Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Orphans and new gene origination, a structural and evolutionary perspective

Sara Light[1,2,3], Walter Basile[1,2] and Arne Elofsson[1,2,4]

The frequency of *de novo* creation of proteins has been debated. Early it was assumed that *de novo* creation should be extremely rare and that the vast majority of all protein coding genes were created in early history of life. However, the early genomics era lead to the insight that protein coding genes do appear to be lineage-specific. Today, with thousands of completely sequenced genomes, this impression remains. It has even been proposed that the creation of novel genes, a continuous process where most *de novo* genes are short-lived, is as frequent as gene duplications. There exist reports with strongly indicative evidence for *de novo* gene emergence in many organisms ranging from Bacteria, sometimes generated through bacteriophages, to humans, where orphans appear to be overexpressed in brain and testis. In contrast, research on protein evolution indicates that many very distantly related proteins appear to share partial homology. Here, we discuss recent results on *de novo* gene emergence, as well as important technical challenges limiting our ability to get a definite answer to the extent of *de novo* protein creation.

Addresses
[1] Science for Life Laboratory, Stockholm University, SE-171 21 Solna, Sweden
[2] Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden
[3] Bioinformatics Infrastructure for Life Sciences, Sweden
[4] Swedish e-Science Research Center (SeRC), Sweden

Corresponding author: Elofsson, Arne (arne@bioinfo.se)

Current Opinion in Structural Biology 2014, **26**:73–83

This review comes from a themed issue on **Sequences and topology**

Edited by **L Aravind** and **Christine A Orengo**

For a complete overview see the Issue and the Editorial

Available online 13th June 2014

http://dx.doi.org/10.1016/j.sbi.2014.05.006

Introduction

Protein coding genes evolve in a number of different ways. Mutations cause changes in the amino acid sequence as well as introduce insertions and deletions (indels). Speciation and population variation cause all genes to exist in multiple copies that diverge by time. Further, gene and whole genome duplications increase the number of members within a gene family and allows for subfunctionalization [1].

In addition to these mechanisms that cause variation within a protein family, there are other mechanisms creating novel protein coding genes. The most common mechanisms include gene-fusion, and deletions, that change the domain architecture of a protein [2]. Because of the frequency of such events it is often more useful to analyze protein evolution from a protein domain perspective, that is considering as the evolutionary unit not the protein-coding gene but a protein domain. Often the evolutionary unit coincides with an independent folding structural unit [3]. From earlier studies it is clear that most protein architecture variation comes from addition or deletion of single domains at the N or C-termini [4], but other rearrangements also occur [5]. The exception to this rule is domain-repeat proteins, where it is frequent to have internal tandem duplications of one or more repeated domains [6]. Using novel sequence search methods it has also been detected that many proteins appear to contain hard to identify internal repeats that appear to have a common evolutionary origin [7,8•,9].

The mechanisms described above generate novel protein-coding genes from existing protein coding gene material. However, at some time in history the first protein coding sequence within a protein family must have been created from non-coding genetic material. In the pre-genomics era it was widely assumed that much of present-day genetic diversity could be traced by common ancestry to a molecular big bang, where all genes evolved at once. Already in 1992 this was challenged [10]. The common origin of all proteins is not well supported by the simple fact that many protein families exist today with no resemblance at all to each other [11]. This argues for that *de novo* creation must have occurred at multiple occasions, but it is still not well understood with what frequency such genes appear nor exactly how.

The mechanism of turning non-coding genetic information into a protein-coding gene can be referred to as *de novo* gene creation [10]. By definition, a recently *de novo* created gene should not have any homologs at all and even after some time there should not exist any homologs outside closely related species. Rephrased: when searching a database of all genes, the *de novo* created genes should only have hits in closely related species. Such genes are referred to as orphan genes.

Given the increasing number of completely sequenced genomes definition of ‘closely related’ is crucial for the number of orphans identified. Unfortunately, because genes evolve at different rates it is not possible to define a strict cutoff but all orphans must be put into a phylogenetic perspective. One of the most difficult tasks when trying to identify orphans is to distinguish between fast evolving genes, genes lost in closely related species and *de novo* created genes. When a large number of closely related genomes are present it is at least easier to distinguish between these groups of genes, but still today many proposed *de novo* created genes might actually be seen in a different perspective when more data and/or better methods are available.

In a similar manner to *de novo* gene creation, novel domains may be incorporated into the protein-coding gene. This may occur by the mutation of a stop or start codon, or it might involve the modification of splice signals. This is a fairly unexplored area of research and few studies focus on the emergence of novel protein domains [12••]. Further, there will always be a matter of definition of how to separate insertions of a few residues, on one hand, from the creation of an orphan domain complicating the problem, on the other hand.

In the last years, identification of *de novo* created genes have been reported in *Drosophila* [13,14•,15], mammals [16], primates [17,18,19•,20,21], Fungi [12••,22••,23,24], plants [25,26], Bacteria [27–29] and viruses [30,31]. Orphan genes are specific to a particular lineage and may therefore provide indications of what distinguishes the genetic repertory of one organism from its close relatives; they also lend important clues as to why the organism in question presents a particular phenotype [32•,33]. Further, these proteins may have structures and functions that are not similar to previously known proteins, but are prime candidates for functional novelty.

Most of previous reviews have dwelled on recently duplicated genes [34•,35,36]. Here, we focus on *de novo* genes rather than gene duplicates. However, it is not always easy to differentiate these two classes of gene birth, in particular not for fast evolving genes. The generation of novel eukaryotic genes may not be the exceedingly rare event [19•] it was once believed to be [1,37], mostly due to the complexity of the eukaryotic genes. We now know that most of the genome is transcribed [38]. Therefore, gene creation might be a fairly common event that provides a steady stream of new protein candidates, but where only a fraction of them will become fixed in the population [32•].

