Evolution of New Functions De Novo and from Preexisting Genes

Dan I. Andersson, Jon Jerlström-Hultqvist, and Joakim Näsvall

Department of Medical Biochemistry and Microbiology, Uppsala University, SE-75123 Uppsala, Sweden

Correspondence: dan.andersson@imbim.uu.se

How the enormous structural and functional diversity of new genes and proteins was generated (estimated to be 10^{10} – 10^{12} different proteins in all organisms on earth [Choi I-G, Kim S-H. 2006. Evolution of protein structural classes and protein sequence families. *Proc Natl Acad Sci* **103**: 14056–14061]) is a central biological question that has a long and rich history. Extensive work during the last 80 years have shown that new genes that play important roles in lineage-specific phenotypes and adaptation can originate through a multitude of different mechanisms, including duplication, lateral gene transfer, gene fusion/fission, and de novo origination. In this review, we focus on two main processes as generators of new functions: evolution of new genes by duplication and divergence of pre-existing genes and de novo gene origination in which a whole protein-coding gene evolves from a noncoding sequence.

How new genes emerge and functionally diversify are very fundamental questions in biology, as new genes provide the raw material for evolutionary innovation that allows organisms to adapt, increase in complexity, and form new species. An organism can acquire new genes through at least three distinct, but potentially overlapping, mechanisms (Fig. 1). Thus, a pre-existing gene can be transferred ready made from another organism by lateral gene transfer (via transformation, transduction, and conjugation), or it can evolve by modification of an already existing gene (by duplication–divergence or gene fusion/fission) or it can be generated de novo from noncoding DNA. It is clear that these mechanisms have generated the diversity of genes and proteins that underlies the existence of all organisms, but their relative importance in new gene evolution and functional diversifica-

tion is unclear. Thus, their importance will depend on several factors, including the organism and gene studied, the time scales involved (e.g., over recent time scales in the majority of eubacteria lateral gene transfer is a more dominant process than the others), and the methodological problems associated with an unambiguous identification of a gene emerging within an organism (a paralog) or being imported from another organism (a xenolog). In this article, we will focus on the roles of the two latter processes (gene duplication–divergence and de novo origination) as generators of new genes, mainly because they address the basic question of how new genes actually emerge (rather than how functional genes are transferred).

The duplication–divergence concept has a long history, and 80 years ago, it was already suggested by Haldane (1932) and Fisher (1935)

Editor: Howard Ochman

Additional Perspectives on Microbial Evolution available at www.cshperspectives.org

Copyright © 2015 Cold Spring Harbor Laboratory Press; all rights reserved; doi: 10.1101/cshperspect.a017996

Cite this article as *Cold Spring Harb Perspect Biol* 2015;7:a017996

D.I. Andersson et al.

Figure 1. Mechanisms of new gene acquisition. (A) Horizontal gene transfer. The foreign gene (yellow) is transferred from another organism and integrated into the genome by recombination. (B) De novo origination. Mutations in a previously nonfunctional sequence create a new gene (yellow). (C) Duplication–divergence. A duplicate of an ancestral gene (green) acquires a new function and becomes a new gene (yellow).

that new genes evolve from pre-existing ones via gene duplication and subsequent divergence of the extra copy to acquire a new function. Susumu Ohno further developed this idea in his seminal book *Evolution by Gene Duplication* (Ohno 1970) and later work modified and refined his model (Jensen 1976; Piatigorsky 1991; Hughes 1994; Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000; Kondrashov et al. 2002; Bergthorsson et al. 2007; Näsval et al. 2012). Subsequently, comparative genomics, genetics, and biochemistry have convincingly shown that duplication–divergence mechanisms are a major contributor to evolution of new genes (Dittmar and Liberles 2010). In contrast, the de novo origination model of new gene evolution is more recent and less supported. De novo origination of protein-coding genes occurs when genes emerge from a nonfunctional DNA sequence that was previously not a gene. Intuitively, it would seem highly improbable that functional proteins could emerge from noncoding DNA as the DNA must both be transcriptionally active and include a translatable open reading frame (ORF). Furthermore, translation of any random ORF devoid of genes is expected to produce insignificant polypeptides rather than proteins with specific functions. Indeed it has been argued that de novo origination of new genes is extremely unlikely (Jacob 1977), but despite these claims, the advent of large-scale sequencing and comparative genomics has provided increasing evidence that new genes have evolved and continuously are originating from noncoding sequences (Cai et al. 2008; Knowles and McLysaght 2009; Tautz and Domazet-Lošo

2011; Wu et al. 2011; Carvunis et al. 2012; Wu and Zhang 2013).

MODIFICATION OF PRE-EXISTING FUNCTIONS

Birth and Fate of Duplications

As shown by comparative genomics, evolution of pre-existing genes by duplication and subsequent divergence plays an important role in the emergence of novel genes (Dittmar and Liberles 2010). Depending on the organism and the duplication mechanism, the size of the region of DNA that is duplicated can vary from just a few bases up to whole chromosomes (aneuploidy) or genomes (polyploidy). In this review, we will not discuss duplications of very short regions (bp) that form via slipped-strand mispairing mechanisms or the very large duplications that include whole chromosomes or genomes and that form by nondisjunction (i.e., failure of chromosome/sister chromatid pairs to separate properly during meiosis or mitosis) of chromosome pairs during mitosis or meiosis in the germ line. Spontaneous duplication of regions of intermediate size (kbp–Mbp) is a common process, and experimental determinations of these rates in different eubacteria, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Drosophila melanogaster* suggest that they are in the range of 10^{-7} to 10^{-3} /gene/cell division (for a review, see Katju and Bergthorsson 2013), several orders of magnitude higher than the rate of point mutation per nucleotide.

After birth of a duplicate gene copy, several fates are possible (the three former will reduce the frequency of duplicate genes in the population, whereas the two latter can preserve them): (1) counterselection against individuals with an extra copy caused by a duplication cost, (2) genetic loss of the extra copy by recombination caused by duplication instability, (3) non-functionalization, in which random mutations accumulate in the gene and inactivate it, (4) subfunctionalization, in which initially neutral mutations accumulate in twinned genes to divide the original gene's work between them, and, finally, (5) neofunctionalization, which involves the generation of a novel function by coding sequence changes in one of the gene copies, whereas one copy retains the old function. Which of these processes will dominate is determined by several factors, including the impact of the gene duplicate on organism fitness, the intrinsic instability of the duplication, and the relative rates of inactivating, neutral, and beneficial mutations and population genetic parameters.

