The discovery of *de novo* gene evolution

Diethard Tautz

Max-Planck Institute for Evolutionary Biology

24306 Plön, Germany

tautz@evolbio.mpg.de

Abstract

Genes can evolve via duplication and divergence mechanisms, but also *de novo* out of non-coding intergenic sequences. This latter mechanism is only now fully appreciated, while the former mechanism became an almost exclusive dogma for a long time. Here I make an attempt to look into the history of this development and why a dogma was developed, although the alternative had hardly been explored. It appears that an important part on our understanding the nature of genes and their evolutionary origin had escaped our attention because of the prevailing dogma. Evidence is now rapidly accumulating that *de novo* evolution is an extremely active mechanism for generating novelty in the genome and this will require a new look at how we envisage genes to arise and become functional.

The genomic revolution has brought about unprecedented large datasets that allow full genome comparisons between many different species. But genome sequences consist solely of strings of the four letters that represent the four nucleotides in DNA. Making sense out of them requires assumptions on what kind of patterns we can expect to find in these nucleotide strings. Most of these assumptions were formed long before the first full genome sequence was available. Finding the genes in the genome sequences is evidently at the forefront of this pattern interpretation task, but it is immediately obvious that this is directly dependent on what we call a "gene" and how we define it. Typically, we currently expect to find in new genome sequences gene regions with open reading frames (ORFs), in eukaryotes usually split into exons and introns, as well as their associated regulatory regions. In addition, there are various types of repetitive elements, usually derived from transposons and viruses, but also genes can occur as more or less perfect repeats. Finally, there are the intergenic regions which cover everything that does not fall into one of the previous classes. Intergenic regions are supposed to have no major function and their evolution is expected to follow more or less random rules. This does not necessarily imply a random base composition, since mutational processes, slippage synthesis and recombination mechanisms shape the base composition as well. But this will not be further discussed here - the important point is that most genomes harbor long strings of nucleotides in addition to the canonically identified gene regions. Interestingly, these intergenic regions now turn out to provide a rich source for the emergence of new genes via *de novo* evolution..

The focus of this article will be on tracing how the ideas on the emergence of genes have developed over time and have thus also determined our expectations of how to identify genes in genome sequences. I will show that there was one predominant dogma for a long time, namely duplication of existing genes followed by divergence into new functions. In its consequence, this dogma implied that all genes existed already in a primordial unicellular ancestor, i.e. the mechanism of their origin was relegated to unknown factors in early evolution that might even have occurred under different biochemical conditions (Lupas 2001). Much evidence was collected to support this dogma, while the alternative, namely *de novo* emergence of genes from intergenic regions was largely neglected. It was proposed early on, but not seriously considered for a long time. This picture is currently completely changing and both, re-analysis of existing data, as well as a wealth of new comparative genome information suggests that *de novo* evolution may actually be a predominant mechanism for gene emergence (Tautz and Domazet-Losos, 2011). This is a radical shift in thinking and has evidently also ramifications for understanding the nature of genes and their transition phases from non-coding to fully functional sequences.

How was it possible that the objective alternatives of gene emergence were not seriously considered and investigated for a long time? In the following I will try to trace the history of the thoughts by looking at some landmark papers, followed by a short description of the current state of knowledge. I will then come back to this question. Since I am not a historian, I can only touch on some aspects of this history, without a claim to have captured it completely. But already these aspects constitute an interesting example of how biological ideas are shaped, how much they depend on intuition and how seemingly stringently argued considerations can distract from finding the full truth.

Evolution by gene duplication

This was the title of a landmark book published by Susumu Ohno (1970). It has to be seen as culmination of a long history of thinking about gene duplication as an evolutionary mechanism that has started in the early 20th century. The initial evidence was largely based on incidental observations, but the ensuing conceptualization of the consequences of gene duplication reached eventually a point where it became a dogma. Taylor and Raes (2004) made an attempt to understand the early history of this development and I am largely following their findings.

