Alternative reading frames are an underappreciated source of protein sequence novelty

Zachary Ardern*
Wellcome Sanger Institute, Cambridge

*zachary.ardern@sanger.ac.uk

Abstract:

Alternative reading frames of protein coding genes are a major contributor to the evolution of novel protein products. Recent studies demonstrating this include examples across the three domains of cellular life and in viruses. Alternative frame sequences both increase the number of trials available for the evolutionary invention of new genes and have unusual properties which may facilitate gene origin. The structure of the standard genetic code contributes to the features and gene-likeness of some alternative frame sequences. These findings have important implications across diverse areas of molecular biology, including for genome annotation, structural biology, and evolutionary genomics.

Introduction

Is evolution a tinkerer, never creating from scratch, as was proposed 45 years ago? (Jacob 1977). This claim both reflected and helped to form a general consensus across the subdisciplines of molecular biology which has been orthodoxy until recently. The growing evidence of highly taxonomically restricted or “orphan” genes formed ‘de novo’ from non-coding sequences however challenges this consensus (Van Oss and Carvunis 2019; Vakirlis, Carvunis, and McLysaght 2020; Weisman 2022). The origin of many new functional genes requires some combination or three factors, as summarised in a recent overview (Weisman 2022): that novel gene sequences are not as rare as earlier assumed, that there are vast numbers of evolutionary trials, or that many of the sequences trialled are non-random in ways which facilitate gene origin. I argue that all three factors likely contribute. In particular, examining the large resource of alternative frame sequences within protein-coding genes suggests both that genomes are involved in more evolutionary trials of potential novel genes than previously realised and that these sequences have useful biases.

The non-coding sequences which are the raw materials of gene birth often have unusual sequence properties which may predispose them towards gene formation (Wilson et al. 2017; Schmitz, Ullrich, and Bornberg-Bauer 2018; Vakirlis et al. 2018; Willis and Masel 2018; L. J. Kosinski and Masel 2020; L. Kosinski et al. 2022). For instance, open reading frames (ORFs) which give rise to proteins in budding yeast are enriched in foldable peptides and have reduced aggregation tendency (Papadopoulos et al. 2021). Another study found that such ORFs tend to encode transmembrane domains (Vakirlis et al. 2020). The specific “pre-adaptation” hypothesis is one of multiple models for de novo gene origin (Van Oss and Carvunis 2019) which proposes particular selective pressures which create sequence biases; the existence of some useful biases in progenitor sequences is a broader claim.

While key details (e.g. regarding foldability vs disorder) remain to be worked out across organisms, the hypothesis that some sequence biases facilitate gene birth would support a key intuition underlying the “tinkering only” framework, that functional protein sequences could only very rarely arise from truly random non-coding sequences. As discussed below, many new protein sequences are at least partially derived from alternative reading frames of existing genes, and these sequences are non-random. This largely unexplored reservoir both greatly increases the total number of evolutionary trials from which gene novelty can arise, and is a source of potentially useful bias in novel sequences.

De novo origins

Proteins have significant functional and biophysical constraints concerning which mutations can be sustained while retaining structural integrity and/or function. Most random changes in a protein sequence tend to be destabilising (Tokuriki et al. 2007). In the early years of molecular biology the observed specificity led to discussion of the “uniqueness of the gene” (Salisbury 1969; J. M. Smith 1970). We now know that any given protein fold can be encoded by vast arrays of different sequences of amino acids, sometimes referred to as the “sequence capacity” or “designability” of a protein fold (Tian and Best 2017; Pan et al. 2021). The sequence specificity however is determined by the ratio of the sequence capacity to the total number of possible protein sequences of the typical sequence length for the fold. The total possibility space is hyperastronomical (Louis 2016), and as such, the specificity ratio may be extremely small for particular protein folds (Axe 2004; Tian and Best 2017). The nature of the sequence specificity of proteins is in general not well understood; for instance, protein sequences appear to be surprisingly close to randomly distributed throughout sequence space (Weidmann et al. 2021). The structure space of both young proteins (Bornberg-Bauer, Hlouchova, and Lange 2021) and small proteins (Kubatova et al. 2020) remains largely unexplored. While many proteins do appear to be highly specified the overall picture is not yet clear, for instance different folds sometimes perform the same functional role (Bork, Sander, and Valencia 1993) and it is not known to what extent the total sequence space includes functional proteins, including folds unsampled in extant biology. Recent advances in computational structure prediction (Jumper et al. 2021; Weissenow, Heinzinger, and Rost 2022; Lin et al. 2022) will greatly improve our knowledge of at least the naturally occurring protein structures.

If proteins have extremely high sequence specificity then de novo birth is expected to be rare, with protein evolution proceeding only through minor divergence from existing sequences. This was the argument of some of the founders of modern molecular evolution such as Maynard Smith (1970), Ohno (1970), and Jacob (1977). Many proteins do indeed have ancient roots, for instance approximately 150 million protein domains are able to be sorted into less than 5500 protein superfamilies in the CATH database (Sillitoe et al. 2021). Each superfamily comprises divergent proteins probably sharing a common ancestor. Some of these different superfamilies are also related by descent (Cheng et al. 2014), and many apparently distinct domains share common short motifs termed ‘themes’, likely via descent (Kolodny et al. 2021). We might conclude from this that most protein families originated before or near the time of the last universal common ancestor. A view like this is indeed sometimes still advocated (Bordin et al. 2021; Weidmann et al. 2021). Even many putative orphan genes without detected homologs do have homologous sequences which are missed

by simple similarity searches due to rapid evolution or differences in genome annotation across species (Arendsee et al. 2019; Weisman, Murray, and Eddy 2022). However, such factors do not account for all orphans (Weisman, Murray, and Eddy 2020), and databases like CATH, by focusing on domains, are inherently biased towards well-conserved protein sequences, so likely underestimate the number of distinct protein families. Many protein-coding genes are genuinely highly taxonomically restricted, and a few have been shown in detail to have originated relatively recently from non-coding sequences (Schmitz and Bornberg-Bauer 2017; Van Oss and Carvunis 2019; Vakirlis, Carvunis, and McLysaght 2020; Weisman 2022).