Detection and quantification of orphan genes

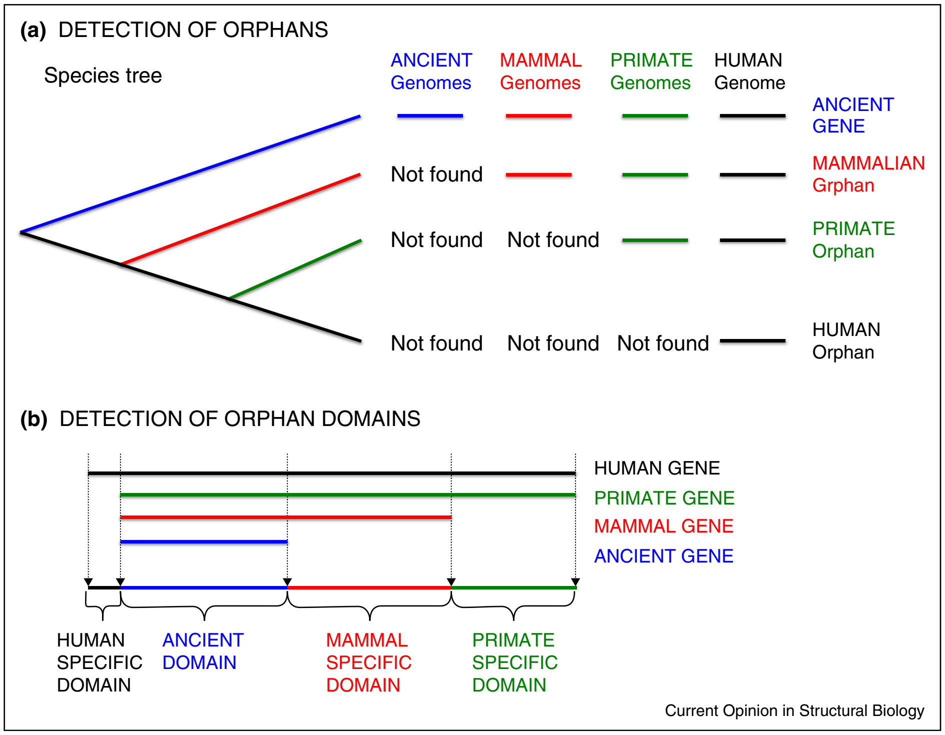
At first, when the yeast chromosomes were sequenced, a large number of genes that had no detectable homologs to any other known genes were discovered [39•]. The general assumption at the time was that once more completely sequenced genomes were accrued, this large number of unique genes would dwindle [40]. Although a more sensitive homologous search detects more distantly related genes [40], the number of orphans (sometimes referred to as ORFans) continued to be 'significantly high' even after dozens of complete genomes had been sequenced [41].

Today, when thousands of genomes have been sequenced, it is necessary to slightly redefine what constitutes an orphan protein (as well as an orphan domain). The definition needs to be put into a phylogenetic perspective, that is a potential orphan gene found in yeast should not be disregarded because it is also found in another, closely related, yeast strain. This method has been referred to as 'phylostratigraphy' [32•]. By taking into account what phylogenetic level the gene orphanicity is confined to, our research group found that 158 orphan genes were unique to *Saccharomyces cerevisiae*, while another 125 genes were unique to the *Saccharomyces* genus [12••].

In theory, the detection of orphan proteins and domains should be straightforward; first detection of all homologs followed by establishing the phylogenetic relationship of the homologs, see Figure 1a. The detection of orphan domains can be done using a similar technique but relying on the alignment of the homologs, see Figure 1b. When applying such a simple method to the current state of eukaryotic reference genomes between 0.2% and 50% of each proteome are found to not have any close homologs, see Figure 2. Clearly genomes that lack closely related organisms, such as Giardia, have a significantly larger fraction of genes without any close homologs than organisms with more closely related species in the database. Even among quite well-characterized genomes, there are notable differences. For instance in *Mus musculus*, 3.03% of the protein coding genes are classified as orphans, while in its closest relative, *Rattus norvegicus*, only 0.75% of the proteins are orphans. It appears likely that these differences are due to the quality of the gene annotations in these genomes, but this requires further analysis. Anyhow, it is clear that even today for most higher eukaryotes we do not have sufficient well-characterized closely related genomes to easily classify the number of true orphans.

Given the existence of thousands of complete genomes today, it might be possible to trace the origin of most protein coding genes and thereby identify most *de novo* created protein coding genes. However, a number of complications exist that need to be addressed when searching for orphans proteins, for example: (i) Also when using the most accurate search methods at our disposal, it is sometimes difficult to detect relationships for fast evolving genes even in closely related species. (ii) The correct identification of protein coding genes from complete genome sequences is non-trivial. In particular, short genes are often missed when assigning genomes. (iii)

Figure 1



Overview of methods for orphan detection. Black represent human genes/domains, green genes/domains found in other primates, red in other mammals and blue genes/domains found outside mammals. **(a)** Detection of orphan genes a at different levels of phylogenetic relationships. The black human orphan is only found within the human genome, the primate specific orphan is also found in other primates etc. **(b)** The lower portion of the figure shows how a similar strategy can be implemented on one gene, resulting in the detection of orphan domains (or regions) at different levels of orphanicity. Here, the black would represent an human specific orphan domain, etc.

Genes are lost or become pseudogenes. (iv) The phylogenetic context for all genes is not always identical for all genes in an organism due to lateral transfers and large-scale genomic rearrangements. (v) Often there is no experimental evidence of a gene-like region being expressed or not, to indicate if it is functional. Below, we discuss these problems and some strategies of how to overcome them.