Plant geneticists were among the first to recognize chromosome number variation between varieties and species and to propose that this could be of evolutionary significance (Kuwada 1911; Tischler 1915), which appears to have led to a first discussion on the significance of the duplication of genetic material. In 1918, Calvin Bridges working in the group of T. H. Morgan on *Drosophila* genetics, suggested at a conference: "*... that the main interest in duplications lay in their offering a method for evolutionary increase in lengths of chromosomes with identical genes which could subsequently mutate separately and diversify their effects*" (cited in Bridges 1935). Hence, he proposed already at that time the essence of the duplication-divergence idea of gene evolution, although there was no material concept for a gene yet. Similarly, even the sub-functionalization concept of gene evolution was proposed before the molecular nature of genes became clear. It was apparently first put forward by Serebrovsky (1938), also a member of the Morgan group. He concluded that "*This principle of loss of duplicate functions by one of the homologues in the process of genic evolution .... should result in a specialization of genes, when each then fulfills only one function which is strictly limited and important for the life of the organism.*". The first ideas of the molecular nature of genes emerged in the 1940s (Gulick 1944; Beadle 1945) allowing to revisit the "Possible significance of duplication in Evolution", which is a title of a paper by Stephens (1951) in which he came to the conclusion that the duplication-divergence mechanism is a very attractive possible mechanism, but that no fully compelling evidence existed at that time. Interestingly, he mentioned also explicitly *de novo* evolution of genes as the obvious alternative mechanism, but did not go on to discuss it, since he considered it too difficult to find evidence for it: "*Present knowledge is quite inadequate to determine whether it is possible for new genetic loci to arise de novo, or, in fact, to test its occurrence if the possibility existed, but the alternative mode of origin (duplication) is a well-established phenomenon*." This was indeed the state for many years to follow, almost like taking a decision at a cross-road. Solid evidence for a role of gene duplication accumulated quickly, while exploring the possibility of *de novo* evolution of genes would have been a path that did not seem very rewarding to take.

Hence, a rich history of thinking about the role of gene duplication in evolution preceded Ohno´s 1970 book (including one of his own landmark papers, Ohno 1968). The book mostly summarized many pieces of evidence, much of which were based on chromosome analyses and the at that time flourishing discipline of enzyme electrophoresis. In addition to the book, Ohno was also a charismatic figure and very active at conferences. Accordingly, he did his best to put forward his message on gene (and genome) duplication, thus consolidating and confirming a view that was already broadly acceptable at that time.

Tinkering and regulatory change

Another influential paper was Jacob´s treatise on "Evolution and Tinkering" published in 1977. It is based on a lecture given at Berkeley in the same year, covering a broad spectrum of ideas, but some key sentences keep being cited until today. One of them is on the origin of novelties: "*Novelties come from previously unseen association of old material. To create is to recombine.*" and "*Evolution does not produce novelties from scratch*" (Jacob, 1977). He went on to expand this specifically to considerations about the origin of new genes: "*The probability that a functional protein would appear de novo by random association of amino acids is practically zero*." and "*...creation of entirely new nucleotide sequences could not be of any importance in the production of new information*." Of course, he acknowledges that genes and proteins must have come from somewhere, but he relegated this to a pre-biotic phase: "*The really creative part of biochemistry must have occurred very early*". Gene duplication and recombination would then have ensured the further evolution: "*... specialization and diversification occurred by using differently the same structural information*". Further he alludes also to regulatory evolution as a mechanism to create diversity (Jacob, 1977).

In fact, the concept of regulatory evolution rather than genic evolution as a factor for explaining diversity was at that time already well developed. It emerged when systematic measurements of DNA content in cells became possible in the 1950s revealing a high constancy within germ cells of a species, but large differences between species (Mirsky and Ris, 1951). Britten and Davidson speculated in 1969 that evolution of regulatory regions may account for this variation in DNA content and they suggested that this would be more important than the evolution of gene numbers (Britten and Davidson, 1969). Another landmark paper suggesting a predominant role of regulatory evolution was published by King and Wilson (1975). They noted a high similarity of protein sequences between human and chimpanzees which they thought would not be enough to explain the anatomical and behavioral differences between these species. They concluded: "*A relatively small number of genetic changes in systems controlling the expression of genes may account for the major organismal differences between humans and chimpanzees*."

This conceptual idea of evolution being based on a set of building blocks that are duplicated and recombined in combination with a predominance of regulatory evolution was then further supported by the results of the research on the evolution of developmental mechanisms (often called evo-devo research). General principles of early development were found to be driven by sets of conserved proteins, put together in differing regulatory contexts to generate differences in embryonic development. Duboule and Wilkins (1998) called this "bricolage", specifically referring to Jacob´s 1977 paper for the origin of these ideas.