Both the ubiquity of orphan genes and proposed models for de novo gene origin have been controversial (Casola 2018; Weisman, Murray, and Eddy 2020). Generalisations about gene origins are typically drawn from studies in single model organisms (usually yeast, mice, fruit flies, or humans), and often results differ when using different organisms, computational approaches or similarity thresholds. As an example of the complexity of this literature, while many examples of de novo gene origin in the fruit fly *Drosophila melanogaster* have been claimed (Levine et al. 2006; Heames, Schmitz, and Bornberg-Bauer 2020), a detailed analysis found just one example arising within the *D. melanogaster* species subgroup which passed stringent thresholds (Zile et al. 2020). A recently published analysis, in contrast, included not only annotated genes but other translated ORFs and found more than 80 de novo gene candidates in this same clade, more than any previous analysis (Zheng and Zhao 2022).

Frameshifted sequences

Biological utilisation of frameshifted sequences has been known since at least the discovery of same-strand overlapping genes in bacteriophages in the mid-1970s (Barrell, Air, and Hutchison 1976). The concept of out-of-phase coding had been discussed earlier, for instance in the context of a manipulated bacteriophage (Contreras et al. 1973). Undergirding the phenomenon of gene overlap is the triplet standard genetic code which uses double-stranded DNA and encodes 20 amino acids among 64 codons. The redundancy of the code allows for significant sequence flexibility at the nucleotide level when encoding a given amino acid sequence. The interchangeability of many amino acids further increases coding flexibility. The triplet nature of the code means that translational frameshifts will produce different protein sequences. The double-stranded nature of DNA further allows for reading protein sequences from either strand or both simultaneously. Both sense and antisense overlapping gene pairs are the subject of an important recent review (Wright, Molloy, and Jaschke 2022). Virus overlapping genes have been discussed in some detail in relation to gene birth (Rancurel et al. 2009; Willis and Masel 2018).

A shift in reading frame may be expected to be highly deleterious. Consider morse code, for instance - it would be surprising if after a message was shifted by one character per morse “codon” (letter code) it was still meaningful. This intuition has been influential; for instance, in an important paper by John Maynard Smith (1970), frameshifts were effectively equated with random sequences. There is also evidence that the genetic code has been optimised to some extent for producing a stop codon as soon as possible after a frameshift, to reduce the damage from frameshifting (Itzkovitz and Alon 2007). Proteins are nonetheless surprisingly

robust to frameshift mutations (Coray et al. 2019). Frameshifted sequences tend to conserve the hydrophobicity profiles of unshifted reference frame sequences (Bartonek, Braun, and Zagrovic 2020) and perhaps biochemical similarity more generally, as summarised in amino acid scoring matrices (Wang et al. 2022). A careful analysis showed that the high frameshift robustness in the standard genetic code compared to alternative codes with a similar degeneracy structure is largely a consequence of the well-documented mismatch robustness in combination with the 'block' structure (Xu and Zhang 2021). Regardless of how it was achieved, the frameshift robustness measured in terms of e.g. polar requirement is a real feature of coding sequences.

Frameshifted sequences are used by diverse organisms. In bacteria, "recoding" of putative pseudogenes involving frameshifting is common, as investigated in a recent in-depth analysis of two serovars of *Salmonella enterica* (Feng et al. 2022). Single-strand RNA bacteriophages regularly evolve sgl (single gene lysis) genes out-of-frame to the phage's core genes (Chamakura et al. 2020). Coronaviruses including SARS-CoV-2 include multiple same-strand overlapping genes (Nelson et al. 2020; Firth 2020; Jungreis et al. 2021; Stewart et al. 2022). In vertebrates an early comparative genomics study showed that frameshifts are commonly retained for millions of years and facilitate the exploration of new sequence space (Raes and Van de Peer 2005). There are a few known examples of out-of-frame ORFs in the human genome (Wright et al. 2022) - a recent study, as reported in supplementary tables, found 98 same strand out-of-frame ORFs with some evidence of protein coding from ribosome profiling experiments (Mudge et al. 2022).

Antisense coding

Coding from each of the three frames antisense to an annotated protein-coding gene has been established across diverse biological systems. In HIV the antisense protein, asp (Cassan et al. 2016; Miller 1988), has been shown to be a structural protein in the viral envelope and a transmembrane protein involved in infection of host cells (Affram et al. 2019). In bacteria, multiple antisense proteins have been characterised (Delaye et al. 2008; Fellner et al. 2014; Hücker et al. 2018; Vanderhaeghen et al. 2018; Zehentner et al. 2020a; Kreitmeier et al. 2022), as has been reviewed (Ardern, Neuhaus, and Scherer 2020). The previously mentioned single confirmed 'de novo' gene in the *D. melanogaster* species group (Zile et al. 2020) arose in antisense to an open reading frame which may be protein coding. Multiple novel antisense protein-coding ORFs have been reported in S. cerevisiae (Blevins et al. 2021). It is apparent that many more antisense coding ORFs are likely already in the extant literature, either in supplementary tables or excluded as candidates at early stages of analysis on account of substantial overlaps. While most reported antisense genes are in frames "-1" or "-2" (i.e. directly antisense or antisense and shifted one nucleotide to the left), the relative proportions haven't been rigorously established. The potential functionality of antisense sequences has long been discussed (Forsdyke 1995), but only relatively recently has been established as a significant phenomenon.