Incomplete and erroneous gene assignment

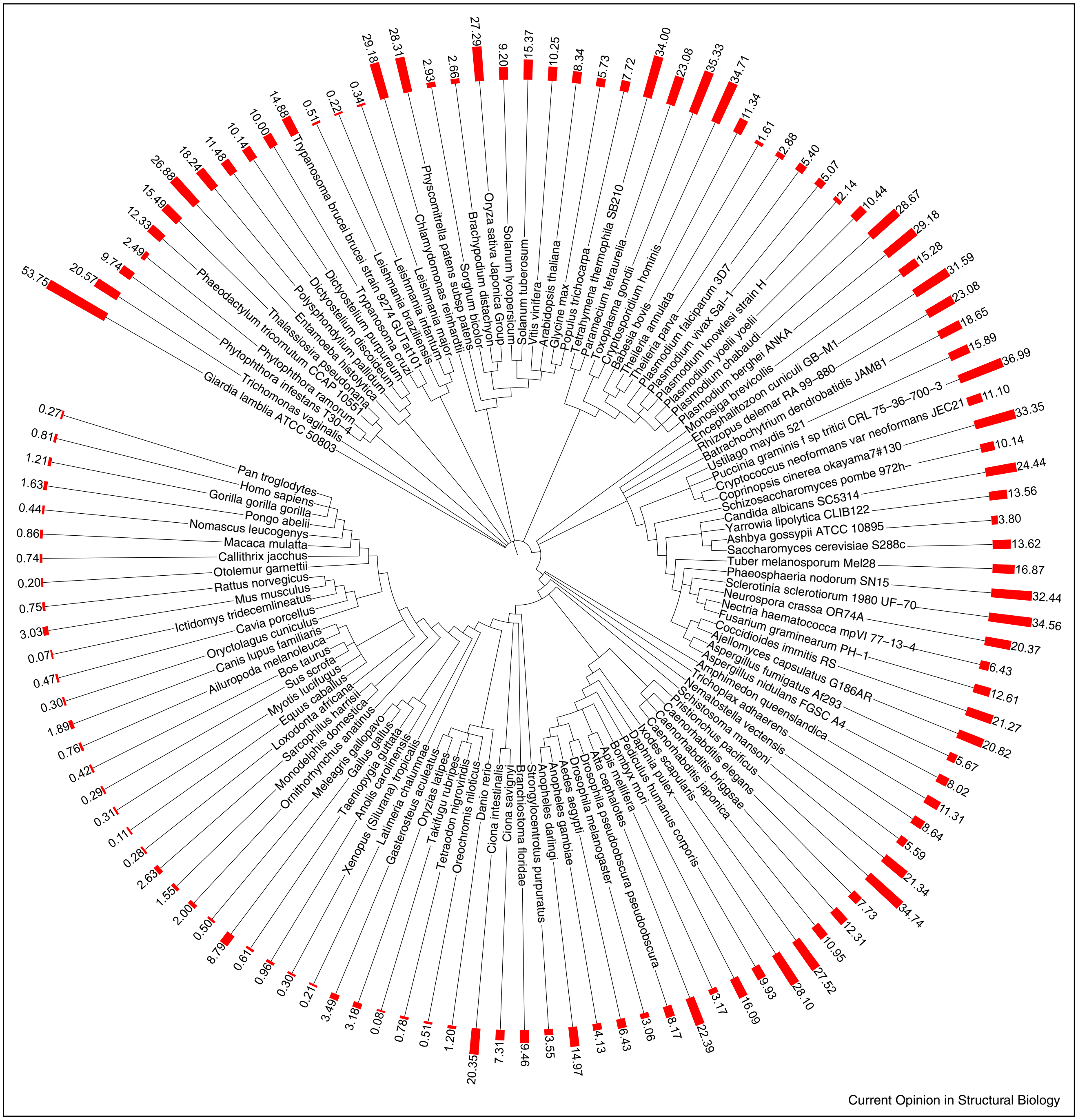
One obvious problem with the approach above is the fact that it relies heavily on current gene prediction programs, that is if a gene is not identified in one organism its homology to another gene cannot be determined and vice-versa. In particular, gene prediction is more difficult in higher eukaryotes and the detection of shorter genes is more difficult. The obvious solution to the problem is to study the DNA sequences directly and not rely on gene predictions.

Further, certain regions of the higher eukaryotic genomes have hitherto been quite difficult to assemble properly, due to large sections of very repetitive regions. With the development of sequencing methods that can provide longer reads, these problems are likely to be resolved in the near future. At the writing of this paper, however, we must note that telomeres and centromeres are currently largely unmappable regions where genes lacking homology to other genes may well be prevalent. This is rather likely since it is known that some of the largest orphan gene families of *S. cerevisiae* reside within the subtelomeric regions [34•].

Eliminating non-orphans by detection of homology between genes.

Although it has been argued that simple BLAST [42] searches should detect most orphans, it is clear that more sophisticated search methods can find distantly related proteins that are missed by BLAST. In particular, the advancements in HMM-HMM search methods have proven to be very powerful [43,44]. However, these methods often break down when it comes to distinguish apparent from true *de novo* orphans due to the, by definition, lack of gene family members needed to build a good HMM. Therefore, one of the best options is to add a search against Pfam [45] or another protein domain database. This methodology detects distant homologs even when homologs in closely related species are missing,

Figure 2



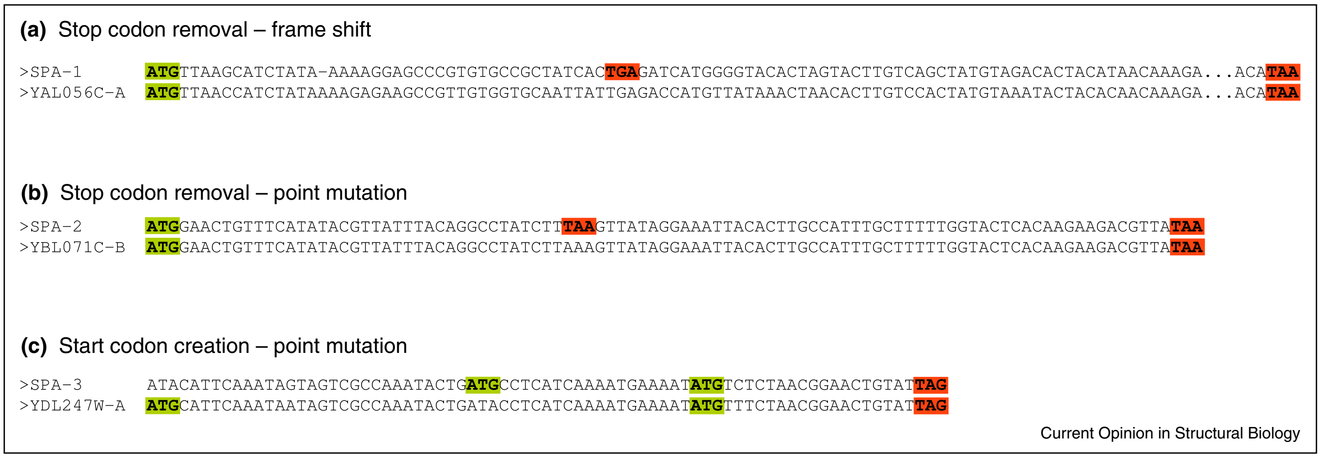
Genes without close homologs in reference eukaryotic proteomes detected using a simple blast search. It can be noted that the species lacking close relatives appear to have more of these genes.

potentially due to gene loss. A match to a domain then needs to be put into its phylogenetic context. Although this method has been used in some papers it is not generally applied in most studies although it helps removing a number of false positives. In our earlier study of orphan genes in yeast we found that about 15% of the potential orphans were removed when Pfam analysis was included [12••].