The emergence of protein domains

A further refinement of the duplication-divergence concept was attained when a large number of protein structures were solved, allowing to do systematic comparisons of structures. In 1991, Cyrus Chothia came up with a bold suggestion that there could be an upper limit of only about 1,000 protein families representing all genes in all organisms (Chotia, 1991). The argument was based on calculating the discovery rate of new protein structures, compared to confirming known structures that were already obtained through crystallography. The possible argument that structures could have arisen independently, rather than by gene duplication, was countered by showing that the majority of them had not only similar structures, but also conserved sequences, which is the hallmark of duplication. Intriguingly, he was very well aware of cases where structural similarity existed in the absence of sequence similarity, but he discussed this away by arguing that structure could be retained even when the underlying amino acid sequences would change during evolution, up to a point where similarity would not be recognizable anymore. He even used this argument for correcting a first calculation of 1,500 families down to 1,000. His paper is remarkable in many aspects. First, although extremely short, in includes all relevant arguments around this topic until today. Second, it reflects a complete adoption of the dogma that new genes can only arise by duplication-divergence, i.e. all current genes must be reducible to a small set of primordial genes. Third, it implies another dogma, namely that proteins can only function when they are stably folded. And it is remarkable that the bold claims turned out to be largely true in at least one respect: there is only a limited number of stable protein folds in today´s protein, although now estimated to be in the order of 1,400 (Orengo et al. 2005) and at least the majority of them can be traced back to the first cellular ancestor of all life on earth. But Chotia´s line of argument did also a rather bad service to any alternative ideas of how new proteins could come about. The simple claim that it would only be a question of time until all structures are solved effectively prohibited creative thinking into other directions.

The mystery of Orphans

The first completed "genome" project was the full sequencing of chromosome III of yeast in 1992 (Oliver et al. 1992). It was actually not a full genome that was sequenced, but only the first long contiguous stretch of eukaryotic DNA. Still, it allowed for the first time to systematically annotate the DNA and count the genes in it. In a comment on the subsequent full sequencing of all other chromosomes, Dujon stated in 1996 that "*The most striking result from the chromosome III sequence was that approximately half of all protein-coding ORFs revealed by the sequence, had no clear-cut sequence homologs in any organism...*". He called this "The mystery of the orphans" (Dujon 1996) and thus introduced a generic name for this class of sequences (later also called "ORFans"). Interestingly, by choosing this name, he implied that these genes should have had "parents" that somehow got lost, i.e. even this choice of a name reflected the prevailing dogma (the name had initially also a second meaning, referring to genes for which no function was known, but this meaning was lost over time).

Why should it have been a mystery to find new genes when it was one of the explicit goals of genome projects to identify all genes in an organism? Indeed much of the discussion in Dujon´s review centered around the question why the sequencing project found so many more genes than the extensive previous genetic analyses in yeast. However, he addressed also the point whether it would eventually be possible to find homologues of the orphans in other organisms, i.e. whether the expectation that all proteins should eventually fall into known gene families would be fulfilled. He concluded: "*...present numerical trends suggest the possibility that there will eventually remain a core of irreducible orphans, specific to the yeast genome, perhaps because they are genes that evolve more rapidly than other*s." (Dujon 1996).

This view was later challenged in a comment by Fischer and Eisenberg (1999), who discussed the outcomes of the genome projects of prokaryotes. They opened their comment by re-stating the dogma: "*Why, if proteins in different organisms have descended from common ancestral proteins by duplication and adaptive variation, do so many today show no similarity to each other?*". In their discussion they ruled out that there are simple explanations for the increasing number of genes found without apparent homologs, i.e. they confirmed that the phenomenon is true. Still, they concluded that it would simply be a matter of generating more data and refined algorithms to eventually remove this phenomenon: "*...until the 3D structures of ORFans are experimentally determined, more sensitive bioinformatic tools will aid in placing genomic ORFans into their proper protein superfamilies*." Hence, they were still not prepared to even consider alternative models of gene and protein emergence.