The structure of the standard genetic code is arranged in a way that may facilitate the origin of functional protein sequences in antisense. Codons for hydrophobic amino acids tend to be complemented by codons for hydrophilic amino acids in the antisense (Blalock and Smith 1984). This entails that the hydrophobicity pattern in the sense strand is templated in mirror

form in the antisense. Protein hydrophobicity is very important for protein structure formation. This suggests an analogy with a method used in designing novel proteins, where combinatorial libraries are created with a defined pattern of polar and non-polar amino acids, increasing the chance of the sequence forming a folded protein (Hecht et al. 2004). Whether antisense sequences really are apt for structure formation deserves further attention. It has similarly been suggested that there is a tendency for secondary structures found in the sense strand to be retained in antisense, in terms of the types of amino acids encoded (Zull and Smith 1990) and a tendency for conservative mutations in one strand to also be conservative in the other (Konecny et al. 1993). This study of conservation in antisense has recently been extended to other frames, finding remarkable similarity across frames (Wichmann and Ardern 2019, 2022).

Mechanisms of protein novelty from alternative reading frames

There are multiple possible mechanisms whereby alternative reading frames can play a role in gene origins - I will briefly describe three (Figure 1). The first is the best known - overlapping genes. The extent of overlapping genes is not yet well quantified and it seems likely that many overlapping genes remain undiscovered across diverse taxa. Overlapping genes were proposed as contributors to gene novelty 30 years ago (Keese and Gibbs 1992). The origin of such gene pairs is termed “overprinting”, and can occur either through the translation of a previously non-coding open reading frame in an alternative reading frame of an existing protein-coding gene, on the same or opposite strand. It is possible that such gene pairs may get copied with one member subsequently lost by pseudogenization, but no such examples have been published to my knowledge. Excitingly, recent studies have demonstrated and proposed other mechanisms by which alternative reading frames play a role in gene origins. Secondly, frameshifts within an existing gene are sometimes retained after compensatory mutations. This phenomenon of “pairs of compensatory mutations” has recently been studied in insects and vertebrates, with a few strong candidates found including three human genes (Biba, Klink, and Bazykin 2022). The general phenomenon of incorporation of frameshifted sequences was studied earlier (Raes and Van de Peer 2005). A third mechanism is frame-shifted gene fusion. Fusions including one or more frame-shifted sequences play a non-trivial role in the origin of new genes in *E. coli*; an estimated ~2.5% of all genes in the species include sequences with an out-of-frame origin (Watson, Lopez, and Bapteste 2022). A fourth possible mechanism is stop codon readthrough, either at the ends of genes (L. J. Kosinski and Masel 2020) or following premature truncation (Feng et al. 2022), leading to the translation of out-of-frame sequences. Other hypotheses have also been proposed. For instance, it has been suggested that frameshifting at the level of translation may produce mosaic peptides in eukaryotes (Çakır et al. 2021).

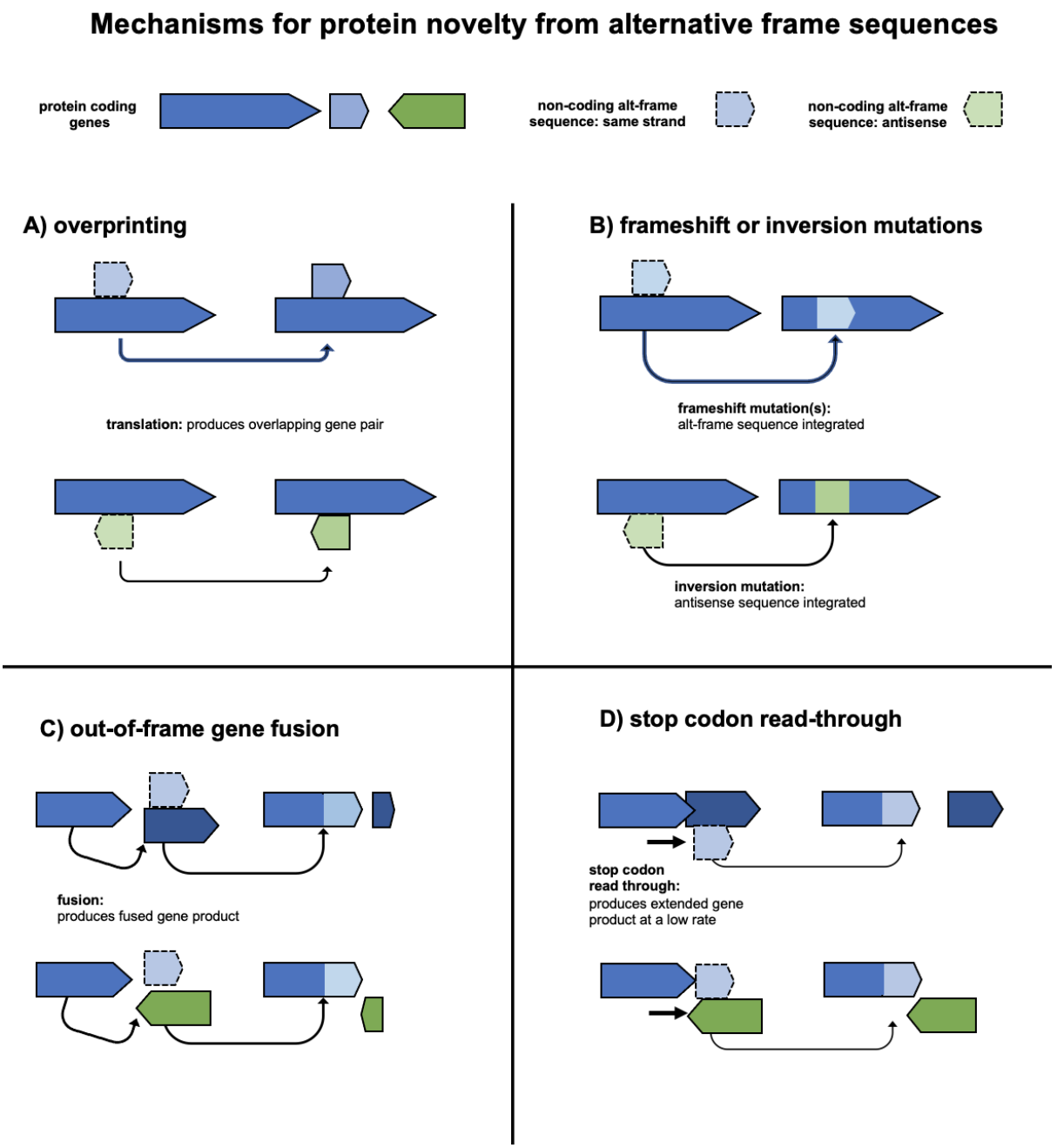

Figure 1: mechanisms for producing protein novelty via alternative frame sequences: 1) overprinting, i.e. translation of out-of-frame sense or antisense sequence produces an overlapping gene pair, or similarly a merged product via programmed ribosomal frameshift. 2) frameshift mutation (potentially with a compensatory frameshift downstream) or inversion incorporates alternative frame sequence 3) fusion of genes, with one or more being out-of-frame creates a fused gene product 4) stop codon read through produces an extended gene product, typically only for a percentage of the protein products.