Mechanisms generating orphans

Many studies of new genes center on duplication, since it is one of the most common ways by which new functions

Figure 3



(a) Stop codon removal – frame shift

```
>SPA-1       ATGTTAAGCATCTATA-AAAAGGAGCCCGTGTGCCGCTATCACTGAGATCATGGGGTACACTAGTACTTGTCAGCTATGTAGACACTACATAACAAAGA...ACATAA
>YAL056C-A   ATGTTAACCATCTATAAAAGAGAAGCCGTTGTGGTGCAATTATTGAGACCATGTTATAAACTAACACTTGTCCACTATGTAAATACTACACAACAAAGA...ACATAA
```

(b) Stop codon removal – point mutation

```
>SPA-2       ATGGAACTGTTTCATATACGTTATTTACAGGCCTATCTTTAAGTTATAGGAAATTACACTTGCCATTTGCTTTTTGGTACTCACAAGAAGACGTTATAA
>YBL071C-B   ATGGAACTGTTTCATATACGTTATTTACAGGCCTATCTTAAAGTTATAGGAAATTACACTTGCCATTTGCTTTTTGGTACTCACAAGAAGACGTTATAA
```

(c) Start codon creation – point mutation

```
>SPA-3       ATACATTCAAATAGTAGTCGCCAAATACTGATGCCTCATCAAAATGAAAATATGTCTCTAACGGAACTGTATTAG
>YDL247W-A   ATGCATTCAAATAATAGTCGCCAAATACTGATACCTCATCAAAATGAAAATATGTTTCTAACGGAACTGTATTAG
```

Current Opinion in Structural Biology

Three possible mechanisms for the proto-gene to gene transition. **(a)** a deletion of one or two nucleotides determines a frame shift, causing the disappearance of a stop codon. **(b)** A stop codon mutates into a coding triplet. **(c)** A point mutation creates a new start codon. YAL056C-A, YBL071C-B and YDL247W-A are *Saccharomyces cerevisiae* genes; SPA-1 to 3 are *Saccharomyces paradoxus* sequences with the following coordinates: SPA-1: AABY01000017.1:88566-88915, SPA-2: AABY01000058.1:43657-43755, SPA-3: AABY01000011.1:3135-3209.

evolve. Duplicated genes are under lowered selective pressure and tend to evolve fast, oftentimes through pseudogenization. Such fast evolving genes are hard to separate from genes that emerge *de novo*. Furthermore, genes sometimes evolve through domain and/or gene fusion and retrotransposition. These important ways of gene birth are, however, distinct from *de novo* gene formation and not the focus of this review.

The first requirement for new gene formation is the physical addition of new genetic material, either through insertion of such material into the genome, or through the acquisition of new coding material from already existing non-coding material. The latter is a process that is more likely to be important in eukaryotic genomes.

Several mechanisms for *de novo* gene formation have been proposed, see Figure 3. First, new domains may emerge through stop codon mutations, generating elongated proteins. Second, new genes may emerge through divergent transcription, where regions in both directions are transcribed although it is just one direction that generates a functional protein coding transcript. Third, in higher eukaryotes, transcription of intergenic regions provides a likely source of *de novo* generated domains. Finally, new genes can emerge from horizontal transfer of non-coding genetic material, as, indeed one of the best documented case of *de novo* gene formation, the antifreeze proteins, illustrates [46•].

The last few years have brought several studies supporting the idea that *de novo* genes emerge through a stepwise procedure, where sporadic transcription constitutes the first step [13,14•,23]. In primates, this has been shown for a number of *de novo* genes by the fact that the transcription profiles of the *de novo* coding gene and the non-coding genes are correlated [18].

De novo formation in different lineages

While the extent of *de novo* gene formation in Bacteria has been known for some time [27], its scope in eukaryotes has only recently been more closely investigated. An important difference between prokaryotes and eukaryotes is the extent of horizontal gene transfer in prokaryotes, which clearly obfuscates the investigation. Bacteria often acquire orphans through bacteriophages [28]. In some cases, this may be an example of horizontal transfer of genetic material, but the material transferred does not always have any detectable similarity to other known protein-coding genes, and therefore could be *de novo* gene formation. However, as viral proteins in general are fast evolving, it is likely that many of the transferred genes actually have distant homologs.

In higher organisms, the corresponding scenario may be more complicated. Indeed, the gene structure is more complicated in higher eukaryotes and an assortment of different signals is required in order for successful transcription and translation. Naturally, the acquisition of novel genes that would possess these properties seems a remote possibility and, therefore, *de novo* gene origination is a difficult scenario to envision. However, in the light of recent findings that much of the genome is transcribed [38], these concerns have lessened. In fact, a recent study showed that there are cryptic signals for transcript processing and regulation in non-coding regions [16], lending credence to the notion of de novo formation out of non-coding regions even in higher eukaryotes. It is possible that *de novo* formation of proto-genes is more common in higher eukaryotes than in yeast due to the lowered selective pressure, small population sizes and large genomes. On the other hand, the more complex cellular structure and interaction networks may pose a threshold to *de novo* formation in higher eukaryotes, as the possibility of disadvantageous effects as a result of the novel genes may increase.

The segmental duplications of the human genome are large and organized into interspersed blocks of duplicated sequences [47,48]. Compared to other organisms, the genomes of great apes are enriched in intrachromosomal duplications [49]. There is a tendency for subtelomeric and pericentromeric clustering of these regions [50]. Primate orphans often originate partially through transposable elements and are located in primate-specific genomic regions. About 1% of primate genes have, according to Toll-Riera *et al.* [17], originated through de novo formation of new genes from non-coding material. This roughly agrees with the numbers presented in Figure 2.

The continuous creation of orphans

By mapping ribosome footprint reads to ORFs, Carvunis *et al.* were able to identify around 1900 proto-genes, genes that provide the reservoir from which *de novo* genes may evolve [22••]. In agreement with these findings, Tautz *et al.* pinpointed that orphans are often short-lived [32•].

The most important questions that remain are: (i) How do orphan genes arise, (ii) what is their fate in the genome and (iii) how do they contribute to the overall phenotype of the organism? The answers to these questions are elusive, but there are a number of promising recent findings, as listed below.