With the completion of more and more genome projects, it became clear that the phenomenon of orphans would not disappear, i.e. it was not a question of obtaining more data to put them into superfamilies. Instead, they had to be considered as true genes, restricted to specific evolutionary lineages. A new term for these genes became now popular, namely "taxonomically restricted genes" or "TRGs" (reviewed in Khalturin et al. 2009). This term implied that the origin of many genes in extant organisms occurred later than at the time of the emergence of the unicellular ancestor and that their origination at later times could be linked to the emergence of lineage specific adaptations. This implies that studying systematically the time horizons at which genes emerged could provide interesting insights into the molecular basis of evolutionary innovations a procedure that we called "phylostratigraphy" (Domazet-Loso et al. 2007).

The discovery of *de novo* gene evolution

The increasing realization that orphan genes are a true class of genes that had no recognizable parental homologs did initially not result in challenging the prevailing dogma that new genes could only arise out of existing genes. Beginning in the 1990s, several research groups had started to specifically explore the origin of new genes. First results in *Drosophila* showed that new gene functions could indeed come from piecing together parts of existing genes (Long et al. 1993) and it was also realized that some genes showed particularly high substitution rates, when compared between closely related lineages (Schmid and Tautz 1997). This implied that a model of duplication (or piecing together) followed by a phase of fast evolution could indeed lead to a loss of similarity to parental genes and thus explain the occurrence of orphans (Domazet and Tautz 2003).

However, with increasing comparative genome information available in *Drosophila*, cases were discovered that suggested that genes could also emerge *de novo* out of non-coding DNA (Levine et al. 2006, Begun et al. 2007, Chen et al. 2007). This was initially considered an oddity, but paved the way of how to study the question of *de novo* evolution - namely by doing systematic genomic and transcriptional comparisons among multiple closely related lineages (Tautz et al. 2013). Toll-Riera et al. (2009) conducted the first comparative analysis that allowed all possible models of gene evolution, including *de novo* evolution. They studied orphan genes in primates and concluded that only a quarter could be associated to highly diverged members of known protein families (i.e. are products of a duplication divergence mechanism). Using very stringent exclusion criteria, they found that 5.5.% of the genes studied had to be considered as true *de novo* genes, emerging out of intergenic DNA and not even containing fragments of other genes or transposable elements. In retrospect, this was an underestimate, since the authors had to use very stringent criteria, since there was no fully documented and unquestionable case of *de novo* evolution of a functional gene at the time where they did their analysis. However, the first such demonstration was published in the same year. Heinen et al. (2009) had studied a transcript in the mouse that showed signs of recent positive selection. They showed that it had emerged out of intergenic DNA through a few mutational steps that had occurred in the mouse lineage, but not in the outgroups. Knockout of the gene revealed that the RNA had already acquired a function in contributing to sperm motility, although it was not possible to show that it coded for a protein. The first demonstration of a functionally important protein to have arisen out of a non-coding RNA transcript came later from yeast (Li et al. 2010).

In parallel, comparative bioinformatic analysis, combined with data from protein mass-spectrometry showed that several protein coding genes in humans had apparently arisen *de novo* in the primate lineage (Knowles and McLysaght 2009). Commenting on this finding, Siepel (2009) pointed out that *de novo* evolution should now be considered as a realistic concept and introduced the term "proto-gene" as an intermediate stage where an arising transcript behaves initially neutrally until positive selection takes on to give it a functional role. A re-analysis of data on RNAs associated with ribosomes in yeast showed then that *de novo* evolved proto-genes could even be found in the otherwise extremely well annotated yeast genome (Wilson and Masel, 2011; Carvunis et al., 2012). Based on their comprehensive analysis, Carvunis et al. (2012) were even able to conclude: "*We identify 1,900 candidate proto-genes among S. cerevisiae ORFs and find that de novo gene birth from such a reservoir may be more prevalent than sporadic gene duplication.*" There is now a rapidly increasing number of papers that show evidence for *de novo* evolution in many lineages, which can not all be cited here. The two most recent additions are a study by Zhao et al. (2014) that shows that *de novo* gene evolution can already be traced at the population level in *Drosophila* and a study by Palmieri et al (2014) showing a fast birth and death processes of orphan genes in *Drosophila* lineages.

Why not earlier?

Now, with the long lasting dogma being broken, we can ask why did it prevail for such a long time and why should it have been so difficult to even consider the relatively simple alternative of *de novo* evolution? In fact, one can find traces in the literature that show that both the idea, as well as supporting data had existed since a long time.

