Gene birth contributes to structural disorder encoded by overlapping genes

S. Willis and J. Masel*
Department of Ecology and Evolutionary Biology, University of Arizona
*Corresponding Author: masel@email.arizona.edu

Abstract

The same nucleotide sequence can encode two protein products in different reading frames. Overlapping gene regions encode higher levels of intrinsic structural disorder (ISD) than non-overlapping genes (39% vs. 25% in our viral dataset). This might be because of the intrinsic properties of the genetic code, because one member per pair was recently born de novo in a process that favors high ISD, or because high ISD relieves increased evolutionary constraint imposed by dual-coding. Here we quantify the relative contributions of these three alternative hypotheses. We estimate that the recency of de novo gene birth explains 32% or more of the elevation in ISD in overlapping regions of viral genes. While the two reading frames within a same-strand overlapping gene pair have markedly different ISD tendencies that must be controlled for, their effects cancel out to make no net contribution to ISD. The remaining elevation of ISD in the older members of overlapping gene pairs, presumed due to the need to alleviate evolutionary constraint, was already present prior to the origin of the overlap. Same-strand overlapping gene birth events can occur in two different frames, favoring high ISD either in the ancestral gene or in the novel gene; surprisingly, most de novo gene birth events contained completely within the body of an ancestral gene favor high ISD in the ancestral gene (23 phylogenetically independent events vs. 1). This can be explained by mutation bias favoring the frame with more start codons and fewer stop codons.

I. Introduction

Protein-coding genes sometimes overlap, i.e. the same nucleotide sequence encodes different proteins in different reading frames. Most of the overlapping pairs of genes that have been characterized to date are found in viral, bacterial and mitochondrial genomes, with emerging research showing that they may be common in eukaryotic genomes as well [10], [20], [31], [32], [37]. Studying overlapping genes can shed light on the processes of de novo gene birth [36] .

Overlapping genes tend to encode proteins with higher intrinsic structural disorder (ISD) than those encoded by non-overlapping genes [36]. The term disorder applies broadly to proteins which, at least in the absence of a binding partner, lack a stable secondary and tertiary structure. There are different degrees of disorder: molten globules, partially unstructured proteins and random coils with regions of disorder spanning from short (less than 30 residues in length) to long. Disorder can be shown experimentally or predicted from amino acid sequences using software [15]. [36] estimated, using the latter approach, that 48% of amino acids in overlapping regions exhibit disorder, compared to only 23% in non-overlapping regions.

In this work we explore three non-mutually-exclusive hypotheses why this might be, and quantify the extent of each. Two have previously been considered: that elevated ISD in overlapping genes is a mechanism that relieves evolutionary constraint, and that elevated ISD is a holdover from the de novo gene birth process. We add consideration of a third, previously-unexplored hypothesis — that elevated ISD with dual-coding may be the result of an artifact of the genetic code — to the mix.

Overlapping genes are particularly evolutionarily constrained because a mutation in an overlapping region simultaneously affects both of the two (or occasionally more) genes involved in that overlap. Because $\sim$70% of mutations that occur in the third codon position are synonymous, versus only $\sim$5% and 0% of mutations in the first and second codon positions respectively [38], a mutation that is synonymous in one reading frame is highly likely to be nonsynonymous in another, so to permit adaptation, overlapping genes must be relatively tolerant of nonsynonymous changes. Demonstrating the higher constraint on overlapping regions, they have lower genetic diversity and d_N/d_S

than non-overlapping regions in RNA viruses [25], [39], [40], [43].

High ISD can alleviate the problem of constraint. Amino acid substitutions that maintain disorder have a reasonable chance of being tolerated, in contrast to the relative fragility of a well-defined three-dimensional structure. This expectation is confirmed by the higher evolutionary rates observed for disordered proteins [4], [13]. The known elevation of evolutionary constraint on overlapping genes is usually invoked as the sole [47], [53] or at least dominant [36] explanation for their high ISD. Given the strength of the evidence for constraint [25], [39], [40], [43], we attribute to constraint, by a process of elimination, what cannot be explained by our other two hypotheses.