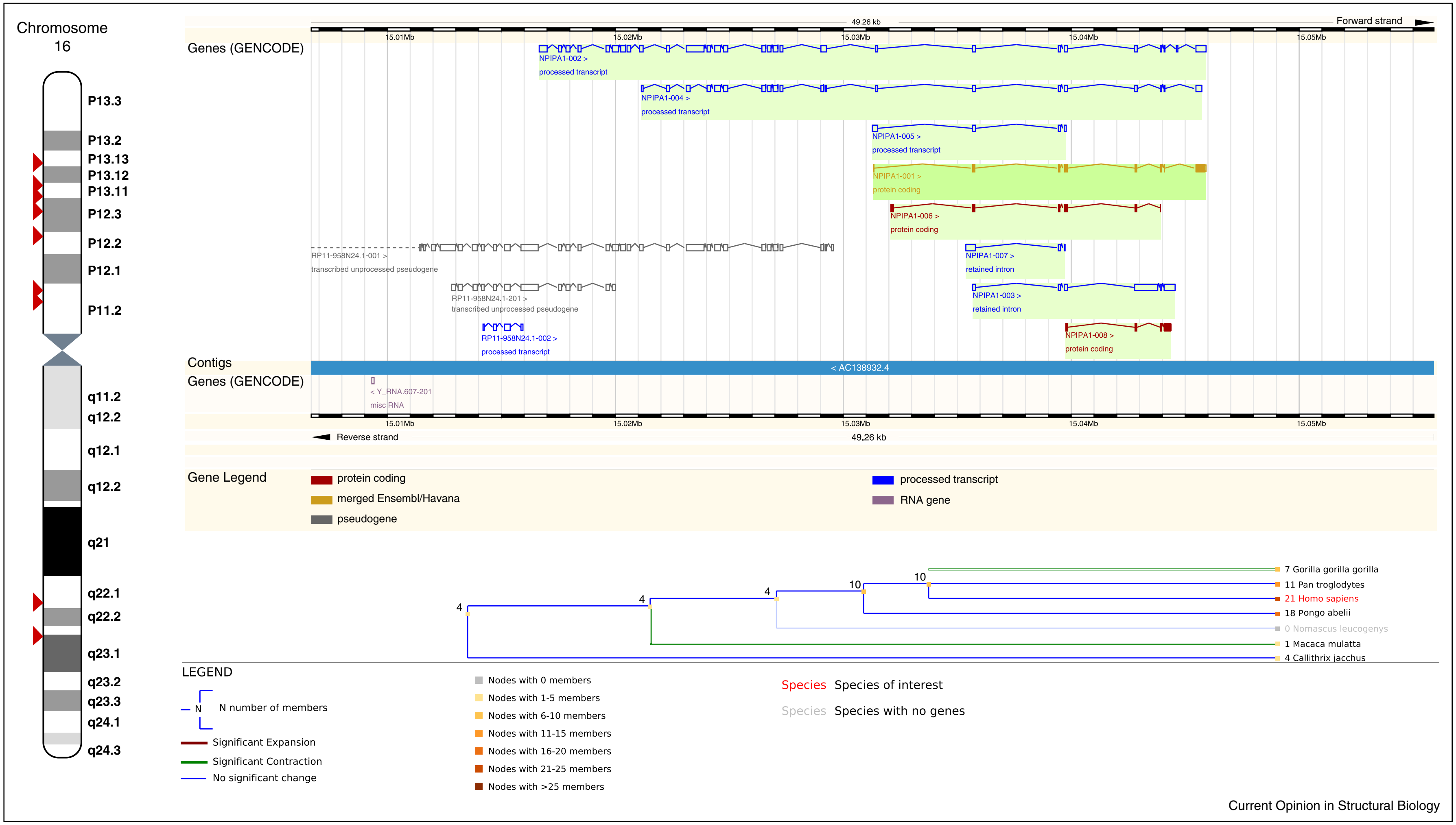
The subtelomeric regions, proximal to the telomeres, in yeast are particularly prone to genetic rearrangements. While most *de novo* genes soon become pseudogenes, there are some rare cases of genes that not only become fixed but also proliferate to form novel gene families. The subtelomeric regions are prone to contain such gene families. Families containing subtelomeric members are evolving much faster, and moreover, are expanding faster than other gene families [34•]. Subtelomeric regions have therefore been proposed as hotbeds for genomic evolution and innovation. Studies of these regions of higher eukaryotes are not readily available since most published genome sequences are largely devoid of subtelomeric sequences due to issues during genome assembly. It is not unlikely that many *de novo* genes originate from subtelomeric regions.

Table 1

Selected examples of reported orphan genes from earlier studies. Although all these proteins have been reported to most likely be *de novo* created it is actually upon careful instruction not as clear as initially believed.

Gene/gene families	Taxa	Functional description
CG15323, CG31406, CG31909, CG32235, CG32582, CG32582, CG32690, CG32690, CG32712, CG34434	*Drosophila*	In *Drosophila*, there is emerging evidence for *de novo* formation of testis-expressed genes [13,14•], indicating that male reproductive functions are under strong selection for novel functions. Additionally, these genes are located in a close neighborhood on the X-chromosome suggesting that there may be some particular genetic mechanism that are responsible for the formation of such genes [14•]. Further experiments on fly show that *de novo* generated genes can arise from non-coding RNAs [65].
O56253_TYMV	Tymovirus	An overlapping protein was found in Tymovirus but no equivalent ORF was found in the closely related potexviruses or carlaviruses nor in outgroups such as the alphaviruses and tobamoviruses and therefore it was assumed to be *de novo* created[10]. However, it clearly bears at least partial similarity to other steroid receptors.
NPIP_HUMAN	primates	In the human genome, one of the most striking reported orphan gene families is the morpheus (NPIP) gene family [66]. This gene cluster appears to be predominantly primate specific, but distantly related homologs can be found in other vertebrates, indicating that it might be fast evolving.
Pfam AFGP family: PF05264	fishes	Another early case of *de novo* formation was presented in 1997 by Chen *et al.* where fish from Arctic and Antarctic waters were shown to contain convergent evolved antifreeze proteins that have arisen *de novo*. The origin of the genetic material was, in this case, microsatellite DNA [46•]. These proteins hold a special place among the *de novo* proteins since both its function and evolutionary history is well studied.
FLJ33706 (B3KR52_HUMAN)	*Homo sapiens*	One protein coding gene, FLJ33706, was proposed to have evolved through an Alu-mediated mechanism and being important for brain function [20]. Although orthologs are primarily found in primates distant homologs can actually be found in most mammals using Pfam, indicating that it might be fast evolving.
C22orf45 AAS1_HUMAN	*Homo sapiens*	One of three human genes identified as *de novo* from a recent study [19•]. The potential function of this protein is unknown and according to UNIPROT it is unclear if it is real coding for a protein, that is it might be a pseudo-gene.

Figure 4



Current Opinion in Structural Biology

The primate orphan genes, the Morpheus family. To the left the chromosome 16 is shown, where all copies of the Morpheus genes reside. The upper panel shows the different transcripts for one of the Morpheus genes, including non-coding transcripts (blue), protein-coding transcripts (red) and Havana transcripts (gold). The lower panel shows the number of Morpheus family members present in each species.