Sequence novelty is a broader concept than gene novelty. As genes are often composed of multiple domains which can be added, subtracted, or replaced, what counts as a new gene resembles a “ship of theseus” problem (“Plutarch, Theseus, Chapter 23, Section 1” n.d.). Of

the four mechanisms of sequence novelty from alternative reading frames highlighted here, i.e. overprinting, fusion, frameshift-compensation, and read-through (Figure 1), the latter three typically do not produce fully new genes except in cases where two genes fuse such that both are frame-shifted), but instead introduce new amino acid sequences into existing genes. Overprinting can also involve extension of existing genes rather than translation of separate open reading frames. Extension or replacement of sequence allows skipping some of the steps involved in forming a fully new gene, which are summarised by Van Oss and Carvunis (2019). However, there is a large pool of translated sequences in addition to canonical genes, giving more templates from which to build. For instance, in both eubacteria and archaea (Gelsinger et al. 2020; Zehentner et al. 2020b; C. Smith et al. 2022), budding yeast (Ingolia et al. 2014; Durand et al. 2019; Blevins et al. 2021), mice (Ruiz-Orera et al. 2018), and the human genome (Mudge et al. 2022) there are many unannotated open reading frames with strong evidence of protein coding but which are generally not under strong purifying selection. Many of these sequences are in alternative frames of older protein-coding genes. Only very seldom in such studies however is purifying selection tested for with methods appropriate for overlapping sequences, of which there are now a few (Firth 2014; Sealfon et al. 2015; Nelson, Ardern, and Wei 2020). As such, the evolution of alternative frame sequences outside of viruses is almost completely unstudied. It is clear that they evolve rapidly (Pavesi 2019), and as such explore large regions of sequence space. Even synonymous mutations in a coding sequence in effect explore sequence space in alternative frames. The surprising similarity in this “collateral” effect size across alternative frames (Wichmann and Ardern 2019) may facilitate evolutionary exploration (Wichmann and Ardern 2022).

Discussion

At the turn of the millennium a helpful critical review of the evidence available at the time for direct sense-antisense coding and its potential role in evolution was published (Boldogköi 2000). The author concluded by favouring the traditional view that proteins arise only by divergence, writing about antisense coding “I do not believe that this capacity, if it ever existed, is utilized by recent genomes.” Bidirectional coding has, similarly, recently been argued to have been important in early life and rare since (Carter 2021). However, the extensive recently accrued evidence for de novo gene origin and out-of-frame coding brings alternative reading frames out of the shadows as a demonstrated mechanism of continued importance in gene origin.

Many research questions are opened up as a result of the increasing prominence of alternative frame protein-coding sequences. In particular, both the relative and absolute contributions of the different mechanisms discussed here, as well as the timeframes on which they operate, are unknown. The structural features of proteins encoded by overlapping genes and novel genes more broadly can now be investigated with new computational methods, along with the contribution of alt-ORF sequences to protein structure. In future research six frame translations should be increasingly used when searching for homologous sequences rather than relying on official annotations, and the gaps in current gene annotations should be taken into account. Improved comprehensive genome annotation will lead to improved understanding of biological processes including critical host-pathogen interactions (Stewart et al. 2022). Understanding evolutionary rates in

novel genes will help in better understanding accessory genes, for instance in both viruses and the large pangenomes of many bacterial species (Brockhurst et al. 2019). The rapid evolution of alternative frame sequences in particular increases the number of evolutionary trials and may facilitate gene origins from these sequences.

Processes underlying the origins of novel proteins remain poorly understood at the level of detailed mechanisms. The evolutionary use of overlapping reading frames potentially provides an elegant mechanism for exploring new territory in sequence and function-space - by both increasing the number of trials and potentially being biased in useful directions. Although the available data is of mixed quality and is often contentious, there is growing evidence that alternative frame ORFs play a major role in gene origin and that the structure of the proteins they encode may be partially templated from pre-existing reference frame proteins. The discovery of this vast coding reservoir contributes towards solving the apparent paradox of protein sequence specificity and the frequent evolution of protein novelty.

Acknowledgements
Thanks to Md. Hassan uz-Zaman for helpful comments on the manuscript.

References

Affram, Yvonne, Juan C. Zapata, Zahra Gholizadeh, William D. Tolbert, Wei Zhou, Maria D. Iglesias-Ussel, Marzena Pazgier, Krishanu Ray, Olga S. Latinovic, and Fabio Romerio. 2019. “The HIV-1 Antisense Protein ASP Is a Transmembrane Protein of the Cell Surface and an Integral Protein of the Viral Envelope.” *Journal of Virology* 93 (21). https://doi.org/10.1128/JVI.00574-19.

Ardern, Zachary, Klaus Neuhaus, and Siegfried Scherer. 2020. “Are Antisense Proteins in Prokaryotes Functional?” *Frontiers in Molecular Biosciences* 7 (August): 187.

Arendsee, Zebulun, Jing Li, Urminder Singh, Priyanka Bhandary, Arun Seetharam, and Eve Syrkin Wurtele. 2019. “Fagin: Synteny-Based Phylostratigraphy and Finer Classification of Young Genes.” *BMC Bioinformatics* 20 (1): 440.