As discussed above, Stephens mentioned already in 1951 *de novo* evolution as an alternative, but felt that it was too difficult to find any proof for it. This was certainly a fair judgment, but many of the unsolved questions at that time had inspired experimentalists to provide the proof for alternative models. Still, at least from the published literature, it is not evident that *de novo* evolution was even considered as an alternative for a long time. It is certainly difficult to find any hint for it in the earlier writings of Ohno, who was otherwise very well prepared to discuss any evidence from various perspectives. Intriguingly, however, it was Ohno himself who published the first seemingly unequivocal evidence for a *de novo* evolved protein (Ohno 1984). He found it by analyzing a protein in Flavobacteria with the capacity to degrade nylon. Given that nylon is a molecule produced by humans, not previously present in the environment, he argued that this protein would be an evolutionary innovation. It turned out that this protein is coded within a previously existing gene, but in an alternative reading frame. Alternative reading frames are almost like random sequences and their corresponding protein products have nothing to do with each other. Hence, finding a functional protein in an alternative reading frame of an existing protein is surely the best possible proof for the effectiveness of *de novo* evolution, since it is clear that any form of gene duplication could not have been involved in it. Hence, from this time onwards, there would have been an experimental route towards proofing *de novo* emergence of genes throughout evolution, but very little was made out of this. Ohno himself was not specifically interested in this aspect. Instead, he focused on the internal repetitious nature of the protein and developed ideas that centered again around primordial evolution (Ohno 1987) [note from today´ perspective: the premises and conclusions of his 1984 paper can not be held up anymore - the assumed *de novo* evolved protein is in fact a ß-lactamase, i.e. a member of an ancient gene family. It is the potential product of the presumed original reading frame that has no similarity to other proteins (Andrei Lupas, MPI Tübingen, pers. comm.)].

The real relevance of the use of an alternative reading frame for the question of gene evolution was only recognized by Keese and Gibbs (1992). They analyzed several genes with double reading frames in viruses and showed that one of the reading frames belonged to an old class of genes, while the other was specific for a given lineage, i.e. must have evolved *de novo*. They called this mechanism "overprinting" and the title of their paper "*Origins of genes: big bang or continuous creation?*" was clearly chosen to attract attention to the conceptual problem. However, its impact was relatively small (cited only a little more than 100 times up to today), possibly because it dealt mostly with viral genes. But a later study focusing on mammalian overprinted genes (Chung et al. 2007) did not have much impact either, although the authors concluded that many more such cases will be discovered. While the literature on overprinted genes in viruses started to flourish (Sabath et al., 2012), it remains an understudied subject in eukaryotic genomes, although it is relatively easy to find candidate loci for overprinting in the databases (Michel et al. 2012; Neme and Tautz, 2013).

The scientists working on the annotation of genomes had of course realized that many transcripts existed that did not code for conserved proteins. However, instead of taking this as a challenge to reconsider models of gene evolution, concepts and analyses were developed that allowed to discuss these away (Clamp et al. 2007). When Wu et al. (2011) eventually published a compelling list of *de novo* evolved proteins in humans, Guerzoni and McLysaght (2011) noted in a comment that many of these were annotated in an earlier versions of the genome sequence, but were removed at later stages by the annotators, most likely because the prevailing dogma had no room for such genes.

Hence, proof for *de novo* evolution from intergenic sequences had to await the increasing evidence from comparative genomic data from closely related species. These have become available only during the last few years, but this has finally brought the issue to the forefront. The 2013 conference of the Society for Molecular Biology and Evolution in Chicago was the first one that had finally a dedicated session on *de novo* evolution and there is little doubt that it will remain a topical issue in the coming years.

