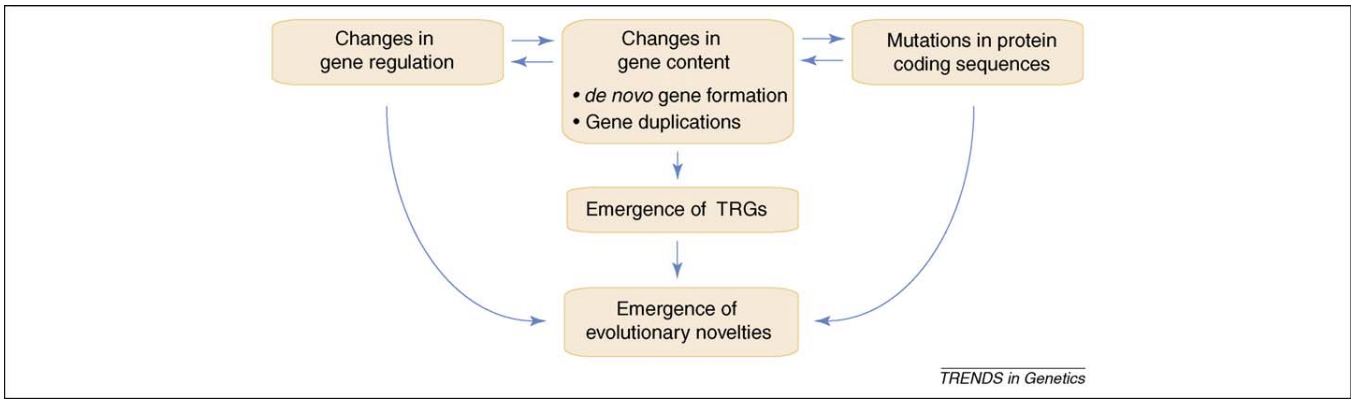

Figure 5. A brief outline of landmark processes involved in the emergence of evolutionary novelties. Several of these processes act simultaneously and cooperatively (i.e. cis-regulatory changes, gene duplications and the emergence of genes *de novo*).

de novo from non-coding regions of the genome by translocation of a DNA segment containing a random ORF into a transcriptionally-active location. It has been remarked that 'novel' genes often encode short secreted proteins (~100 amino acids in length). This would be expected in genes that arise *de novo* because of the low probability of randomly generating long random ORFs. It is possible that the activity of transposable elements (TE) is the driving force behind this mode of *de novo* gene formation. In support of this view, more than half of the orphan genes in primates were found to contain TE-like sequences [12]. However, ~5% of primate orphan genes (62 genes) did not contain any traces of TEs. Interestingly, most of these genes were found in genomic regions that are syntenic with other mammalian genomes. The corresponding genomic regions from non-primates do not contain any sequence which, when artificially translated, would encode a primate protein. Thus, the corresponding ORFs must have originated from distant regions of the genome by as yet unknown mechanisms [12].

The second plausible scenario is that TRGs emerge as a result of gene duplications. These are known to provide raw genetic material for evolutionary novelties [64,65]. Gene copies generated by duplication can remain active and functionally identical to the original, become a pseudogene, or diverge and obtain a new function. It is therefore reasonable to propose that TRGs are rapidly evolving genes that appear as a result of gene duplication followed by the period of rapid sequence divergence [18]. In this scenario, once a certain evolutionary time has elapsed, sequence similarity to 'parent' genes will become undetectable by current bioinformatics algorithms. Such mode of TRG generation was initially proposed to explain the high number of orphan genes in *Drosophila* [18]. Indeed, most *Drosophila* 'orphans' belong to the group of rapidly evolving genes showing higher substitution rates than non-orphans [18]. However, there are several interesting exceptions: some 'orphan' genes in the *Drosophila* lineage show divergence rates that are extremely low and are comparable to the rates of divergence of proteins that are highly conserved across the animal kingdom. An excellent example is the *flightin* gene (Z18858) that encodes a myofibrillar protein of the flight muscle [66]. *Flightin* is clearly involved in taxon-specific adaptation and is present only in insects – mutation of the gene in *Drosophila* has no effect on fecundity or viability but results in loss of flight ability.

The emergence of evolutionary novelties appears to stem from several processes that act simultaneously and cooperatively: cis-regulatory changes, gene duplications, and the emergence of genes *de novo* (Figure 5). Our understanding of the evolutionary significance for TRGs does not overturn a general consensus concerning the role of deep homology and conserved transcription factors in generating diverse adaptations and in the evolution of novelties. Instead, it uncovers previously unknown components, revealing a more complex picture of morphological evolution in early metazoans.