Fitness Costs of Duplications

It is often assumed that duplications are cost free and that they are stably inherited, and that their fate is largely determined by the relative rates of the non-, sub- and neofunctionalization pathways. However, there is increasing evidence that these assumptions are incorrect for both eubacteria and eukaryotes. The costs of duplications could manifest at several different levels: (1) costs of duplicated DNA, (2) costs owing to gene expression of RNA and protein, (3) costs owing to the involvement of the expressed protein in an energy-requiring reaction, or (4) costs owing to imbalances in RNA/protein levels that lead to improper gene regulation or unwanted molecular interactions. It is experimentally difficult to tease apart the relative importance of these costs for duplications, but it is likely that the first is of small importance because DNA (and RNA) synthesis constitutes very minor costs as compared with protein synthesis (Neidhardt et al. 1990). Furthermore, to distinguish the cost of (2) from (3) and (4) would require

that the proteins included in the duplication are functionally inactivated to remove costs owing to the normal activity of the protein although still maintaining normal gene expression. Irrespective of the reason for the costs, recent detailed studies in *Escherichia coli* and *Salmonella typhimurium* show that duplications in a size range of 20 kbp to 1246 kbp are associated with costs on the order of $s = 10^{-3}$ per 1 kbp of extra DNA (Pettersson et al. 2009; Adler et al. 2014). As spontaneous duplications range in size from a few kbp up to Mbp in eubacteria (Anderson and Roth 1981; Andersson and Hughes 2009), with similar sizes in *C. elegans*, *D. melanogaster*, and *S. cerevisiae* (Katju and Bergthorsson 2013) and effective population sizes are typically $>10^4$, it is likely the majority of duplications are counterselected (i.e., $s > 1/N_e$). To our knowledge, no detailed experimental analyses of duplication costs have been made in eukaryotes but other lines of indirect evidence support the notion that eukaryotic duplications are also deleterious (Katju and Bergthorsson 2013).

Stability of Duplications

Another factor that will influence the fate of duplications is their intrinsic genetic stability. The stability will largely be determined by whether the duplicated copies can recombine with each other to result in segregational loss of one of the copies. Often, duplications/amplifications are formed by unequal crossing-over mechanisms or rolling-circle amplification that use as recombination regions various types of imperfect or perfect homologous sequences normally present in any genome (Fig. 2). These mechanisms generate tandem arrays that generally are intrinsically unstable because of the presence of long, directly repeated regions with perfect homology (i.e., the duplicated regions) that will allow homologous recombination and segregational loss of the amplified region down to one copy (Anderson and Roth 1977, 1979, 1981; Andersson and Hughes 2009; Hastings et al. 2009). In contrast, if the duplication mechanism generates duplicated copies that are inversely oriented (e.g., by duplication-inversion) (Kugelberg et al. 2010), or where the

D.I. Andersson et al.

Figure 2. Formation and fates of duplications and amplifications. (A) Unequal crossover between two direct repeats (green rectangles) on sister chromatids results in a duplication of the intervening sequence. Unequal exchange between the two copies results in either loss or further amplification. (B) Amplification through rolling circle replication. A double-strand break leads to single strand invasion of a homologous sequence (green rectangle) on the same chromosome. Replication from the site of invasion leads to rolling circle amplification of the sequence between the repeats. Homologous recombination with another chromosome completes the amplification. The thickness of the arrows reflects the rates of duplication (k_{dupl}), amplification (k_{ampl}), and segregation (k_{loss}).

copies are located at widely separated sites (e.g., by retrotransposition), the amplified state can be stabilized.

Duplicate Maintenance

The above description suggests that the most likely outcome for a gene duplicate in a population is its loss or inactivation by recombination, counterselection, or random mutational inactivation (nonfunctionalization), raising the question of how duplicates can be maintained for any longer times in a population. Basically, two types of solutions have been proposed to resolve this problem. One class of models, best exemplified by the subfunctionalization model (also known as the DDC model, duplication–degeneration–complementation), suggests that the duplicates are initially functionally redundant and that they are preserved by neutral mutation processes rather than by selection. Thus,

both gene copies accumulate mutations that will affect the function of each copy, but the two copies together can complement each other and perform the function of the original copy. This functional partitioning over the duplicates can, by purifying the selection, preserve a duplicate, and as this model requires the accumulation and fixation of several successive, initially neutral mutations, it is more likely to occur in organisms with small population sizes (Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000).

Another class of models posits that gene duplicates are maintained by selection for increased gene dosage of a function that increases organism fitness, for example, by conferring detoxification, improved nutrient uptake, or better carbon source utilization (Hartley 1984; Romero and Palacios 1997; Andersson and Hughes 2009; Sandegren and Andersson 2009). Adaptive gene duplication has been widely observed

in all kingdoms of life, and in principle it could maintain the duplicated state to allow for beneficial mutations to accumulate in one duplicate copy (neofunctionalization). However, this model generates another problem. Thus, mutation(s) that diverges one copy to perform a new function will typically decrease selection for preservation of the duplicate because of functional trade-offs between the old and new function. A central problem (designated “Ohno’s dilemma”) is, therefore, how may a duplication be selectively maintained but at the same time free to accumulate mutations and acquire a new beneficial function? (Bergthorsson et al. 2007).

Apart from these models, other types of hybrid models that include both subfunctionalization and neofunctionalization, in which an initial phase of rapid subfunctionalization is followed by prolonged neofunctionalization, have been proposed (Kellis et al. 2004; He and Zhang 2005; Rastogi and Liberles 2005; Byrne and Wolfe 2007; Scannell and Wolfe 2008).

A Solution to Ohno’s Dilemma: The Innovation–Amplification–Divergence Model

The starting point of the innovation–amplification–divergence (IAD) model is a gene that has a weak trace of another function in addition to its main function (Fig. 3) (Hendrickson et al. 2002; Francino 2005; Bergthorsson et al. 2007). For the following explanation of the model, we use the example of a gene encoding an enzyme, but that is not a prerequisite of the model as the function could also be structural or regulatory. The main function is that the enzyme has been selected for through its previous evolutionary history, and the side activity is a fortuitous activity that has not previously been positively selected. The side activity may be catalysis of the same reaction on a different substrate, or a different reaction on the original substrate. If a change in the environment makes this activity beneficial (innovation), positive selection favors mutants that express more of the enzyme. The environmental change could be either external (i.e., the appearance of a new nutrient or a toxic compound) or internal (i.e., fixation of a dele-

terious mutation somewhere else in the genome). As duplications are very common compared with other mutations, positive selection for increasing a weak beneficial activity tends to lead to enrichment of duplications and higher order amplifications in the population (amplification). The fitness cost and inherent instability of tandem duplications are thus outweighed by positive selection, solving two of the problems described above. Purifying selection purges common deleterious mutations from the population, whereas rare beneficial mutations are enriched by positive selection. As the copy number increases, so does the target size where beneficial mutations can occur, facilitating the accumulation of rare beneficial mutations. Improved variants may be further amplified, whereas less-improved variants may be lost (divergence). As beneficial mutations accumulate, positive selection to keep the amplified state is relaxed, until a variant that provide enough activity on its own arises, and the amplification can segregate. If there exists an adaptive conflict between the two functions, and selection to maintain the original activity is present throughout the process, the end result is a new pair of paralogous genes, in which one copy has the new function and the other copy retains the original function.

Innovation

The IAD model presumes that a weak trace of the “new” function is already there by chance in the ancestral gene. All that is needed to start evolving the new gene is a change in the environment that makes this previously unnoticed promiscuous activity beneficial. One of the questions that need to be answered to evaluate the IAD model is “how widespread is promiscuity”? A number of biochemical and genetic evidence supports the idea that many enzymes can catalyze adventitious secondary reactions (i.e., they have promiscuous activities), and that these activities are sometimes high enough to contribute to the fitness of the organism with the proper selective pressure and when overexpressed (Copley 2003; Khersonsky et al. 2006; Khersonsky and Tawfik 2010). For example, various β -lactamases, with

D.I. Andersson et al.