Axe, Douglas D. 2004. “Estimating the Prevalence of Protein Sequences Adopting Functional Enzyme Folds.” *Journal of Molecular Biology* 341 (5): 1295–1315.

Barrell, B. G., G. M. Air, and C. A. Hutchison. 1976. “Overlapping Genes in Bacteriophage φX174.” *Nature* 264 (5581): 34–41.

Bartonek, Lukas, Daniel Braun, and Bojan Zagrovic. 2020. “Frameshifting Preserves Key Physicochemical Properties of Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (11): 5907–12.

Biba, Dmitry, Galya Klink, and Georgii Bazykin. 2022. “Pairs of Mutually Compensatory Frameshifting Mutations Contribute to Protein Evolution.” *Molecular Biology and Evolution*, February. https://doi.org/10.1093/molbev/msac031.

Blalock, J. E., and E. M. Smith. 1984. “Hydropathic Anti-Complementarity of Amino Acids Based on the Genetic Code.” *Biochemical and Biophysical Research Communications* 121 (1): 203–7.

Blevins, William R., Jorge Ruiz-Orera, Xavier Messeguer, Bernat Blasco-Moreno, José Luis Villanueva-Cañas, Lorena Espinar, Juana Díez, Lucas B. Carey, and M. Mar Albà. 2021. “Uncovering de Novo Gene Birth in Yeast Using Deep Transcriptomics.” *Nature Communications* 12 (1): 604.

Boldogköi, Z. 2000. “Coding in the Noncoding DNA Strand: A Novel Mechanism of Gene Evolution?” *Journal of Molecular Evolution* 51 (6): 600–606.

Bordin, Nicola, Ian Sillitoe, Jonathan G. Lees, and Christine Orengo. 2021. “Tracing

Evolution Through Protein Structures: Nature Captured in a Few Thousand Folds." *Frontiers in Molecular Biosciences* 8 (May): 668184.
Bork, P., C. Sander, and A. Valencia. 1993. "Convergent Evolution of Similar Enzymatic Function on Different Protein Folds: The Hexokinase, Ribokinase, and Galactokinase Families of Sugar Kinases." *Protein Science: A Publication of the Protein Society* 2 (1): 31–40.
Bornberg-Bauer, Erich, Klara Hlouchova, and Andreas Lange. 2021. "Structure and Function of Naturally Evolved de Novo Proteins." *Current Opinion in Structural Biology* 68 (June): 175–83.
Brockhurst, Michael A., Ellie Harrison, James P. J. Hall, Thomas Richards, Alan McNally, and Craig MacLean. 2019. "The Ecology and Evolution of Pangenomes." *Current Biology: CB* 29 (20): R1094–1103.
Çakır, Umut, Noujoud Gabed, Marie Brunet, Xavier Roucou, and Igor Kryvoruchko. 2021. "Mosaic Translation Hypothesis: Chimeric Polypeptides Produced via Multiple Ribosomal Frameshifting as a Basis for Adaptability." *The FEBS Journal*, November. https://doi.org/10.1111/febs.16269.
Carter, Charles W., Jr. 2021. "Simultaneous Codon Usage, the Origin of the Proteome, and the Emergence of de-Novo Proteins." *Current Opinion in Structural Biology* 68 (June): 142–48.
Casola, Claudio. 2018. "From De Novo to 'De Nono': The Majority of Novel Protein-Coding Genes Identified with Phylostratigraphy Are Old Genes or Recent Duplicates." *Genome Biology and Evolution* 10 (11): 2906–18.
Cassan, Elodie, Anne-Muriel Arigon-Chifolleau, Jean-Michel Mesnard, Antoine Gross, and Olivier Gascuel. 2016. "Concomitant Emergence of the Antisense Protein Gene of HIV-1 and of the Pandemic." *Proceedings of the National Academy of Sciences of the United States of America* 113 (41): 11537–42.
Chamakura, Karthik R., Jennifer S. Tran, Chandler O'Leary, Hannah G. Lisciandro, Sophia F. Antillon, Kameron D. Garza, Elizabeth Tran, Lorna Min, and Ry Young. 2020. "Rapid de Novo Evolution of Lysis Genes in Single-Stranded RNA Phages." *Nature Communications* 11 (1): 6009.
Cheng, Hua, R. Dustin Schaeffer, Yuxing Liao, Lisa N. Kinch, Jimin Pei, Shuoyong Shi, Bong-Hyun Kim, and Nick V. Grishin. 2014. "ECOD: An Evolutionary Classification of Protein Domains." *PLoS Computational Biology* 10 (12): e1003926.
Contreras, R., M. Ysebaert, W. M. Jou, and W. Fiers. 1973. "Bacteriophage Ms2 RNA: Nucleotide Sequence of the End of the a Protein Gene and the Intercistronic Region." *Nature: New Biology* 241 (108): 99–101.
Coray, Dorien S., Nellie Sibaeva, Stephanie McGimpsey, and Paul P. Gardner. 2019. "The Genetic Robustness of RNA and Protein from Evolutionary, Structural and Functional Perspectives." *bioRxiv*. https://doi.org/10.1101/480087.
Delaye, Luis, Alexander Deluna, Antonio Lazcano, and Arturo Becerra. 2008. "The Origin of a Novel Gene through Overprinting in Escherichia Coli." *BMC Evolutionary Biology* 8 (January): 31.
Durand, Éléonore, Isabelle Gagnon-Arsenault, Johan Hallin, Isabelle Hatin, Alexandre K. Dubé, Lou Nielly-Thibault, Olivier Namy, and Christian R. Landry. 2019. "Turnover of Ribosome-Associated Transcripts from de Novo ORFs Produces Gene-like Characteristics Available for de Novo Gene Emergence in Wild Yeast Populations." *Genome Research* 29 (6): 932–43.
Fellner, Lea, Niklas Bechtel, Michael A. Witting, Svenja Simon, Philippe Schmitt-Kopplin, Daniel Keim, Siegfried Scherer, and Klaus Neuhaus. 2014. "Phenotype of htgA (mbiA), a Recently Evolved Orphan Gene of Escherichia Coli and Shigella, Completely Overlapping in Antisense to yaaW." *FEMS Microbiology Letters* 350 (1): 57–64.
Feng, Ye, Zeyu Wang, Kun-Yi Chien, Hsiu-Ling Chen, Yi-Hua Liang, Xiaoting Hua, and Cheng-Hsun Chiu. 2022. "'Pseudo-pseudogenes' in bacterial genomes: Proteogenomics reveals a wide but low protein expression of pseudogenes in Salmonella enterica." *Nucleic Acids Research* 50 (9): 5158–70.