In asking the question why it took so long to realize that the alternative model of *de novo* evolution could be viable, I may add my own experience, which may be typical for others in the field. In our 2003 paper on the evolutionary analysis of orphan genes in *Drosophila*, we discussed *de novo* evolution as an option, but dismissed it with specific reference to the notion that all protein domains must have arisen early on in evolution. But we were at that time also not aware of the Keese and Gibbs (1992) paper and the evidence for overprinting, which could have changed our assessment. Instead, we opted for the canonical duplication-divergence model, which remains a viable option, but not the only one. Also, our work on the *Pldi* gene, which became the first functionally studied *de novo* evolved gene (Heinen et al. 2009), was initially not motivated to find evidence for *de novo* evolution, but by trying to understand the function of an orphan gene that evolved fast between closely related species. That *de novo* evolution was involved became only clear after we had collected a large amount of comparative data. Even then it remained a single case with unclear generality, but it was now at least clear how one would have to search systematically for similar cases. It was only during writing a review on orphan evolution (Tautz and Domazet-Loso, 2011) where we started to combine the increasing evidence and had a new look on the data. My co-author, Tomislav Domazet-Loso produced a variant of his phylostratigraphy by plotting emergence rates of new genes over time for three major evolutionary lineages (plants, insect and mammals). Each of them showed consistently a very high emergence of new genes in the youngest lineage, which could not be explained by gene duplication mechanisms. This implied most directly that *de novo* evolution must be a very powerful mechanism for generating new genes and that there must be a continuous birth and death process of such genes. Most interestingly, the above mentioned study by Palmieri et al. (2014) provides now direct evidence for this birth and death model (Neme and Tautz 2014).

Consequences for our understanding of the gene concept

Accepting the reality of frequent *de novo* evolution of genes has several practical consequences for our understanding of genes.

1) Functional novelty can arise out of randomness. This implies that we will have to re-think what a functional protein space is. It is generally clear that proteins do not have to be folded to be functional (Dyson and Wright 2005), but it will also be important to revisit the question whether folds can evolve convergently.

2) There is a continuum in gene emergence from neutrally expressed transcripts over non-coding, but functional RNA transcripts, to protein coding genes (Carvunis et al. 2012; Tautz et al. 2013). In addition, even very short reading frames have been shown to be functional in some cases (Tautz 2009). Hence, the frequent practice to consider a gene only when it has a minimal open reading frame needs to be re-considered.

3) The duplication-divergence concept for orphan gene emergence needs to be revisited. Although there are a few well studied cases where high structural similarity exists in the virtual absence of sequence similarity, it will need to be more systematically assessed whether fast protein evolution can indeed effectively conceal evolutionary ancestry, or whether this is rare.

But apart of these insights into the nature of genes, the discovery process of *de novo* evolution may also serve as an excellent example of how concepts shape our understanding of biology - and how much they can be in the way when it comes to seeing the full picture.

Acknowledgements

I thank Rafik Neme for pointing out papers on the historical background, as well as him and Tomislav Domazet-Loso for frequent discussion of the ideas worked out in this article and Andrei Lupas for reanalysing data. My research is funded by institutional funds of the Max-Planck Society.

Cited Literature

Beadle GW (1945) BIOCHEMICAL GENETICS. *Chemical Reviews* **37**, 15-96.

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.

Bridges CB (1935) Salivary chromosome maps with a key to the banding of the chromosomes of Drosophila melanogaster. *Journal of Heredity* **26**, 60-64.

Britten RJ, Davidson EH (1969) GENE REGULATION FOR HIGHER CELLS - A THEORY. *Science* **165**, 349-&.

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

Chen ST, Cheng HC, Barbash DA, Yang HP (2007) Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in Drosophila melanogaster. *Plos Genetics* **3**, 1131-1143.

Chothia C (1991). One thousand families for the molecular biologist. Nature **357**, 543-544.

Chung WY, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A (2007) A first look at ARFome: Dual-coding genes in mammalian Genomes. *Plos Computational Biology* **3**, 855-861.

Clamp M, Fry B, Kamal M, *et al.* (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 19428-19433.

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

Domazet-Loso T, Tautz D (2003) An evolutionary analysis of orphan genes in Drosophila. *Genome Research* **13**, 2213-2219.

Duboule D, Wilkins AS (1998) The evolution of 'bricolage'. *Trends in Genetics* **14**, 54-59.

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

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

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

Guerzoni D, McLysaght A (2011) De Novo Origins of Human Genes. *Plos Genetics* **7**.

Gulick A (1944) The chemical formulation of gene structure and gene action. *Advances in Enzymology and Related Subjects of Biochemistry* **4**, 1-39.

Heinen T, Staubach F, Haming D, Tautz D (2009) Emergence of a New Gene from an Intergenic Region. *Current Biology* **19**, 1527-1531.

Jacob F (1977) EVOLUTION AND TINKERING. *Science* **196**, 1161-1166.

Keese PK, Gibbs A (1992) ORIGINS OF GENES - BIG-BANG OR CONTINUOUS CREATION. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 9489-9493.