Figure 3. The innovation–amplification–divergence (IAD) model. An ancestral protein possesses a promiscuous side-activity “b” in addition to its main activity “A.” An environmental change makes the b activity beneficial (Innovation). Selection to increase expression of the b activity leads to enrichment of duplications and amplifications (Amplification). Mutant variants (yellow) with improved b activity “B” are selectively amplified, whereas less-improved variants may be lost. When a mutant variant with sufficient B activity appears, selection to maintain the amplification is relaxed, and the amplification segregates. If selection to keep the original A activity is present throughout the process, the end result is a duplication in which one copy has the new B function and the other has the original A function. Positive selection is the driving force for the entire process.

primary penicillin-hydrolyzing activities, can also contribute to decreased susceptibility toward other classes of β -lactam antibiotics (Sun et al. 2009; Adler et al. 2013). Patrick et al. (2007) identified a number of *E. coli* genes that, when overexpressed, could complement the auxotrophy caused by lack of another gene. In 11

cases, the identified genes encoded enzymes that were suggested to act by having a weak promiscuous activity that could replace the missing enzyme by performing the missing enzyme’s reaction, and in an additional six cases, the rescuing genes encoded metabolite transporters that promiscuously imported or exported

a noncognate metabolite. For example, a mutant lacking phosphoserine phosphatase (SerB) could be independently suppressed by overexpression of three different enzymes (phosphoglycolate phosphatase, histidinol phosphatase, and phosphoglyceromutase 2), which all could substitute for SerB in the last step in serine biosynthesis, albeit with poor efficiency (Patrick et al. 2007). Yip and Matsumura (2013) showed that all three of these enzymes could acquire significantly better SerB activity with single amino acid substitutions. Similarly, an *E. coli* mutant blocked in an early step in the synthesis of pyridoxal-5'-phosphate (PLP) from sedoheptulose-7-phosphate and glyceraldehyde-3-phosphate can still generate PLP through two independent mechanisms. One is the action of promiscuous activities in two different glycolytic enzymes that lead to a bypass of the missing enzymatic step (Nakahigashi et al. 2009). The other is through overexpression of either of two enzymes (YeaB or ThrB), that through promiscuous activities produces an intermediate for PLP synthesis in a pathway starting from phosphoserine (Kim et al. 2010). Four sugar kinases with promiscuous activity can rescue *E. coli* cells defunct in the ability to phosphorylate glucose when expressed from strong extrachromosomal promoters. The k_{cat}/k_m values of the promiscuous activities on glucose were 10^4 – 10^5 times lower than the endogenous glucokinase (Miller and Raines 2004, 2005). Single mutations in two of these enzymes increase the glucokinase activity 12- and 60-fold (Larion et al. 2007). In summary, these examples and other studies show that promiscuous activities are common among enzymes.

Apart from the type of enzymatic promiscuity mentioned above, it is also well established that so-called moonlighting, in which a single protein performs more than one function, is a common phenomenon that has been described for hundreds of proteins (Henderson and Martin 2011; Copley 2012; Nam et al. 2012). However, in contrast to adventitious promiscuous reactions, moonlighting enzymes often involve well-evolved activities in which different regions of a protein or alternative protein structures perform efficient and different reactions.

Amplification

Duplications in bacterial genomes are typically arranged in tandem, and often seem to be formed through illegitimate homologous recombination between direct repeat sequences. The frequency of unselected duplications in bacterial populations (at least in *Salmonella enterica*) rapidly reaches a steady state after a few tens of generations. The steady-state frequencies typically range between 10^{-5} and 10^{-2} per cell per gene (Anderson and Roth 1981; Reams et al. 2010). The frequency of any specific duplication is set by the balance between its rate of formation, its mechanistic loss rate, and its inherent fitness cost. As recombination between the copies of a tandem duplication is equally likely to increase the copy number as it is to decrease the copy number, the rate of further increase in copy number once a duplication has been formed is probably as high as the mechanistic loss rate, and will thus reach as high as 10^{-2} /cell/generation. At any given time, up to several percent of a bacterial population will have a tandem duplication somewhere in their genome (Anderson and Roth 1981; Reams et al. 2010; Sun et al. 2012) and among these, a significant fraction will carry amplifications with three or more copies. As the mechanistic loss rate typically is as high as 10^{-2} /cell/generation and most duplications are costly, these copy number variations are expected to be short lived. If any of these transient duplications and amplifications would confer a positive fitness effect that outweighs the cost and mechanistic loss rate, it would be enriched in the population. In eukaryotes, tandem duplications are also very frequent (Gelbart and Chovnick 1979; Shapira and Finnerty 1986; Lam and Jeffreys 2007), and there are additional mechanisms (i.e., retrotransposition that can place a new copy of a gene on a new location in the genome) (Kazazian 2004).

Divergence

The rate of adaptation toward a new function will depend on the rate and magnitude of beneficial mutations in the gene under selection. Several studies of the distribution of fitness ef-

D.I. Andersson et al.

fects of both whole organisms and individual genes have shown that a large majority of mutations are deleterious (for a review, see Gordo et al. 2011). However, the rate and distribution of effects of beneficial mutations have been more difficult to assess by experiments. Experimental studies of fitness effects of mutations at the genome level suggest that beneficial mutation rates can vary between 2×10^{-9} and 4.8×10^{-4} per genome per generation depending on the experimental system and method used to estimate these rates (Imhof and Schlotterer 2001; Hegreness et al. 2006; Perfeito et al. 2007; Gordo et al. 2011).

When testing the IAD model, ~20 single amino acid substitution mutations were found to be beneficial in this specific experimental context (i.e., need for the *hisA* gene to provide a dual HisA/TrpF function). Assuming that we had reached saturation (many mutations were independently found more than once) and found the majority of beneficial mutations, and using experimentally determined rates of mutation per nucleotide per generation in *S. typhimurium* (Hudson et al. 2002), we can estimate that the beneficial mutation rate is on the order of 10^{-9} to 10^{-8} per generation in the *hisA* gene.

Predictions from and Supporting Evidence for the IAD Model

From the IAD model, one can make several predictions. First, new paralogs should often be found clustered together in tandem arrays, with some copies that retain the original function and other copies that have the new function. Second, the paralogs should show evidence for positive selection in the copy that has evolved the new function. Third, in cases in which the ancestral protein can be found (e.g., as an ortholog in a closely related species) or deduced from bioinformatics, it should show some trace of the new function. The first two predictions may be general for most models of evolution through duplication–divergence, whereas the third is specific to the IAD model. A compelling example of a system that may have evolved through IAD is the evolution of export-

ed antifreeze proteins (AFPs) from a carboxy-terminal portion of the cytoplasmic enzyme sialic acid synthase (SAS) in the fish Antarctic eelpout (*Lycodichthys dearborni*) (Deng et al. 2010). Closely related fish that lack AFPs have two copies of the gene encoding SAS in a locus in the genome (Fig. 4). In *L. dearborni*, a retrotransposon is inserted between the two SAS genes (SAS-A and SAS-B), and in another locus in the genome, a fragment of this retrotransposon plus an array of >30 tandem copies of the evolved AFP genes (evolved from part of the SAS-B gene) is inserted. Further, the purified SAS-B protein has a weak ice-binding activity. From these facts, one can suggest a model for the evolution of AFPs in *L. dearborni*. A change in the environment (cooling of the sea) made the weak ice-binding activity of SAS-B beneficial. Selection for increased expression of SAS-B led to duplication of the SAS-B gene through a retrotransposition event. During the evolution to the AFPs that can be found today, the extra copy of the SAS-B gene has gained an export signal through mutations that extends the amino terminal by a few amino acids, and lost all exons except the one encoding the very carboxy-terminal domain. In addition to this, the gene has been amplified to high copy number and gained several point mutations that contribute to improved and efficient ice-binding activity. This example has been used as support for the “escape from adaptive conflict” (EAC) model of new gene evolution, which bears several similarities to the IAD model, except it does not consider the possibility of positive selection for the initial duplication of the ancestral gene.