Acknowledgements

Work in our laboratory is supported in part by grants from the Deutsche Forschungsgemeinschaft (DFG), and grants from the DFG Cluster of Excellence programs *The Future Ocean* and *Inflammation at Interfaces*.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2009.07.006.

References

1 Carroll, S.B. *et al.* (2001) From DNA to diversity. (Carroll S, ed). p. 214 London: Blackwell Science
2 Shubin, N. *et al.* (2009) Deep homology and the origins of evolutionary novelty. *Nature* 457, 818–823
3 Kortschak, R.D. *et al.* (2003) EST analysis of the cnidarian *Acropora millepora* reveals extensive gene loss and rapid sequence divergence in the model invertebrates. *Curr. Biol.* 13, 2190–2195
4 Putnam, N.H. *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317, 86–94
5 Miller, D.J. *et al.* (2005) Cnidarians and ancestral genetic complexity in the animal kingdom. *Trends. Genet.* 21, 536–539
6 Technau, U. *et al.* (2005) Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. *Trends. Genet.* 21, 633–639
7 Wilkins, A. (1998) Evolutionary developmental biology: where is it going? *BioEssays* 20, 783–784
8 Prud'homme, B. (2007) Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci. USA* 104, 8605–8612
9 Ohno, S. (1970) *Evolution by gene duplication.* Springer-Verlag, (New York)
10 Wilson, G.A. *et al.* (2005) Orphans as taxonomically restricted and ecologically important genes. *Microbiology* 151, 2499–2501
11 Wilson, G.A. *et al.* (2007) Large-scale comparative genomic ranking of taxonomically restricted genes (TRGs) in bacterial and archaeal genomes. *PLoS One* 2, e324
12 Toll-Riera, M. *et al.* (2009) Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* 26, 603–612
13 Sunagawa, S. *et al.* (2009) Identification and gene expression analysis of a taxonomically restricted cysteine-rich protein family in reef-building corals. *PLoS One* 4, e4865
14 Emerson, J.J. *et al.* (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303, 537–540
15 Betran, E. *et al.* (2006) Fast protein evolution and germ line expression of a Drosophila parental gene and its young retroposed paralog. *Mol. Biol. Evol.* 23, 2191–2202
16 Wang, W. *et al.* (2006) High rate of chimeric gene origination by retroposition in plant genomes. *Plant. Cell.* 18, 1791–1802
17 Wang, H.C. and Hickey, D.A. (2007) Rapid divergence of codon usage patterns within the rice genome. *BMC Evol Biol* 7 (Suppl 1), S6
18 Domazet-Loso, T. and Tautz, D. (2003) An evolutionary analysis of orphan genes in Drosophila. *Genome. Res.* 13, 2213–2219
19 Daubin, V. and Ochman, H. (2004) Start-up entities in the origin of new genes. *Curr. Opin. Genet. Dev.* 14, 616–619
20 Clamp, M. *et al.* (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19428–19433
21 Oliver, S.G. *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature* 357, 38–46
22 Dujon, B. (1996) The yeast genome project: what did we learn? *Trends. Genet.* 12, 263–270
23 Casari, G. *et al.* (1996) Bioinformatics and the discovery of gene function. *Trends. Genet.* 12, 244–245
24 Schmid, K.J. and Tautz, D. (1997) A screen for fast evolving genes from Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9746–9750
25 Schmid, K.J. and Aquadro, C.F. (2001) The evolutionary analysis of 'orphans' from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. *Genetics* 159, 589–598
26 Fukuchi, S. and Nishikawa, K. (2004) Estimation of the number of authentic orphan genes in bacterial genomes. *DNA Res.* 11, 219–231
27 Minezaki, Y. *et al.* (2005) Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. *DNA Res.* 12, 269–280
28 Kunin, V. *et al.* (2003) Myriads of protein families, and still counting. *Genome. Biol.* 4, 401
29 Goffeau, A. *et al.* (1997) The yeast genome directory. *Nature* 387 (6632 Suppl), 5