The second hypothesis that we consider is that high ISD in overlapping genes is an artifact of the process of de novo gene birth [36]. There is no plausible path by which two non-overlapping genes could re-encode an equivalent protein sequence as overlapping; instead, an overlapping pair arises either when a second gene is born de novo within an existing gene, or when the boundaries of an existing gene are extended to create overlap [38]. In the latter case of ”overprinting” [7], [19], [36], the extended portion of that gene, if not the whole gene, is born de novo. One overlapping protein-coding sequence is therefore always evolutionarily younger than the other; we refer to these as ”novel” versus ”ancestral” overlapping genes or portions of genes. Genes may eventually lose their overlap through a process of gene duplication followed by subfunctionalization [19], enriching overlapping genes for relatively young genes that have not yet been through this process. However, gene duplication may be inaccessible to many viruses (in particular, many RNA, ssDNA, and retroviruses), due to intrinsic geometric constraints on maximum nucleotide length [6], [9], [11].

Young genes are known to have higher ISD than old genes, with high ISD at the moment of gene birth facilitating the process [52], perhaps because cells tolerate them better [48]. Domains that were more recently born de novo also have higher ISD [3], [5], [14], [26]. High ISD could be helpful in itself in creating novel function, or it could be a byproduct of a hydrophilic amino acid composition whose function is simply the avoidance of harmful protein aggregation [16], [22]. Regardless of the cause of high ISD in young genes, the ”facilitate birth” hypothesis makes a distinct prediction from the constraint hypothesis, namely that the novel overlapping reading frames will tend to encode higher ISD than the ancestral overlapping reading frames.

Under the constraint hypothesis, ancestral overlapping proteins will still have elevated ISD relative to non-overlapping proteins, even if it is less elevated than that of novel overlapping proteins. Elevated ISD in the ancestral member of the gene pair might have already been there at the moment of gene birth, or it might have subsequently evolved, representing two subhypotheses within the overall hypothesis of constraint. The overlapping gene pairs that we observe are those that have been retained; if either member of the overlapping pair was born with low ISD, then constraint makes it difficult to adapt to a changing environment, and that pair is less likely to be retained. When the ancestral member of the pair already has high ISD at the moment at which the novel gene is functionalized, long-term maintenance of both genes in the face of constraint is more likely. The ”already there” or ”preadaptation” [52] version of the constraint hypothesis predicts that the pre-overlapping ancestors of today’s ancestral overlapping genes had higher ISD than other genes, perhaps because these gene pairs are the ones to have been retained. While these ancestral sequences are not available, as a proxy we use homologous sequences from basal lineages whose most recent common ancestry predates the origin of the overlap. For simplicity, we refer to these sequences as ”pre-overlapping” to distinguish them from ”non-overlapping” genes never known to have overlap. The preadaptation version of the constraint hypothesis predicts higher ISD in pre-overlapping genes than in non-overlapping genes.

Finally, here we also consider the possibility that the high ISD observed in overlapping genes might simply be an artifact of the genetic code [21]. We perform for the first time the appropriate control, by predicting what the ISD would be if codons were read from alternative reading frames of existing non-overlapping genes. Any DNA sequence can be read in three reading frames on each of the two strands, for a total of 6 reading frames. We focus only on same-strand overlap, due to superior availability of reliable data on same-strand overlapping gene pairs. We classify the reading frame of each gene in an overlapping pair relative to its counterpart; if gene A is in the +1 frame with respect to gene B, this means that gene B is in the +2 frame with respect to gene A. We use the +0 frame designation just for non-overlapping or pre-overlapping genes in their original frame. If the high ISD of overlapping genes is primarily driven by the intrinsic properties of the genetic code,

then we expect their ISD values to closely match those expected from translation in the +1 vs. +2 frames of non-overlapping genes.