Another interesting example is the evolution of the MALS family of α -glucosidases in *S. cerevisiae* and its close relatives (Voordeckers et al. 2012). *S. cerevisiae* has seven closely related α -glucosidases (*MAL12* and 32, *IMA1–5*) with different substrate specificities. *MAL12* and 32 hydrolyze maltose, sucrose, and similar sugars, whereas the *IMA* proteins hydrolyze isomaltose and some similar sugars. Reconstruction of the hypothetical single ancestor of all seven MALS genes from *S. cerevisiae* resulted in an enzyme that has fair activity toward sucrose and maltose, but very poor activity toward isomaltose.

Figure 4. Antifreeze proteins (AFPs) evolution in Antarctic eelpout. (A) AFPs in Antarctic eelpout have evolved from the carboxy terminal of sialic acid synthase (SAS). Closely related fish without AFPs have two SAS genes next to each other in the genome. In Antarctic eelpout, a transposon (LdCR1-3) has been inserted in between the SAS genes. (B) In another chromosomal locus in Antarctic eelpout, there is a truncated copy of the transposon (LdCR1-3) along with >30 tandem copies of the newly evolved AFP. This transposition event is not found in closely related fish not adapted to cold conditions. The AFPs derive from the carboxyl terminus of the SAS-B gene that shows a weak intrinsic ice-binding activity. The amplified copies have accumulated point mutations that contribute to ice binding and a secretion signal in the amino terminal to direct extracellular secretion. *L. dearborni*, *Lycodichthys dearborni*; *G. aculeatus*; *Gasterosteus aculeatus*.

Although ancestors along the branches toward *MAL12* and *MAL32* improved the activities toward maltose and sucrose, ancestors along the branches toward the *IMA* proteins gained activity on isomaltose and similar sugars at the expense of the activities on maltose-like sugars. Only two pairs of the genes in this family are close to each other on the same chromosomes, but they are all located in the subtelomeric regions of several chromosomes. During the adaptation of *Saccharomyces* to a fermentative lifestyle, the ancestor of the *MALS* genes has been duplicated several times by different mechanisms. This could have been positively selected to increase the promiscuous activities toward several substrates, increasing the fitness in the presence of those substrates. Most or all of the traces of any tandem amplifications, if they occurred at any point, could since have been erased by segregation after the appearance of

beneficial mutations. In parallel to the expansion of the *MALS* genes, the *MALR* (regulators), and *MALT* (permeases) gene families also expanded in subtelomeric regions of different chromosomes (Brown et al. 2010).

An Experimental Test of the IAD Model

We recently tested the IAD model in a laboratory experiment (Fig. 5) (Näsvall et al. 2012). To simulate the “Innovation” part of the model, we generated *HisA^{dual}*, a mutant variant of the *Salmonella* *HisA* enzyme that catalyzes a reaction in tryptophan biosynthesis (normally catalyzed by TrpF) in addition to its original function in histidine biosynthesis. A *Salmonella* strain lacking the normal *hisA* and *trpF* genes, but carrying a *hisA^{dual}* gene grows very poorly on medium lacking histidine or tryptophan, because of poor *HisA* and *TrpF* activities. We grew several

D.I. Andersson et al.

Figure 5. An experimental test of the innovation–amplification–divergence (IAD) model. *Salmonella enterica* carrying a bifunctional gene, *hisA^{dual}*, with two weak activities in histidine biosynthesis (original activity) and tryptophan biosynthesis (new activity) were placed under selection for 3000 generations to improve both activities. After 1000 and 2000 generations, some chosen lineages were split up into new lineages as indicated in the white text boxes. (A) Trajectories of evolution. Green symbols indicate gene variants that are sufficient for growth in the absence of both histidine and tryptophan (generalists). Blue symbols indicate variants that are sufficient to support growth in the absence of histidine but not tryptophan (HisA specialists). Yellow symbols indicate variants that are sufficient for supporting growth in the absence of tryptophan but not histidine (TrpF specialists). Numbers to the left indicate after how many generations the indicated variants were observed. The black and white bars on the gene symbols indicate mutations in the evolved variants compared with the ancestral *hisA^{dual}* gene. (B) Trajectories of evolution of enzyme activities. The circled letters indicate examples of evolved gene variants (highlighted with the same letters in A). Each gene variant was placed as a single copy on the *Salmonella* chromosome. The growth rate in the presence of histidine but absence of tryptophan was used as a measure of TrpF activity on the y-axis. The growth rate in the presence of tryptophan but absence of histidine was used as a measure of HisA activity on the x-axis. gen., Generations.

populations of this bacterium in minimal medium lacking both amino acids for ~3000 generations to select improved histidine and tryptophan synthesis, and thus faster growth. As predicted from the IAD model, duplication and amplification of the *hisA^{dual}* gene were the predominant mechanism of early adaptation, and was apparent very early in the experiment. After as few as 50 or less generations, cells having two or more copies of the *hisA^{dual}* gene dominated most populations. Appearance of beneficial mutations other than copy number changes were slower, but eventually they did turn up. In some of the populations, cells that had two different *hisA* genes evolved, in which one gene had an improved TrpF activity, whereas the other had an improved HisA activity. After 3000 generations of continuous selection, most cultures contained cells with two specialized HisA enzymes that had diverged from their ancestor by up to three amino acid changes. We could, thus, after constructing the “innovation,” follow evolution of the new genes through the “amplification” and early stages of “divergence” in just 3000 generations of bacterial growth, or just a bit over 300 days.

DE NOVO EVOLUTION OF NEW GENES AND FUNCTIONS

The evolution of an entirely new gene without any functional progenitor has been regarded as an exceedingly rare and improbable event. Considering the vast number of possible protein species and the rather limited number of configurations found in biological systems, this notion might seem to make sense. Any random sequence generated by chance must end up as a stable polypeptide with a function accessible by selection or end up being purged from the genome by neutral mutational processes. François Jacob famously stated that nature is but a “tinkerer” of already available protein hardware, advocating the importance of duplications in gene emergence (Jacob 1977). With the advent of large-scale genome sequencing, it soon became apparent that a large part of extant genomes contain genes without any recognized homologous sequences that nature had “tin-

kered” genes from. The origin and status of these genes, termed orphan or ORFan genes or “taxonomically restricted genes” (TRGs), have posed an evolutionary conundrum because of their high abundance and almost ubiquitous presence in genomes.