30 *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode C.elegans: a platform for Investigating biology. *Science* 282, 2012–2018
31 Adams, M. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195
32 Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
33 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
34 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
35 Wood, V. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415, 871–880
36 Stein, L.D. *et al.* (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS. Biol.* 1, E45
37 Kusserow, A. *et al.* (2005) Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* 433, 156–160
38 Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450, 203–218
39 Levine, M.T. *et al.* (2006) Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* 103, 9935–9939
40 Bosch, T.C.G. and Khalturin, K. (2002) Patterning and cell differentiation in Hydra: novel genes and the limits to conservation. *Can. J. Zool.* 80, 1670–1677
41 Milde, S. *et al.* (2009) Characterization of taxonomically-restricted genes in a phylum-restricted cell type. Characterization of taxonomically-restricted genes in a phylum-restricted cell type. *Genome. Biol.* 10, R8
42 Hemmrich, G. *et al.* (2007) Molecular phylogenetics in Hydra, a classical model in evolutionary developmental biology. *Mol. Phylogenet. Evol.* 44, 281–290
43 Lindgens, D. *et al.* (2004) Hyzic, the Hydra homolog of the zic/odd-paired gene, is involved in the early specification of the sensory nematocytes. *Development* 131, 191–201
44 Grens, A. *et al.* (1995) Evolutionary conservation of a cell fate specification gene: the Hydra achaete-scute homolog has proneural activity in Drosophila. *Development* 121, 4027–4035
45 Hwang, J.S. *et al.* (2007) The evolutionary emergence of cell type-specific genes inferred from the gene expression analysis of Hydra. *Proc. Natl. Acad. Sci. U. S. A.* 104, 14735–14740
46 David, C.N. *et al.* (2008) Evolution of complex structures: minicollagens shape the cnidarian nematocyst. *Trends. Genet.* 24, 431–438
47 Hellstern, S. *et al.* (2006) Structure/function analysis of spinalin, a spine protein of Hydra nematocysts. *FEBS J.* 273, 3230–3237
48 Takahashi, T. *et al.* (1997) Systematic isolation of peptide signal molecules regulating development in hydra: LWamide and PW families. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1241–1246
49 Bosch, T.C.G. and Fujisawa, T. (2001) Polyps, peptides and patterning. *BioEssays* 23, 420–427
50 Fujisawa, T. (2008) Hydra peptide project 1993–2007. *Dev. Growth Differ.* 50 (Suppl 1), S257–S268
51 Weinziger, R. *et al.* (1994) Ks1, an epithelial cell specific gene, responds to early signals of head formation in Hydra. *Development* 120, 2511–2517
52 Genikhovich, G. *et al.* (2006) Discovery of genes expressed in Hydra embryogenesis. *Dev. Biol.* 289, 466–481
53 Takahashi, T. *et al.* (2000) A novel neuropeptide, Hym-355, positively regulates neuron differentiation in Hydra. *Development* 127, 997–1005
54 Lohmann Jan, U. *et al.* (1999) Silencing of developmental genes in Hydra. *Dev. Biol.* 214, 211–214
55 Wittlieb, J. *et al.* (2006) Transgenic Hydra allow in vivo tracking of individual stem cells during morphogenesis. *Proc. Natl. Acad. Sci. U. S. A.* 16, 6208–6211
56 Takahashi, T. *et al.* (2005) Hym-301, a novel peptide, regulates the number of tentacles formed in hydra. *Development* 132, 2225–2234
57 Khalturin, K. *et al.* (2008) A novel gene family controls species-specific morphological traits in Hydra. *PLoS Biol.* 6, e278
58 Bosch, T.C.G. *et al.* (2009) Uncovering the evolutionary history of innate immunity: the simple metazoan hydra uses epithelial cells for host defence. *Dev. Comp. Immunol.* 33, 559–569
59 Jung, S. *et al.* (2009) Hydramacin-1: Structure and antibacterial activity of a protein from the basal metazoan hydra. *J. Biol. Chem.* 284, 1896–1905
60 Augustin, R. *et al.* (2009) Identification of a kazal-type serine protease inhibitor with potent anti-staphylococcal activity as part of hydra's innate immune system. *Dev. Comp. Immunology* 33, 830–837
61 Ovchinnikova, T.V. *et al.* (2006) Aurelin, a novel antimicrobial peptide from jellyfish Aurelia aurita with structural features of defensins and channel-blocking toxins. *Biochem. Biophys. Res. Commun.* 348, 514–523
62 Lemaitre, B. and Hoffmann, J. (2007) The host defense of Drosophila melanogaster. *Annu. Rev. Immunol.* 25, 697–743
63 Gläser, R. *et al.* (2004) Antimicrobial psoriasin (S100A7) protects human skin from Escherichia coli infection. *Nat. Immunol.* 6, 57–64
64 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
65 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404
66 Reedy, M.C. *et al.* (2000) *Flightin* is essential for thick filament assembly and sarcomere stability in *Drosophila* flight muscles. *J. Cell. Biol.* 151, 1483–1500
67 Bosch, T.C.G. *et al.* (1988) Thermotolerance and synthesis of heat shock proteins: these responses are present in *Hydra attenuata* but absent in *Hydra oligactis*. *Proc. Natl. Acad. Sci. U. S. A.* 85, 7927–7931
68 Gellner, K. *et al.* (1992) Cloning and expression of an heat inducible hsp70 gene in two species of hydra which differ in their stress response. *Eur. J. Biochem.* 210, 683–691