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

King MC, Wilson AC (1975) EVOLUTION AT 2 LEVELS IN HUMANS AND CHIMPANZEES. *Science* **188**, 107-116.

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

Kuwada Y (1911) Meiosis in the pollen mother cells of *Zea Mays* L. *Bot. Mag.* **25**, 163

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

Li D, Dong Y, Jiang Y, *et al.* (2010) A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Research* **20**, 408-420.

Long MY, Langley CH (1993) NATURAL-SELECTION AND THE ORIGIN OF JINGWEI, A

CHIMERIC PROCESSED FUNCTIONAL GENE IN DROSOPHILA. *Science* **260**, 91-95.

Lupas AN, Ponting CP, Russell RB (2001) 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? *Journal of Structural Biology* **134**, 191-203.

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

Mirsky AE, Ris H (1951) THE DESOXYRIBONUCLEIC ACID CONTENT OF ANIMAL CELLS AND ITS EVOLUTIONARY SIGNIFICANCE. *Journal of General Physiology* **34**, 451-462.

Neme R, Tautz D (2013) Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**.

Neme R, Tautz D (2014) Evolution: Dynamics of de novo gene emergence. *Curr Biol* 24, R 238-240.

Ohno S (1970) Evolution by Gene Duplication. NewYork: Springer

Ohno S, Wolf U, Atkin NB (1968) EVOLUTION FROM FISH TO MAMMALS BY GENE DUPLICATION. *Hereditas-Genetiskt Arkiv* **59**, 169-&.

Ohno S (1984) BIRTH OF A UNIQUE ENZYME FROM AN ALTERNATIVE READING FRAME OF THE PREEXISTED, INTERNALLY REPETITIOUS CODING SEQUENCE. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **81**, 2421-2425.

Ohno S (1987) EVOLUTION FROM PRIMORDIAL OLIGOMERIC REPEATS TO MODERN CODING SEQUENCES. *Journal of Molecular Evolution* **25**, 325-329.

Oliver SG, Vanderaart QJM, Agostonicarbone ML, *et al.* (1992) THE COMPLETE DNA-SEQUENCE OF YEAST CHROMOSOME-III. *Nature* **357**, 38-46.

Orengo CA, Thornton JM (2005) Protein families and their evolution - A structural perspective. *Annual Review of Biochemistry* **74***,* 867-900.

Palmieri N, Kosiol C, Schlötterer C (2014). The life cycle of Drosophila orphan genes. *eLife* (in press).

Sabath N, Wagner A, Karlin D (2012) Evolution of Viral Proteins Originated De Novo by Overprinting. *Molecular Biology and Evolution* **29**, 3767-3780.

Schmid KJ, Tautz D (1997) A screen for fast evolving genes from Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9746-9750.

Serebrovsky AS (1938). Genes *scute* and *achaete* in *Drosophila melanogaster* and a hypothesis of gene divergency. *C. R. Acad. Sci. URSS* **19**, 77–81.

Siepel A (2009) Darwinian alchemy: Human genes from noncoding DNA. *Genome Research* **19**, 1693-1695.

Stephens SG (1951) POSSIBLE SIGNIFICANCE OF DUPLICATION IN EVOLUTION. *Advances in Genetics Incorporating Molecular Genetic Medicine* **4**, 247-265.

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

Tautz D, Domazet-Loso T (2011) The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**, 692-702.

Tautz, D., Neme, R. and Domazet-Lošo, T. 2013. Evolutionary Origin of Orphan Genes. eLS. DOI: 10.1002/9780470015902.a0024601

Taylor JS, Raes J (2004) Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics* **38**, 615-643.

Tischler G (1915). Chromosomenzahl, Form und Individualit im Planzenreiche. *Progr. Rei Bot.* **5**, 164

Toll-Riera M, Bosch N, Bellora N, *et al.* (2009) Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Molecular Biology and Evolution* **26**, 603-612.

Wilson BA, Masel J (2011) Putatively Noncoding Transcripts Show Extensive Association with Ribosomes. *Genome Biology and Evolution* **3**, 1245-1252.

Wu DD, Irwin DM, Zhang YP (2011) De Novo Origin of Human Protein-Coding Genes. *Plos Genetics* **7**.

Zhao L, Saelao P, Jones CD, Begun DJ (2014). Origin and spread of de novo genes in Drosophila melanogaster populations," *Science*, doi:10.1126/science.1248286, 2014.