Relation to Orphan Genes

Orphan genes are suspected coding sequences that lack identifiable homologs in all sequenced genomes. Substantial parts of the gene catalog from bacteria, archaea, eukaryotes (10%–30%), and particularly viruses (on average 30%) have already been recognized as orphan genes since the dawn of genome sequencing (Yin and Fischer 2008; Wissler et al. 2013). The definition of orphan genes is complex and depends on both the detection method used and the reference set used. The detection method should be sensitive enough to differentiate between specific and random matches and at the same time not conservative enough to bias the discovery of novel proteins. The determination of a gene lineage can be accomplished by phylostratigraphy, a framework in which the protein coding complement of increasingly divergent species acts as temporal layers that can be used to determine the evolutionary age of all genes in the organisms under study. This dating allows the assignment of orphan genes as genes only populating the terminal layer of the strata. Such stratigraphy has revealed the presence of many tens to hundreds of potential orphan genes in diverse lineages (Ladoukakis et al. 2011; Yu and Stoltzfus 2012; Neme and Tautz 2013; Wissler et al. 2013).

The source of orphans is not immediately clear but several mechanisms have been proposed to explain their prevalence. One possibility is that orphans might be spurious ORFs that are artefacts of genome annotation. However, many orphan genes display slower evolutionary rates than noncoding regions arguing for selective processes acting for their retention. At the same time, many orphan genes are evolving faster than core genes. Orphan genes might also be genes that have diverged beyond the point where we are able to infer their evolutionary origin. Structural analysis of orphan proteins might

D.I. Andersson et al.

be used to ascertain whether they adopt already known folds, either indicating a common origin or convergent evolution, or whether they represent unique polypeptides. Orphan genes could also be pseudogenes, frame-shifted, or rearranged genes. They might even be the vestiges of gene loss since a common ancestor; however, this would impose impossibly large genomes in ancestral organisms. They might also be horizontally transferred genes in which the donor lineage is yet to be sampled by genomics. Finally, they might have originated *de novo* in the organism under study from noncoding sequences. Orphan genes have fewer paralogous gene family members, which might indicate that they have a shorter history in a genome than other genes (Fischer and Eisenberg 1999). Studies in *E. coli* show that characteristics of orphan genes are that they are shorter and more A/T rich than genes of a broader phylogenetic distribution (Yu and Stoltzfus 2012). The presence of many orphans in clusters close to repetitive elements and phage integration sites suggests that mobile genetic elements are responsible for the influx of orphan genes, at least in prokaryotes (Yu and Stoltzfus 2012). The relation between orphans and *de novo* genes is, at the moment, not entirely clear. It is quite possible that many orphan genes will be found to have homologs as genome databases increase the taxonomic sampling density.

How Do Genes Emerge *De Novo*?

A few basic tenets need to be met to give rise to a gene *de novo*. A stretch of DNA encoding an ORF needs to acquire signals that lead to the transcription and subsequent translation of encoded polypeptide. That polypeptide, in turn, needs to be biologically active for enough time to allow selective forces to operate for its retention in the genome. Studies of *de novo* genes in *Drosophila* indicate that different routes are possible in the formation of *de novo* genes. Some regions that give rise to *de novo* genes were transcribed as novel noncoding RNAs before the appearance of an ORF suitable for translation. In other cases, genes showed evidence of transcription and an ORF being formed in the same time interval. In contrast, there were no cases of

proto-ORFs existing long before formation of the novel gene (Reinhardt et al. 2013).

Large-scale expression studies have, in recent years, catalogued dense transcription of many characterized genomes (Dinger et al. 2008). Many of the transcribed loci are not predicted to give rise to proteins but rather encode noncoding RNA as they mainly contain short ORFs. Surprisingly, ribosome-profiling data have indicated that many long noncoding RNA with short ORFs are in fact exported to the cytoplasm and engaged by the ribosome (Wilson and Masel 2011; Carvunis et al. 2012). The presence of such translation events on noncoding RNA has been used to formulate an evolutionary theory in which transcribed noncoding regions act as protogenes whose transcription and translation might become under selection to be kept as protein (Fig. 6). With increasing age of the protogene, it will transition into being a “*bona fide*” gene by the action of selective forces. These changes might be incorporation into regulatory systems for stage-specific expression or interaction with other gene products. The protogene model of gene birth serves as an elaboration of previous models of gene evolution and gene turnover through duplication–divergence and pseudogenization. This model allows there to be a continuum of evolutionary intermediates from noncoding ORFs to genes via protogenes as opposed to the strict cutoffs imposed by gene-calling algorithms and filtering procedures.

Bifunctional RNAs that are both translated and function as noncoding RNAs might be viewed as special examples of how intermediates between noncoding RNAs and *de novo* evolved genes can coexist simultaneously (Dinger et al. 2008). Expression of protogenes offers a route to select against variants that are cytotoxic caused by aggregation and low solubility. (The IAD model would work well to explain the emergence of *de novo* genes if selective functions are common enough.) *De novo* genes can also form through overprinting using alternative reading frames of already functional genes. The *de novo* proteins formed in this manner would be in a transcriptional permissive environment with signals in place for cor-

Evolution of New Genes

Figure 6. De novo gene evolution. Genes can evolve de novo through several mechanisms. Transcription of protogenes (noncoding RNAs with open reading frames (ORFs), overlapping gene ORFs, intergenic regions) lead to ribosomal association and translation of the message. Translated peptides might confer a selective advantage through potential weak promiscuous activities. The mechanism described in the innovation–amplification–divergence (IAD) model would operate to increase the effectiveness of the protogenes through positive selection and promote the birth of novel genes. Some protogenes never make the transition to become selectively advantageous in the long run and become pseudogenized.

rect expression and translation. This process has been documented in prokaryotes (Delaye et al. 2008) and in eukaryotes (Neme and Tautz 2013) but appears to be uncommon. In contrast, this mechanism appears to be widely used in many groups of viruses (Sabath et al. 2012; Pavesi et al. 2013).

Evidence for De Novo Emergence of Genes

De novo genes have mostly been attested in eukaryotes in which the availability of high-quality genome sequences with well-supported gene models from closely related species have allowed the delineation of the evolutionary emergence of all genes in a given genome. The gold standard for calling a de novo evolved gene is showing that a gene is a noncoding region in multiple out-group species and that the in-group species transcribes an RNA in the equivalent genomic region, which is translated into a protein species that is validated by proteomic techniques.

In practice, evidence fulfilling the stringent criteria outlined above might be difficult to obtain. Typically, de novo gene candidates are identified by bioinformatic filtering of predicted gene catalogs, transcript expression libraries and shotgun proteomic data. Evolutionary divergence characteristics are commonly used as a filter to distinguish de novo gene candidates from neutrally evolving genomic regions. De novo gene emergence have been reported from many organisms such as insects (Begun et al. 2007; Reinhardt et al. 2013), yeast (Cai et al. 2008; Li et al. 2010b), *Hydra* (Khalturin et al. 2008), primates (Johnson et al. 2001; Knowles and McLysaght 2009; Toll-Riera et al. 2009; Li et al. 2010a; Wu et al. 2011; Xie et al. 2012), mouse (Murphy and McLysaght 2012; Neme and Tautz 2013), *Plasmodium* (Yang and Huang 2011), and plants (Donoghue et al. 2011).