It has been proposed that many viral proteins are created *de novo* and as mentioned above, the first *de novo* identified gene was found in a virus [10]. One noted mechanisms for the creation is overprinting of a new reading frame onto an existing ('ancestral') frame [31]. Second, an 'out of testis' hypothesis (i.e. that the meiotic state in the testis makes it easier for orphans to appear) has been proposed for the creation of orphan proteins [13,14•,51]. The out of the testis hypothesis is based on that transcription of new genes might be easier in some types of testis cells because of the overall permissive chromatin state and overexpression of key components of the transcriptional machinery in these cells [51]. When transcribed, new genes may evolve more efficient promoters and eventually also evolve more diverse expression patterns and thus also obtain functions in other tissues. 28 new potentially orphan human transcripts formed by exon exaptation and exon shuffling primarily expressed in testis has been detected [50]. Finally, mammalian promoters and enhancers often direct transcription initiation in opposite orientations, a phenomenon called divergent transcription. This has been suggested as a possible mechanism by which new genes may emerge [52]. For instance, the human gene MYEOV, see Table 1, was probably derived from an intergenic enhancer [52] and the Morpheus family has probably expanded in the human genome, see Figure 4. However, when carefully analyze potential distant homologs, both these proteins might actually examples of fast evolving proteins.

Characteristics of orphan proteins and orphan domains

As described above the separation of truly *de novo* created proteins from rapidly evolving proteins is difficult and can only reliably be made for genomes with many closely related fully sequenced genomes. Given the problem of lateral transfers in prokaryotes and the difficulty of gene assignments in higher eukaryotes, most studies of properties of true orphans are performed in *S. cerevisiae*. When comparing different levels of orphanicity in yeast it was found that orphans detected at the *S. cerevisiae* level had other features than those detected at the *Saccharomyces* level [12••]. In particular, it is now evident that features such as low complexity and intrinsic disorder are underrepresented among the most recent orphans. This indicate that within the more ancient group many proteins were most likely not *de novo* created and therefore rather rapidly evolving genes rather than orphans. Thus, for example disordered sequences could appear to be orphan because they often evolve more rapidly than other sequences [53]. Another observation regarding recently created orphans is that these are enriched in transmembrane proteins and have a higher average hydropathy [22••].

A similar trend can be found for orphan domains, that is the most recent ones do not appear to be enriched in intrinsically disorder residues but older ones do. Protein elongation does to a large extent occur within disordered regions [54]. Therefore, disordered regions might appear to be orphan because they evolve more rapidly than other regions [55,56]. Further, short (or even long) nucleotide motifs may induce nucleotide duplication and such repeats might be more common in disordered proteins [53].

Final conclusions

When analyzing relationships between genes in closely related genomes it appears that *de novo* creation of protein coding genes is a continuous process, at least in yeast [22••]. This process seems to have generated protein coding genes in *S. cerevisiae* [12••]. Further, a number of suggestive cases in other organisms have also been reported, see Table 1. This is in strong contrast to the earlier picture where most protein families were created by a 'big bang' of genetic diversity followed by very scant novel gene origination. However, still the difficulty to reliably detect distantly related proteins blurs the picture of how frequent *de novo* creation of genes is.

An alternative, and to some extent contradictory, picture to quantify orphan proteins can be obtained by clustering all known protein sequences. This has been used to estimate the number of superfamilies (i.e. homologous protein families) and folds (i.e. structurally similar but possibly analogous proteins). Today the protein family databases as integrated into Interpro release 45.0 [57] describes 17,085 protein families, 7133 protein domains and 274 repeat signatures. These families cover about 80–90% of all proteins in UniProt [58]. Further, current structural databases classify proteins into 2626 superfamilies [59] or 1195 folds [11]. There seems to be a consensus today, after years of alternating views, that in total there are in the order of 10^5 superfamilies and 10^3 folds [60•]. Given this rather limited set of protein domain families it can be assumed that even if there is a continuous flow of novel protein coding genes only a very small fraction of these *de novo* created genes gets fixed in the population and expand to become large protein families. This highlights the well known fact that protein domain frequencies follows a power-law distribution [61].

From what has been learned about protein evolution during the last decades there are additional evolutionary events that have to be taken into account when addressing the number of folds. There are a number of evolutionary mechanisms including 3D domain swaps [62], circular permutations [5], strand invasions, structural variations within superfamilies [63] and repeat duplications [6] that change the structure of a protein to such a degree that it prevents the detection of homology even using the most sensitive methods.

Actually, many distant homologous relationships can be found between proteins even of different folds [60•] and it is also proposed that repeated units in for example outer membrane β-barrels [7], Tim-Barrels [8•] and α-helical membrane proteins [9] appear to all have a common ancestry. These findings point toward the idea that at least a significant fraction of protein coding genes have arisen from a limited set of short peptide ancestors (antecedent domain segments) [64]. Assuming that such a mechanism is present, this further blurs the distinction between fast evolving genes and orphans, as different parts of a domain might have different homologous relationships. A similar complication obviously exist for multi-domain proteins as each domain might have its own evolutionary history.

Although the extent to which *de novo* creation occurs is not yet clear, in particular for higher eukaryotes, we can say that the consensus view of *de novo* creation has changed, much thanks to, first, studies of prokaryotes in the early 2000s, then by detailed studies of higher eukaryotes that can be performed in the light of next generation sequencing and, finally, by the large scale studies on yeast by Carvunis *et al.* [22••].

It is now time to revise that idea once proposed by Francois Jacob — 'The probability that a functional protein would appear *de novo* by random association of amino acids is practically zero. In organisms as complex and integrates as those that were already living along time ago, creation of entirely new nucleotide sequences could not be of any importance in the production of new information' [37]. Not only is evolution a tinkerer working on already extant gene duplicates, as Jacob proposed in his seminal paper, but it is also blindly tinkering on non-coding regions as well as coding regions, not only to the extent of mere mutations, but also occasionally sending in the bench warmers to the big game. However, the extent to which the *de novo* generated genes are functional, or indeed detrimental, in particular in highly complex organisms, remains to be seen.

Acknowledgements

This work was supported by grants from the Swedish Research Council (VR-NT 2009-5072, 2012-5046, VR-M 2010-3555), SSF, the Foundation for Strategic Research, Science for Life Laboratory the EU 7th framework through the EDICT project, contract No: FP7-HEALTH-F4-2007-201924. Funding for SL was provided by BILS, Bioinformatics Infrastructure for Life Science.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ohno S: *Evolution by Gene Duplication*. New York: Springer; 1970, .