Firth, Andrew E. 2014. “Mapping Overlapping Functional Elements Embedded within the Protein-Coding Regions of RNA Viruses.” *Nucleic Acids Research* 42 (20): 12425–39.
———. 2020. “A Putative New SARS-CoV Protein, 3c, Encoded in an ORF Overlapping ORF3a.” *The Journal of General Virology* 101 (10): 1085–89.
Forsdyke, D. R. 1995. “Sense in Antisense?” *Journal of Molecular Evolution* 41 (5): 582–86.
Gelsinger, Diego Rivera, Emma Dallon, Rahul Reddy, Fuad Mohammad, Allen R. Buskirk, and Jocelyne DiRuggiero. 2020. “Ribosome Profiling in Archaea Reveals Leaderless Translation, Novel Translational Initiation Sites, and Ribosome Pausing at Single Codon Resolution.” *Nucleic Acids Research* 48 (10): 5201–16.
Heames, Brennen, Jonathan Schmitz, and Erich Bornberg-Bauer. 2020. “A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila.” *Journal of Molecular Evolution* 88 (4): 382–98.
Hecht, Michael H., Aditi Das, Abigail Go, Luke H. Bradley, and Yinan Wei. 2004. “De Novo Proteins from Designed Combinatorial Libraries.” *Protein Science: A Publication of the Protein Society* 13 (7): 1711–23.
Hücker, Sarah M., Sonja Vanderhaeghen, Isabel Abellan-Schneyder, Siegfried Scherer, and Klaus Neuhaus. 2018. “The Novel Anaerobiosis-Responsive Overlapping Gene Ano Is Overlapping Antisense to the Annotated Gene ECs2385 of Escherichia Coli O157:H7 Sakai.” *Frontiers in Microbiology* 9 (May): 931.
Ingolia, Nicholas T., Gloria A. Brar, Noam Stern-Ginossar, Michael S. Harris, Gaëlle J. S. Talhouarne, Sarah E. Jackson, Mark R. Wills, and Jonathan S. Weissman. 2014. “Ribosome Profiling Reveals Pervasive Translation outside of Annotated Protein-Coding Genes.” *Cell Reports* 8 (5): 1365–79.
Itzkovitz, Shalev, and Uri Alon. 2007. “The Genetic Code Is Nearly Optimal for Allowing Additional Information within Protein-Coding Sequences.” *Genome Research* 17 (4): 405–12.
Jacob, F. 1977. “Evolution and Tinkering.” *Science*. https://www.jstor.org/stable/1744610?casa_token=915UyeD8Q7AAAAAA:pDqkNrpyNa6H6enkNB0j0StEF5x6tlONkC1H6hbOUoviRxBFJSq6HKAczRkCljySNCUVAc21iIhoUP3Cp-KjxiGivK9wWOnjq-2drxZ7hwwfTbSMWQY.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.
Jungreis, Irwin, Chase W. Nelson, Zachary Ardern, Yaara Finkel, Nevan J. Krogan, Kei Sato, John Ziebuhr, et al. 2021. “Conflicting and Ambiguous Names of Overlapping ORFs in the SARS-CoV-2 Genome: A Homology-Based Resolution.” *Virology* 558 (June): 145–51.
Keese, P. K., and A. Gibbs. 1992. “Origins of Genes:‘ Big Bang’ or Continuous Creation?” *Of the National Academy of Sciences*. https://www.pnas.org/content/89/20/9489.short.
Kolodny, Rachel, Sergey Nepomnyachiy, Dan S. Tawfik, and Nir Ben-Tal. 2021. “Bridging Themes: Short Protein Segments Found in Different Architectures.” *Molecular Biology and Evolution* 38 (6): 2191–2208.
Konecny, J., M. Eckert, M. Schöniger, and G. L. Hofacker. 1993. “Neutral Adaptation of the Genetic Code to Double-Strand Coding.” *Journal of Molecular Evolution* 36 (5): 407–16.
Kosinski, Luke, Nathan Aviles, Kevin Gomez, and Joanna Masel. 2022. “Random Peptides Rich in Small and Disorder-Promoting Amino Acids Are Less Likely to Be Harmful.” *Genome Biology and Evolution*, June. https://doi.org/10.1093/gbe/evac085.
Kosinski, Luke J., and Joanna Masel. 2020. “Readthrough Errors Purge Deleterious Cryptic Sequences, Facilitating the Birth of Coding Sequences.” *Molecular Biology and Evolution* 37 (6): 1761–74.
Kreitmeier, Michaela, Zachary Ardern, Miriam Abele, Christina Ludwig, Siegfried Scherer, and Klaus Neuhaus. 2022. “Spotlight on Alternative Frame Coding: Two Long Overlapping Genes in Pseudomonas Aeruginosa Are Translated and under Purifying Selection.” *iScience* 25 (2): 103844.
Kubatova, Nina, Dennis J. Pyper, Hendrik R. A. Jonker, Krishna Saxena, Laura Remmel,