De novo genes are often characterized by being short, often overlapping other genes or being present within intronic sequences. Many de novo genes show weak expression. However, in animals, the highest expression is often found

D.I. Andersson et al.

in the testes. This finding might be correlated to the hyperactive transcription in this sexual organ, a feature that leads to overall higher transcription. The increased transcription might lead to an increased selective potential as outlined in the “out of the testes” hypothesis (Kaessmann 2010). De novo genes in plants are often tissue-specifically expressed with stress-induced responsiveness (Donoghue et al. 2011). Even with transcriptional evidence and evolutionary conservation, some de novo genes might be examples of unrecognized noncoding RNAs that experience weaker structural constraints than most coding sequences. Because of the obvious risk of misidentification, it is imperative that proteomic evidence is gathered to support the claim of identified de novo genes. Proteomic evidence has indeed been integral in many de novo gene prediction pipelines, although the amount of proteomic evidence available is often limited.

Several clearly defined cases of de novo gene emergence have been attested by extensive functional characterization but most identified genes remain uncharacterized. A de novo originated gene has been described in yeast that acts to depress the mating pathway. Curiously, this gene in turn represses the protein encoded on its antisense strand (Li et al. 2010b). A second example comes from *Hydra* in which the expression of a small, secreted orphan protein recapitulates phenotypic differences seen between species (Khalturin et al. 2008). Recently, six de novo evolved *Drosophila* genes were investigated for their contribution to organismal fitness. The genes show higher expression in the males than in females, notably in the testes, but knockdown of the genes lead to inhibition of metamorphosis in both sexes (Reinhardt et al. 2013).

Protein Folds; Limited Number of Solution; Unstructured Proteins

Tightly linked to the concept of de novo gene formation is the protein-folding problem. What is the possibility of arriving at a biologically active polypeptide through random exploration of sequence space. The evolution of novel proteins capable of molecular recognition and cat-

alytic activity has been explored by selection for activity or small molecule binders in several experiments. Keefe and Szostak derived ATP-binding proteins from 80 amino acid random sequences using high-diversity mRNA display (6×10^{12} variants) without relying on stabilizing sequence motifs (Keefe and Szostak 2001). In their selection, an estimated frequency of 1 in 10^{11} random sequence proteins was able to bind ATP. Interestingly, the selected proteins showed no significant homology with biological ATP-binding motifs suggesting that there are novel folds not sampled by nature for this specific purpose. Streptavidin-binding proteins selected by mRNA-display showed nanomolar affinities and a consensus HPQ motif that mimics biotin (Wilson et al. 2001). Twenty different binders were identified in a library of 10^{13} peptides of 88 amino acids in length, showing a similar frequency of appearance as the ATP-binding ability. Totally random peptide libraries have been shown to produce few soluble proteins, and many researchers have opted to maximize the likelihood of obtaining soluble proteins by providing a stable scaffold with randomized positions elsewhere in the sequence (4-helix bundles) (Patel et al. 2009) or limiting the amino acid repertoire in the libraries (Tanaka et al. 2010). In the case of the 4-helix bundles, positionally randomized but otherwise unevolved proteins were generated and a high proportion showed heme binding as a cofactor. This imparted enzymatic activity to the binders, some of which displayed several promiscuous activities even in the absence of the cofactor (Patel et al. 2009). The same group also showed that a library of 10^6 positionally randomized 4-helix bundle proteins carries enough activities to rescue four out of 27 conditionally essential knockout mutants in *E. coli* (Fisher et al. 2011). Three naïve 4-helix bundles were screened for binding to 10,000 small molecules and indicated that each protein displayed a distinct binding profile (Cherny et al. 2012).

Even though a stable 3D structure is often needed for proper function, some natural polypeptides are inherently unstructured (Itzhaki and Wolynes 2008; Pavlović-Lažetić et al. 2011). These proteins adopt a folded shape only on

contact with binding partners or cofactors. The evolution of a stably folded protein could proceed by an unstructured stage in which folding is stabilized by chaperone activity of other polypeptides or cofactors in the cellular environment.

Divergent or Convergent Evolution of Folds; Limited Number of Solutions

There appears to be limited number of folded polypeptides that nature uses to create the available repertoire of proteins. To understand how protein evolution works, it is important to understand the fold diversity of proteins. Although no systematic large-scale orphan structural study has been reported, there has been a clear decline in the discovery of novel protein folds. A recent study of 248 domains of unknown function (DUFs) by the protein structure initiative found that 27% of the domains represented a novel fold and the remainder belonged to already known folds or variations thereof (Jaroszewski et al. 2009). Estimates of fold diversities indicate that probably between 1000 and 10,000 protein folds exist (Orengo and Thornton 2005), which is on the same order as the number of recognized folds today (1208 folds, SCOP, released February 2015). Structural studies of proteins have noticed a decreased pace of novel folds being discovered. Many folds show regions of local similarity, a feature that might be the result of convergent evolution or a reflection of ancient evolutionary processes dating to the emergence of the modern protein with defined domains (Ponting and Russell 2002).

The presence of proteins with identical folds but no discernible sequence identity has been noted by structural biologists throughout the years (Ponting and Russell 2002). One example is adenylate cyclase and DNA polymerase that share the central palm catalytic domain despite showing no significant homology (Artymiuk et al. 1997). Both of these enzymes perform a similar enzymatic reaction, which involves the elimination of pyrophosphate using the 3' OH of a ribose unit to attack the α -phosphate of a nucleotide 5' triphosphate. Whether these cases

are examples of extreme divergence from a common ancestor or convergent evolution is difficult to address from genome sequences alone. Convergent evolution has been noted in the case of the Ser/His/Asp catalytic triad, which is found in at least five distinct folds (Dodson and Wlodawer 1998). How would divergence and convergence of protein folds be tested in the laboratory? One way to approach this question is to generate synthetic proteins de novo by selecting for rescue of cellular auxotrophies. De novo proteins that can be recovered and shown to possess the same function as the missing protein causing the auxotrophy should be structurally determined. Initial screens might have to be undertaken by mRNA display or augmented by computer-based prescreens to gain libraries deep enough for recovery of multiple hits. If this can be accomplished at a large enough scale, it might be possible to gain information on how many possible solutions exist for a given activity and whether highly similar folds develop independently catalyzing the same reaction. If one fold is recovered repeatedly, it might indicate that highly similar folds are far more likely to have arisen independently without a common ancestor. It should also be possible ascertain whether these folds are reached using an intermediate unfolded stage that is gradually stabilized by evolutionary adaptation. This would allow the distinction of what is principally selectable in early protein evolution: structural stability, activity, or a balancing of both. Of course, this matter is complicated by the fact that only subsets of sequence space might be accessible even for short proteins, and the results might be radically different depending on the region of sequence space being queried.

ACKNOWLEDGMENTS

Our work is supported by the Swedish Research Council and the Wallenberg Foundation.

REFERENCES

Adler M, Anjum M, Andersson DI, Sandegren L. 2013. Influence of acquired β -lactamases on the evolution of

D.I. Andersson et al.