2. Ekman D, Bjorklund A, Frey-Skott J, Elofsson A: **Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions**. *J Mol Biol* 2005, **348**:231-243.

3. Elofsson A, Sonnhammer E: **A comparison of sequence and structure protein domain families as a basis for structural genomics**. *Bioinformatics* 1999, **15**:480-500.

4. Bjorklund A, Ekman D, Light S, Frey-Skott J, Elofsson A: **Domain rearrangements in protein evolution**. *J Mol Biol* 2005, **353**:911-923.

5. Weiner J 3rd, Bornberg-Bauer E: **Evolution of circular permutations in multidomain proteins**. *Mol Biol Evol* 2006, **23**:734-743.

6. Bjorklund A, Ekman D, Elofsson A: **Expansion of protein domain repeats**. *PLoS Comput Biol* 2006, **2**:e114.

7. Remmert M, Biegert A, Linke D, Lupas A, Soding J: **Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin**. *Mol Biol Evol* 2010, **27**:1348-1358.

8. • Soding J, Remmert M, Biegert A: **HHrep: de novo protein repeat detection and the origin of TIM barrels**. *Nucleic Acids Res* 2006, **34(Web Server issue)**:W137-42.
Here it is shown by using improved sequence search methods that all repeat units in TIM-barrels might have a common origin.

9. Hennerdal A, Falk J, Lindahl E, Elofsson A: **Internal duplications in alpha-helical membrane protein topologies are common but the nonduplicated forms are rare**. *Protein Sci* 2010, **19**:2305-2318.

10. Keese P, Gibbs A: **Origins of genes: "big bang" or continuous creation?** *Proc Natl Acad Sci U S A* 1992, **89**:9489-9493 Probably the first report of a well-characterized orphan gene.

11. Murzin A, Brenner S, Hubbard T, Chothia C: **Scop: a structural classification of proteins database for the investigation of sequences and structures**. *J. Mol. Biol.* 1995, **247**:536-540.

12. •• Ekman D, Elofsson A: **Identifying and quantifying orphan protein sequences in fungi**. *J Mol Biol* 2010, **396**:396-405.
One of the first studies that use closely related genomes to identify de novo by using closely related genomes. It was noted that orphans detected at the *S. Cerevisiae* level had other features than those detected at the Cerevisiae level, and it was concluded that many at the later level most likely were not de novo created and therefore rather rapidly evolving genes rather than orphans. One of the papers that discuss orphan domains.

13. Begun D, Lindfors H, Kern A, Jones C: **Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/ Drosophila erecta clade**. *Genetics* 2007, **176**:1131-1137.

14. • Levine M, Jones C, Kern A, Lindfors H, Begun D: **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* 2006, **103**:9935-9939.
Documentation of orphan genes predominantly found expressed in testis in Drosophila.

15. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W: **On the origin of new genes in Drosophila**. *Genome Res* 2008, **18**:1446-1455.

16. Heinen T, Staubach F, Haming D, Tautz D: **Emergence of a new gene from an intergenic region**. *Curr Biol* 2009, **19**:1527-1531.

17. Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Alba M: **Origin of primate orphan genes: a comparative genomics approach**. *Mol Biol Evol* 2009, **26**:603-612.

18. Xie C, Zhang Y, Chen J, Liu C, Zhou W, Li Y, Zhang M, Zhang R, Wei L, Li C: **Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs**. *PLoS Genet* 2012, **8**:e1002942.

19. • Knowles D, McLysaght A: **Recent de novo origin of human protein-coding genes**. *Genome Res* 2009, **19**:1752-1759.
Here it is reported that 0.075% of human genes appear to be de novo created.

20. Li C, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang P, Lu S, Li X, Yu Q, Zheng X, Du Q, Uhl G, Liu Q, Wei L: **A human-specific de novo protein-coding gene associated with human brain functions**. *PLoS Comput Biol* 2010, **6**:e1000734.

21. Wu D, Irwin D, Zhang Y: **De novo origin of human protein-coding genes**. *PLoS Genet* 2011, **7**:e1002379.

22. •• Carvunis A, Rolland T, Wapinski I, Calderwood M, Yildirim M, Simonis N, Charloteaux B, Hidalgo C, Barbette J, Santhanam B, Brar G, Weissman J, Regev A, Thierry-Mieg N, Cusick M, Vidal M: **Proto-genes and de novo gene birth**. *Nature* 2012, **487**:370-374.

This paper proposes that continuous gene birth occurs in *S. Cerevisiae* by turning non-coding proto-ORFs to functional ORFs. They study about 6000 annotated and 261,000 unannotated ORFs in *S. Cerevisiae* and identify their presence in ten other yeast genomes. Each PRF is then classified by its conservation level and analyzed further.

23. Cai J, Zhao R, Jiang H, Wang W: **De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae***. *Genetics* 2008, **179**:487-496.

24. Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W: **A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand**. *Cell Res* 2010, **20**:408-420.

25. Felippes F, Schneeberger K, Dezulian T, Huson D, Weigel D: **Evolution of arabidopsis thaliana microRNAs from random sequences**. *RNA* 2008, **14**:2455-2459.

26. Xiao W, Liu H, Li Y, Li X, Xu C, Long M, Wang S: **A rice gene of de novo origin negatively regulates pathogen-induced defense response**. *PLoS One* 2009, **4**:e4603.

27. Daubin V, Ochman H: **Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli***. *Genome Res* 2004, **14**:1036-1042.

28. Daubin V, Ochman H: **Start-up entities in the origin of new genes**. *Curr Opin Genet Dev* 2004, **14**:616-619.

29. Delaye L, Deluna A, Lazcano A, Becerra A: **The origin of a novel gene through overprinting in *Escherichia coli***. *BMC Evol Biol* 2008, **8**:31.

30. Sabath N, Wagner A, Karlin D: **Evolution of viral proteins originated de novo by overprinting**. *Mol Biol Evol* 2012, **29**:3767-3780.

31. Pavesi A, Magiorkinis G, Karlin D: **Viral proteins originated de novo by overprinting can be identified by codon usage: application to the ''gene nursery'' of delta retroviruses**. *PLoS Comput Biol* 2013, **9**:e1003162.

32. Tautz D, Domazet-Loso T: **The evolutionary origin of orphan**
• **genes**. *Nat Rev Genet* 2011, **12**:692-702.
A recent review that described both de novo and duplicated genes.