Christian Richter, Sabine Brantl, et al. 2020. “Rapid Biophysical Characterization and NMR Spectroscopy Structural Analysis of Small Proteins from Bacteria and Archaea.” *Chembiochem: A European Journal of Chemical Biology* 21 (8): 1178–87.
Levine, Mia T., Corbin D. Jones, Andrew D. Kern, Heather A. Lindfors, and David J. Begun. 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*. https://doi.org/10.1073/pnas.0509809103.
Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. 2022. “Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model.” *bioRxiv*. https://doi.org/10.1101/2022.07.20.500902.
Louis, Ard A. 2016. “Contingency, Convergence and Hyper-Astronomical Numbers in Biological Evolution.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 58 (August): 107–16.
Miller, R. H. 1988. “Human Immunodeficiency Virus May Encode a Novel Protein on the Genomic DNA plus Strand.” *Science* 239 (4846): 1420–22.
Mudge, Jonathan M., Jorge Ruiz-Orera, John R. Prensner, Marie A. Brunet, Ferriol Calvet, Irwin Jungreis, Jose Manuel Gonzalez, et al. 2022. “Standardized Annotation of Translated Open Reading Frames.” *Nature Biotechnology* 40 (7): 994–99.
Nelson, Chase W., Zachary Ardern, Tony L. Goldberg, Chen Meng, Chen-Hao Kuo, Christina Ludwig, Sergios-Orestis Kolokotronis, and Xinzhu Wei. 2020. “Dynamically Evolving Novel Overlapping Gene as a Factor in the SARS-CoV-2 Pandemic.” *eLife* 9 (October). https://doi.org/10.7554/eLife.59633.
Nelson, Chase W., Zachary Ardern, and Xinzhu Wei. 2020. “OLGenie: Estimating Natural Selection to Predict Functional Overlapping Genes.” *Molecular Biology and Evolution* 37 (8): 2440–49.
Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Springer Berlin Heidelberg.
Pan, Feng, Yuan Zhang, Xiuwen Liu, and Jinfeng Zhang. 2021. “Estimating the Designability of Protein Structures.” *bioRxiv*. https://doi.org/10.1101/2021.11.03.467111.
Papadopoulos, Chris, Isabelle Callebaut, Jean-Christophe Gelly, Isabelle Hatin, Olivier Namy, Maxime Renard, Olivier Lespinet, and Anne Lopes. 2021. “Intergenic ORFs as Elementary Structural Modules of de Novo Gene Birth and Protein Evolution.” *Genome Research*, November. https://doi.org/10.1101/gr.275638.121.
Pavesi, Angelo. 2019. “Asymmetric Evolution in Viral Overlapping Genes Is a Source of Selective Protein Adaptation.” *Virology* 532 (June): 39–47.
“Plutarch, Theseus, Chapter 23, Section 1.” n.d. Accessed November 23, 2022. https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A2008.01.0067%3Achapter%3D23%3Asection%3D1.
Raes, Jeroen, and Yves Van de Peer. 2005. “Functional Divergence of Proteins through Frameshift Mutations.” *Trends in Genetics: TIG* 21 (8): 428–31.
Rancurel, Corinne, Mahvash Khosravi, A. Keith Dunker, Pedro R. Romero, and David Karlin. 2009. “Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into de Novo Protein Creation.” *Journal of Virology* 83 (20): 10719–36.
Ruiz-Orera, Jorge, Pol Verdaguer-Grau, José Luis Villanueva-Cañas, Xavier Messeguer, and M. Mar Albà. 2018. “Translation of Neutrally Evolving Peptides Provides a Basis for de Novo Gene Evolution.” *Nature Ecology & Evolution* 2 (5): 890–96.
Salisbury, F. B. 1969. “Natural Selection and the Complexity of the Gene.” *Nature* 224 (5217): 342–43.
Schmitz, Jonathan F., and Erich Bornberg-Bauer. 2017. “Fact or Fiction: Updates on How Protein-Coding Genes Might Emerge de Novo from Previously Non-Coding DNA.” *F1000Research* 6 (January): 57.
Schmitz, Jonathan F., Kristian K. Ullrich, and Erich Bornberg-Bauer. 2018. “Incipient de Novo Genes Can Evolve from Frozen Accidents That Escaped Rapid Transcript Turnover.” *Nature Ecology & Evolution* 2 (10): 1626–32.
Sealfon, Rachel S., Michael F. Lin, Irwin Jungreis, Maxim Y. Wolf, Manolis Kellis, and Pardis C. Sabeti. 2015. “FRESCo: Finding Regions of Excess Synonymous Constraint in