- spontaneous carbapenem resistance in *Escherichia coli*. *J Antimicrob Chemother* **68**: 51–9.
- Adler M, Anjum M, Berg OG, Andersson DI, Sandegren L. 2014. High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Mol Biol Evol* **31**: 1526–1535.
- Anderson R, Roth J. 1977. Tandem genetic duplications in phage and bacteria. *Annu Rev Microbiol* **31**: 473–505.
- Anderson R, Roth J. 1979. Gene duplication in bacteria: Alteration of gene dosage by sister-chromosome exchanges. *Cold Spring Harb Symp* **43**: 1083–1087.
- Anderson P, Roth J. 1981. Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (rrn) cistrons. *Proc Natl Acad Sci* **78**: 3113–3117.
- Andersson DI, Hughes D. 2009. Gene amplification and adaptive evolution in bacteria. *Annu Rev Genet* **43**: 167–195.
- Artymiuk PJ, Poirrette AR, Rice DW, Willett P. 1997. A polymerase I palm in adenyl cyclase? *Nature* **388**: 33–34.
- Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. *Genetics* **176**: 1131–1137.
- Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: Evolution of new genes under continuous selection. *Proc Natl Acad Sci* **104**: 17004–17009.
- Brown CA, Murray AW, Verstrepen KJ. 2010. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr Biol* **20**: 895–903.
- Byrne KP, Wolfe KH. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* **175**: 1341–1350.
- Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179**: 487–496.
- Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charlotiaux B, Hidalgo CA, Barrette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* **487**: 370–374.
- Cherny I, Korolev M, Koehler AN, Hecht MH. 2012. Proteins from an unevaluated library of de novo designed sequences bind a range of small molecules. *ACS Synth Biol* **1**: 130–138.
- Choi I-G, Kim S-H. 2006. Evolution of protein structural classes and protein sequence families. *Proc Natl Acad Sci* **103**: 14056–14061.
- Copley SD. 2003. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Curr Opin Chem Biol* **7**: 265–272.
- Copley SD. 2012. Moonlighting is mainstream: Paradigm adjustment required. *Bioessays* **34**: 578–588.
- Delaye L, Deluna A, Lazcano A, Becerra A. 2008. The origin of a novel gene through overprinting in *Escherichia coli*. *BMC Evol Biol* **8**: 31.
- Deng C, Cheng CH, Ye H, He X, Chen L. 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc Natl Acad Sci* **107**: 21593–21598.
- Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Comput Biol* **4**: e1000176.
- Dittmar K, Liberles DA. 2010. *Evolution after gene duplication*. Wiley, Hoboken, NJ.
- Dodson G, Wlodawer A. 1998. Catalytic triads and their relatives. *Trends Biochem Sci* **23**: 347–352.
- Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol Biol* **11**: 47.
- Fischer D, Eisenberg D. 1999. Finding families for genomic ORFans. *Bioinformatics* **15**: 759–762.
- Fisher R. 1935. The sheltering of lethals. *Am Nat* **69**: 446–455.
- Fisher MA, McKinley KL, Bradley LH, Viola SR, Hecht MH. 2011. De novo designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth. *PLoS ONE* **6**: e15364.
- Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.
- Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. *Nat Genet* **37**: 573–577.
- Gelbart W, Chovnick A. 1979. Spontaneous unequal exchange in the rosy region of *Drosophila melanogaster*. *Genetics* **92**: 849–859.
- Gordo I, Perfeito L, Sousa A. 2011. Fitness effects of mutations in bacteria. *J Mol Microbiol Biotechnol* **21**: 20–35.
- Haldane JBS. 1932. *The causes of evolution*. Longmans, London.
- Hartley B. 1984. Experimental evolution of ribitol dehydrogenase. In *Microorganisms as model systems for studying evolution* (ed. Mortlock RP), pp. 23–54. Plenum, New York.
- Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**: 551–564.
- He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**: 1157–1164.
- Hegreness M, Shores N, Hartl D, Kishony R. 2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311**: 1615–1617.
- Henderson B, Martin A. 2011. Bacterial virulence in the moonlight: Multitasking bacterial moonlighting proteins are virulence determinants in infectious disease. *Infect Immun* **79**: 3476–3491.
- Hendrickson H, Slechts ES, Bergthorsson U, Andersson DI, Roth JR. 2002. Amplification-mutagenesis: Evidence that “directed” adaptive mutation and general hypermutability result from growth with a selected gene amplification. *Proc Natl Acad Sci* **99**: 2164–2169.
- Hudson RE, Bergthorsson U, Roth JR, Ochman H. 2002. Effect of chromosome location on bacterial mutation rates. *Mol Biol Evol* **19**: 85–92.
- Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc R Soc London B* **256**: 119–124.

- Imhof M, Schlotterer C. 2001. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. *Proc Natl Acad Sci* **98**: 1113–1117.
- Itzhaki L, Wolynes P. 2008. The quest to understand protein folding. *Curr Opin Struct Biol* **18**: 1–3.
- Jacob F. 1977. Evolution and tinkering. *Science* **196**: 1161–1166.
- Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A. 2009. Exploration of uncharted regions of the protein universe. *PLoS Biol* **7**: e1000205.
- Jensen R. 1976. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* **30**: 409–425.
- Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**: 514–519.
- Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–1326.
- Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: Rates, fitness effects and adaptive significance. *Front Genet* **4**: 273.
- Kazazian HH. 2004. Mobile elements: Drivers of genome evolution. *Science* **303**: 1626–1632.
- Keefe AD, Szostak JW. 2001. Functional proteins from a random-sequence library. *Nature* **410**: 715–718.
- Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.
- Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TC. 2008. A novel gene family controls species-specific morphological traits in *Hydra*. *PLoS Biol* **6**: e278.
- Khersonsky O, Tawfik DS. 2010. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu Rev Biochem* **79**: 471–505.
- Khersonsky O, Roodveldt C, Tawfik DS. 2006. Enzyme promiscuity: Evolutionary and mechanistic aspects. *Curr Opin Chem Biol* **10**: 498–508.
- Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD. 2010. Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5'-phosphate synthesis. *Mol Syst Biol* **6**: 436.
- Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res* **19**: 1752–1759.
- Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome Biol* **3**: 8.
- Kugelberg E, Kofoed E, Andersson DI, Lu Y, Mellor J, Roth FP, Roth JR. 2010. The tandem inversion duplication in *Salmonella enterica*: Selection drives unstable precursors to final mutation types. *Genetics* **185**: 65–80.
- Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol* **12**: R118.
- Lam K-WG, Jeffreys AJ. 2007. Processes of de novo duplication of human α -globin genes. *Proc Natl Acad Sci* **104**: 10950–10955.
- Larion M, Moore LB, Thompson SM, Miller BG. 2007. Divergent evolution of function in the ROK sugar kinase superfamily: Role of enzyme loops in substrate specificity. *Biochemistry* **46**: 13564–13572.
- Li CY, Zhang Y, Wang Z, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X, Du Q, et al. 2010a. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput Biol* **6**: e1000734.
- Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. 2010b. 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.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.
- Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**: 459–473.
- Miller BG, Raines RT. 2004. Identifying latent enzyme activities: Substrate ambiguity within modern bacterial sugar kinases. *Biochemistry* **43**: 6387–6392.
- Miller BG, Raines RT. 2005. Reconstitution of a defunct glycolytic pathway via recruitment of ambiguous sugar kinases. *Biochemistry* **44**: 10776–10783.
- Murphy DN, McLysaght A. 2012. De novo origin of protein-coding genes in murine rodents. *PLoS ONE* **7**: e48650.
- Nakahigashi K, Toya Y, Ishii N, Soga T, Hasegawa M, Watanabe H, Takai Y, Honma M, Mori H, Tomita M. 2009. Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol Syst Biol* **5**: 306.
- Nam H, Lewis NE, Lerman JA, Lee DH, Chang RL, Kim D, Palsson BO. 2012. Network context and selection in the evolution to enzyme specificity. *Science* **337**: 1101–1104.
- Näsval J, Sun L, Roth JR, Andersson DI. 2012. Real-time evolution of new genes by innovation, amplification, and divergence. *Science* **338**: 384–387.
- Neidhardt FC, Ingraham JL, Schaechter M. 1990. *Physiology of the bacterial cell: A molecular approach*. Sinauer Associates, Sunderland, MA.
- Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**: 117.
- Ohno S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin.
- Orengo CA, Thornton JM. 2005. Protein families and their evolution—A structural perspective. *Annu Rev Biochem* **74**: 867–900.
- Patel SC, Bradley LH, Jinadasa SP, Hecht MH. 2009. Cofactor binding and enzymatic activity in an unevolved superfamily of de novo designed 4-helix bundle proteins. *Protein Sci* **18**: 1388–1400.
- Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. 2007. Multicopy suppression underpins metabolic evolvability. *Mol Biol Evol* **24**: 2716–2722.
- Pavesi A, Magiorkinis G, Karlin DG. 2013. Viral proteins originated de novo by overprinting can be identified by codon usage: Application to the “gene nursery” of Deltaretroviruses. *PLoS Comput Biol* **9**: e1003162.
- Pavlović-Lazetić GM, Mitić NS, Kovačević JJ, Obradović Z, Malkov SN, Beljanski MV. 2011. Bioinformatics analysis