33. Cai J, Petrov D: **Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes**. *Genome Biol Evol* 2010, **2**:393-409.

34. Brown C, Murray A, Verstrepen K: **Rapid expansion and**
• **functional divergence of subtelomeric gene families in yeasts**. *Curr Biol* 2010, **20**:895-903.
Reporting the importance of subtelomeric regions for the creation of orphans in yeast.

35. Capra J, Pollard K, Singh M: **Novel genes exhibit distinct patterns of function acquisition and network integration**. *Genome Biol* 2010, **11**:R127.

36. Chen S, Krinsky B, Long M: **New genes as drivers of phenotypic evolution**. *Nat Rev Genet* 2013, **14**:645-660.

37. Jacob F: **Evolution and tinkering**. *Science* 1977, **196**:1161-1166.

38. Djebali S, Davis C, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov G, Khatun J, Williams B, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid R, Alioto T, Antoshechkin I, Baer M, Bar N, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood M, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo O, Park E, Persaud K, Preall J, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L, Shahab A, Skancke J, Suzuki A, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis S, Hannon G, Giddings M, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras T: **Landscape of transcription in human cells**. *Nature* 2012, **489**:101-108.

39. Tanaka S, Isono K: **Correlation between observed transcripts**
• **and sequenced ORFs of chromosome III of *Saccharomyces cerevisiae***. *Nucleic Acids Res* 1993, **21**:1149-1153.
The first report trying to quantify the amount of orphan genes from a genomic perspective.

40. Dujon B: **The yeast genome project: what did we learn?** *Trends Genet* 1996, **12**:263-270.

41. Fischer D, Eisenberg D: **Finding families for genomic ORFans**. *Bioinformatics* 1999, **15**:759-762.

42. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

43. Soding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**:951-960.

44. Johnson L, Eddy S, Portugaly E: **Hidden markov model speed heuristic and iterative HMM search procedure**. *BMC Bioinformatics* 2010, **11**:431.

45. Punta M, Coggill P, Eberhardt R, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer E, Eddy S, Bateman A, Finn R: **The Pfam protein families database**. *Nucleic Acids Res* 2012, **40(Database issue)**:D290-D301.

46. Chen L, DeVries A, Cheng C: **Evolution of antifreeze**
• **glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish**. *Proc Natl Acad Sci U S A* 1997, **94**:3811-3816.
One of the best characterized orphan proteins in a higher eukaryote.

47. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson R, Paabo S, Rocchi M, Eichler E: **A genome-wide comparison of recent chimpanzee and human segmental duplications**. *Nature* 2005, **437**:88-93.

48. Eichler E, Johnson M, Alkan C, Tuzun E, Sahinalp C, Misceo D, Archidiacono N, Rocchi M: **Divergent origins and concerted expansion of two segmental duplications on chromosome 16**. *J Hered* 2001, **92**:462-468.

49. She X, Liu G, Ventura M, Zhao S, Misceo D, Roberto R, Cardone M, Rocchi M, NISC Comparative Sequencing Program, Green E, Archidiacano N, Eichler E: **A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications**. *Genome Res* 2006, **16**:576-583.

50. She X, Horvath J, Jiang Z, Liu G, Furey T, Christ L, Clark R, Graves T, Gulden C, Alkan C, Bailey J, Sahinalp C, Rocchi M, Haussler D, Wilson R, Miller W, Schwartz S, Eichler E: **The structure and evolution of centromeric transition regions within the human genome**. *Nature* 2004, **430**:857-864.

51. Kaessmann H, Vinckenbosch N, Long M: **RNA-based gene duplication: mechanistic and evolutionary insights**. *Nat Rev Genet* 2009, **10**:19-31.

52. Wu X, Sharp P: **Divergent transcription: a driving force for new gene origination?** *Cell* 2013, **155**:990-996.

53. Kellis M, Patterson N, Endrizzi M, Birren B, Lander E: **Sequencing and comparison of yeast species to identify genes and regulatory elements**. *Nature* 2003, **423**:241-254.

54. Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A: **Protein expansion is primarily due to indels in intrinsically disordered regions**. *Mol Biol Evol* 2013, **30**:2645-2653.

55. Romero P, Obradovic Z, Li X, Garner E, Brown C, Dunker A: **Sequence complexity of disordered protein**. *Proteins* 2001, **42**:38-48.

56. Brown C, Takayama S, Campen A, Vise P, Marshall T, Oldfield C, Williams C, Dunker A: **Evolutionary rate heterogeneity in proteins with long disordered regions**. *J Mol Evol* 2002, **55**:104-110.

57. Apweiler R, Attwood T, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning M, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder N, Oinn T, Pagni M, Servant F, Sigrist C, Zdobnov E: **InterPro Consortium Interpro-an integrated documentation resource**

for protein families domains and functional sites. *Bioinformatics* 2000, **16**:1145-1150.

58. Apweiler R, Bairoch A, Wu C, Barker W, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin M, Natale D, O'Donovan C, Redaschi N, Yeh L: **Uniprot: the universal protein knowledgebase**. *Nucleic Acids Res* 2004, **32(Database issue)**:D115-D119.

59. Marsden R, Lee D, Maibaum M, Yeats C, Orengo C: **Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space**. *Nucleic Acids Res* 2006, **34**:1066-1080.

60. • Alva V, Remmert M, Biegert A, Lupas A, Soding J: **A galaxy of folds**. *Protein Sci* 2010, **19**:124-130.
A paper showing that many distinct folds have detectable sequence similarity between peptides.

61. Qian J, Luscombe N, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model**. *J Mol Biol* 2001, **313**:673-681.

62. Bennett M, Choe S, Eisenberg D: **Domain swapping: entangling alliances between proteins**. *Proc Natl Acad Sci U S A* 1994, **91**:3127-3131.

63. Reeves G, Dallman T, Redfern O, Akpor A, Orengo C: **Structural diversity of domain superfamilies in the CATH database**. *J Mol Biol* 2006, **360**:725-741.

64. Lupas A, Ponting C, Russell R: **On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?** *J Struct Biol* 2001, **134**:191-203.

65. Reinhardt J, Wanjiru B, Brant A, Saelao P, Begun D, Jones C: **De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences**. *PLoS Genet* 2013, **9**:e1003860.

66. Johnson M, Viggiano L, Bailey J, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler E: **Positive selection of a gene family during the emergence of humans and African apes**. *Nature* 2001, **413**:514-519.