Diverse Viruses." *Genome Biology* 16 (February): 38.
Sillitoe, Ian, Nicola Bordin, Natalie Dawson, Vaishali P. Waman, Paul Ashford, Harry M. Scholes, Camilla S. M. Pang, et al. 2021. "CATH: Increased Structural Coverage of Functional Space." *Nucleic Acids Research* 49 (D1): D266–73.
Smith, Carol, Jill G. Canestrari, Archer J. Wang, Matthew M. Champion, Keith M. Derbyshire, Todd A. Gray, and Joseph T. Wade. 2022. "Pervasive Translation in Mycobacterium Tuberculosis." *eLife* 11 (March). https://doi.org/10.7554/eLife.73980.
Smith, J. M. 1970. "Natural Selection and the Concept of a Protein Space." *Nature* 225 (5232): 563–64.
Stewart, Hazel, Yongxu Lu, Sarah O'Keefe, Anusha Valpadashi, Luis Daniel Cruz-Zaragoza, Hendrik A. Michel, Samantha K. Nguyen, et al. 2022. "The SARS-CoV-2 Protein ORF3c Is a Mitochondrial Modulator of Innate Immunity." *bioRxiv*. https://doi.org/10.1101/2022.11.15.516323.
Tian, Pengfei, and Robert B. Best. 2017. "How Many Protein Sequences Fold to a Given Structure? A Coevolutionary Analysis." *Biophysical Journal* 113 (8): 1719–30.
Tokuriki, Nobuhiko, Francois Stricher, Joost Schymkowitz, Luis Serrano, and Dan S. Tawfik. 2007. "The Stability Effects of Protein Mutations Appear to Be Universally Distributed." *Journal of Molecular Biology* 369 (5): 1318–32.
Vakirlis, Nikolaos, Omer Acar, Brian Hsu, Nelson Castilho Coelho, S. Branden Van Oss, Aaron Wacholder, Kate Medetgul-Ernar, et al. 2020. "De Novo Emergence of Adaptive Membrane Proteins from Thymine-Rich Genomic Sequences." *Nature Communications* 11 (1): 781.
Vakirlis, Nikolaos, Anne-Ruxandra Carvunis, and Aoife McLysaght. 2020. "Synteny-Based Analyses Indicate That Sequence Divergence Is Not the Main Source of Orphan Genes." *eLife* 9 (February). https://doi.org/10.7554/eLife.53500.
Vakirlis, Nikolaos, Alex S. Hebert, Dana A. Opulente, Guillaume Achaz, Chris Todd Hittinger, Gilles Fischer, Joshua J. Coon, and Ingrid Lafontaine. 2018. "A Molecular Portrait of De Novo Genes in Yeasts." *Molecular Biology and Evolution* 35 (3): 631–45.
Vanderhaeghen, Sonja, Barbara Zehentner, Siegfried Scherer, Klaus Neuhaus, and Zachary Ardern. 2018. "The Novel EHEC Gene Asa Overlaps the TEGT Transporter Gene in Antisense and Is Regulated by NaCl and Growth Phase." *Scientific Reports* 8 (1): 17875.
Van Oss, Stephen Branden, and Anne-Ruxandra Carvunis. 2019. "De Novo Gene Birth." *PLoS Genetics* 15 (5): e1008160.
Wang, Xiaolong, Quanjiang Dong, Gang Chen, Jianye Zhang, Yongqiang Liu, and Yujia Cai. 2022. "Frameshift and Wild-Type Proteins Are Often Highly Similar Because the Genetic Code and Genomes Were Optimized for Frameshift Tolerance." *BMC Genomics*. https://doi.org/10.1186/s12864-022-08435-6.
Watson, Andrew K., Philippe Lopez, and Eric Bapteste. 2022. "Hundreds of Out-of-Frame Remodeled Gene Families in the Escherichia Coli Pangenome." *Molecular Biology and Evolution* 39 (1). https://doi.org/10.1093/molbev/msab329.
Weidmann, Laura, Tjeerd Dijkstra, Oliver Kohlbacher, and Andrei N. Lupas. 2021. "Minor Deviations from Randomness Have Huge Repercussions on the Functional Structuring of Sequence Space." *bioRxiv*. bioRxiv. https://doi.org/10.1101/706119.
Weisman, Caroline M. 2022. "The Origins and Functions of De Novo Genes: Against All Odds?" *Journal of Molecular Evolution* 90 (3-4): 244–57.
Weisman, Caroline M., Andrew W. Murray, and Sean R. Eddy. 2020. "Many, but Not All, Lineage-Specific Genes Can Be Explained by Homology Detection Failure." *PLoS Biology* 18 (11): e3000862.
———. 2022. "Mixing Genome Annotation Methods in a Comparative Analysis Inflates the Apparent Number of Lineage-Specific Genes." *Current Biology: CB* 32 (12): 2632–39.e2.
Weissenow, Konstantin, Michael Heinzinger, and Burkhard Rost. 2022. "Protein Language-Model Embeddings for Fast, Accurate, and Alignment-Free Protein Structure Prediction." *Structure* , May. https://doi.org/10.1016/j.str.2022.05.001.

Wichmann, Stefan, and Zachary Ardern. 2019. “Optimality in the Standard Genetic Code Is Robust with Respect to Comparison Code Sets.” *Bio Systems* 185 (November): 104023.
———. 2022. “Why Is the Average Collateral Effect of Synonymous Mutations so Similar across Alternative Reading Frames?” *bioRxiv*. https://doi.org/10.1101/2022.03.22.485379.
Willis, Sara, and Joanna Masel. 2018. “Gene Birth Contributes to Structural Disorder Encoded by Overlapping Genes.” *Genetics* 210 (1): 303–13.
Wilson, Benjamin A., Scott G. Foy, Rafik Neme, and Joanna Masel. 2017. “Young Genes Are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth.” *Nature Ecology & Evolution* 1 (6): 0146–0146.
Wright, Bradley W., Mark P. Molloy, and Paul R. Jaschke. 2022. “Overlapping Genes in Natural and Engineered Genomes.” *Nature Reviews. Genetics* 23 (3): 154–68.
Wright, Bradley W., Zixin Yi, Jonathan S. Weissman, and Jin Chen. 2022. “The Dark Proteome: Translation from Noncanonical Open Reading Frames.” *Trends in Cell Biology* 32 (3): 243–58.
Xu, Haiqing, and Jianzhi Zhang. 2021. “On the Origin of Frameshift-Robustness of the Standard Genetic Code.” *Molecular Biology and Evolution* 38 (10): 4301–9.
Zehentner, Barbara, Zachary Ardern, Michaela Kreitmeier, Siegfried Scherer, and Klaus Neuhaus. 2020a. “A Novel pH-Regulated, Unusual 603 Bp Overlapping Protein Coding Gene Pop Is Encoded Antisense to ompA in Escherichia Coli O157:H7 (EHEC).” *Frontiers in Microbiology* 11 (March): 377.
———. 2020b. “Evidence for Numerous Embedded Antisense Overlapping Genes in Diverse E. Coli Strains.” *bioRxiv*. https://doi.org/10.1101/2020.11.18.388249.
Zheng, Eric B., and Li Zhao. 2022. “Protein Evidence of Unannotated ORFs in Drosophila Reveals Diversity in the Evolution and Properties of Young Proteins.” *eLife* 11 (September). https://doi.org/10.7554/eLife.78772.
Zile, Karina, Christophe Dessimoz, Yannick Wurm, and Joanna Masel. 2020. “Only a Single Taxonomically Restricted Gene Family in the Drosophila Melanogaster Subgroup Can Be Identified with High Confidence.” *Genome Biology and Evolution* 12 (8): 1355–66.
Zull, J. E., and S. K. Smith. 1990. “Is Genetic Code Redundancy Related to Retention of Structural Information in Both DNA Strands?” *Trends in Biochemical Sciences* 15 (7): 257–61.