D.I. Andersson et al.

- of disordered proteins in prokaryotes. *BMC Bioinformatics* **12**: 66.
- Perfeito L, Fernandes L, Mota C, Gordo I. 2007. Adaptive mutations in bacteria: High rate and small effects. *Science* **317**: 813–815.
- Pettersson ME, Sun S, Andersson DI, Berg OG. 2009. Evolution of new gene functions: Simulation and analysis of the amplification model. *Genetica* **135**: 309–324.
- Piatigorsky J. 1991. The recruitment of crystallins: New functions precede gene duplication. *Science* **252**: 1078–1079.
- Ponting CP, Russell RR. 2002. The natural history of protein domains. *Annu Rev Biophys Biomol Struct* **31**: 45–71.
- Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* **5**: 28.
- Reams AB, Kofoid E, Savageau M, Roth JR. 2010. Duplication frequency in a population of *Salmonella enterica* rapidly approaches steady state with or without recombination. *Genetics* **184**: 1077–1094.
- Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet* **9**: e1003860.
- Romero D, Palacios R. 1997. Gene amplification and genomic plasticity in prokaryotes. *Annu Rev Genet* **31**: 91–111.
- Sabath N, Wagner A, Karlin D. 2012. Evolution of viral proteins originated de novo by overprinting. *Mol Biol Evol* **29**: 3767–3780.
- Sandegren L, Andersson DI. 2009. Bacterial gene amplification: Implications for the evolution of antibiotic resistance. *Nat Rev Microbiol* **7**: 578–588.
- Scannell DR, Wolfe KH. 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res* **18**: 137–147.
- Shapira SK, Finnerty VG. 1986. The use of genetic complementation in the study of eukaryotic macromolecular evolution: Rate of spontaneous gene duplication at two loci of *Drosophila melanogaster*. *J Mol Evol* **23**: 159–167.
- Sun S, Berg OG, Roth JR, Andersson DI. 2009. Contribution of gene amplification to evolution of increased antibiotic resistance in *Salmonella typhimurium*. *Genetics* **182**: 1183–1195.
- Sun S, Ke R, Hughes D, Nilsson M, Andersson DI. 2012. Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PLoS ONE* **7**: e42639.
- Tanaka J, Doi N, Takashima H, Yanagawa H. 2010. Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci* **19**: 786–795.
- Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12**: 692–702.
- Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Alba MM. 2009. Origin of primate orphan genes: A comparative genomics approach. *Mol Biol Evol* **26**: 603–612.
- Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, Verstrepen KJ. 2012. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol* **10**: e1001446.
- Wilson Ba, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol* **3**: 1245–1252.
- Wilson DS, Keefe AD, Szostak JW. 2001. The use of mRNA display to select high-affinity protein-binding peptides. *Proc Natl Acad Sci* **98**: 3750–3755.
- Wissler L, Gadau J, Simola DE, Helmkamp M, Bornberg-Bauer E. 2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biol Evol* **5**: 439–455.
- Wu D, Zhang Y. 2013. Evolution and function of de novo originated genes. *Mol Phylogenet Evol* **67**: 541–545.
- Wu D-D, Irwin DM, Zhang Y-PP. 2011. De novo origin of human protein-coding genes. *PLoS Genet* **7**: e1002379.
- Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, Zhang M, Zhang R, Wei L, Li CY. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet* **8**: e1002942.
- Yang Z, Huang J. 2011. De novo origin of new genes with introns in *Plasmodium vivax*. *FEBS Lett* **585**: 641–644.
- Yin Y, Fischer D. 2008. Identification and investigation of ORFans in the viral world. *BMC Genomics* **9**: 24.
- Yip SH-C, Matsumura I. 2013. Substrate ambiguous enzymes within the *Escherichia coli* proteome offer different evolutionary solutions to the same problem. *Mol Biol Evol* **30**: 2001–2012.
- Yu G, Stoltzfus A. 2012. Population diversity of ORFan genes in *Escherichia coli*. *Genome Biol Evol* **4**: 1176–1187.

Cold Spring Harbor Perspectives in Biology

Evolution of New Functions De Novo and from Preexisting Genes

Dan I. Andersson, Jon Jerlström-Hultqvist and Joakim Näsvall

Cold Spring Harb Perspect Biol 2015; doi: 10.1101/cshperspect.a017996

Subject Collection [Microbial Evolution](#)

Not So Simple After All: Bacteria, Their Population Genetics, and Recombination

William P. Hanage

Realizing Microbial Evolution

Howard Ochman

Thoughts Toward a Theory of Natural Selection: The Importance of Microbial Experimental Evolution

Daniel Dykhuizen

Coevolution of the Organization and Structure of Prokaryotic Genomes

Marie Touchon and Eduardo P.C. Rocha

Mutation—The Engine of Evolution: Studying Mutation and Its Role in the Evolution of Bacteria

Ruth Hershberg

The Origin of Mutants Under Selection: How Natural Selection Mimics Mutagenesis (Adaptive Mutation)

Sophie Maisnier-Patin and John R. Roth

Evolution of New Functions De Novo and from Preexisting Genes

Dan I. Andersson, Jon Jerlström-Hultqvist and Joakim Näsvall

Genome-Based Microbial Taxonomy Coming of Age

Philip Hugenholtz, Adam Skarshewski and Donovan H. Parks

Horizontal Gene Transfer and the History of Life

Vincent Daubin and Gergely J. Szöllosi

Early Microbial Evolution: The Age of Anaerobes

William F. Martin and Filipa L. Sousa

Microbial Speciation

B. Jesse Shapiro and Martin F. Polz

The Evolution of *Campylobacter jejuni* and *Campylobacter coli*

Samuel K. Sheppard and Martin C.J. Maiden

Paleobiological Perspectives on Early Microbial Evolution

Andrew H. Knoll

For additional articles in this collection, see <http://cshperspectives.cshlp.org/cgi/collection/>

**All Modifications and
Oligo Types Synthesized**
Long Oligos • Fluorescent • Chimeric • DNA • RNA • Antisense

Oligo Modifications?
Your wish is our command.