Hypothesis	ISD Prediction
Artifact of the Genetic Code	+1 Frame Overlapping = +1 Controls +2 Frame Overlapping = +2 Controls
Conflict Resolution	Ancestral Overlapping* > Non-Overlapping *Controlling for Frame Effects
Preadaptation	Pre-Overlap Ancestral Homologs = Overlapping
Facilitate Birth	Novel > Ancestral

Fig. 1. Three non-mutually-exclusive hypotheses about why overlapping genes have high ISD. The column on the right describes the ISD patterns we would expect if the hypotheses were true.

Here we test the predictions of all three hypotheses, as summarized in Figure 1, and find that both the birth-facilitation and conflict-resolution hypotheses play a role. The artifact hypothesis plays no appreciable role in elevating the ISD of overlapping regions; while reading frame (+1 vs. +2) strongly affects the ISD of individual genes, each overlapping gene pair has one of each, and the two cancel out such that there is no net contribution to the high ISD found in overlapping regions. Surprisingly, novel genes are more likely to be born in the frame prone to lower ISD; this seems to be a case where mutation bias in the availability of open reading frames (ORFs) is more important than selection favoring higher ISD for novel than ancestral genes.

II. Materials and Methods

A. Overlapping Viral Genes

A total of 102 viral same-strand overlapping gene pairs were compiled from the literature [35], [36], [40], [42], [43], [51]. Of these, ten were discarded because one or both of the genes involved in the overlap were not found in the ncbi databases, either because the accession number had been removed, or because the listed gene could not be located. This left 92 gene pairs for analysis from 80 different species, spanning 33 viral families. Six of these pairs were ssDNA, five were retroviruses, while the remaining 81 were RNA viruses: 7 dsRNA, 61 positive sense RNA and 13 negative sense RNA.

B. Relative Gene Age

For 39 of the remaining 92 gene pairs available for analysis, the identity of the older vs. younger member of the pair had been classified in the literature [27], [36], [40], [42] via phylogenetic analysis. There was disagreement in the literature regarding the TGBp2/TGBp3 overlap; we followed [27] rather than [36].

We also used relative levels of codon bias to classify the relative ages of members of each pair. Because all of the overlapping genes are from viral genomes, we can assume that they are highly expressed, leading to a strong expectation of codon bias in general. Novel genes are expected to have less extreme codon bias than ancestral genes due to evolutionary inertia [35], [40].

For each viral species, codon usage data [29], [55] were used to calculate a relative synonymous codon usage (RSCU) value for each codon [17]:

$$\text{RSCU}_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

where X_i is the number of occurrences of codon i in the viral genome, and $1 \leq n \leq 6$ is the number of synonymous codons which code for the same amino acid as codon i. The relative adaptedness value (w_i) for each codon in a viral species was then calculated as:

$$w_i = \frac{\text{RSCU}_i}{\text{RSCU}_{\text{max}}}$$

where RSCU_{max} is the RSCU value for the most frequently occurring codon corresponding to the amino acid associated with codon i. The codon adaptation index (CAI) was then calculated for the overlapping portion of each gene, as the geometric mean of the relative adaptedness values:

$$\text{CAI} = \left(\prod_{i=1}^{L} w_i\right)^{\frac{1}{L}}$$

where L is the number of codons in the overlapping portion of the gene, excluding ATG and TGG codons. This exclusion is because ATG and TGG are the only codons that code for their respective amino acids and so their relative adaptedness values are always 1, thereby introducing no new information. To ensure sufficient statistical power to differentiate between CAI values, we did not analyze CAI for gene pairs with overlapping sections less than 200 nucleotides long.

Within each overlapping pair, we provisionally classified the gene with the higher CAI value as ancestral and the gene with lower CAI value as novel. We then compared the two sets of relative adaptedness

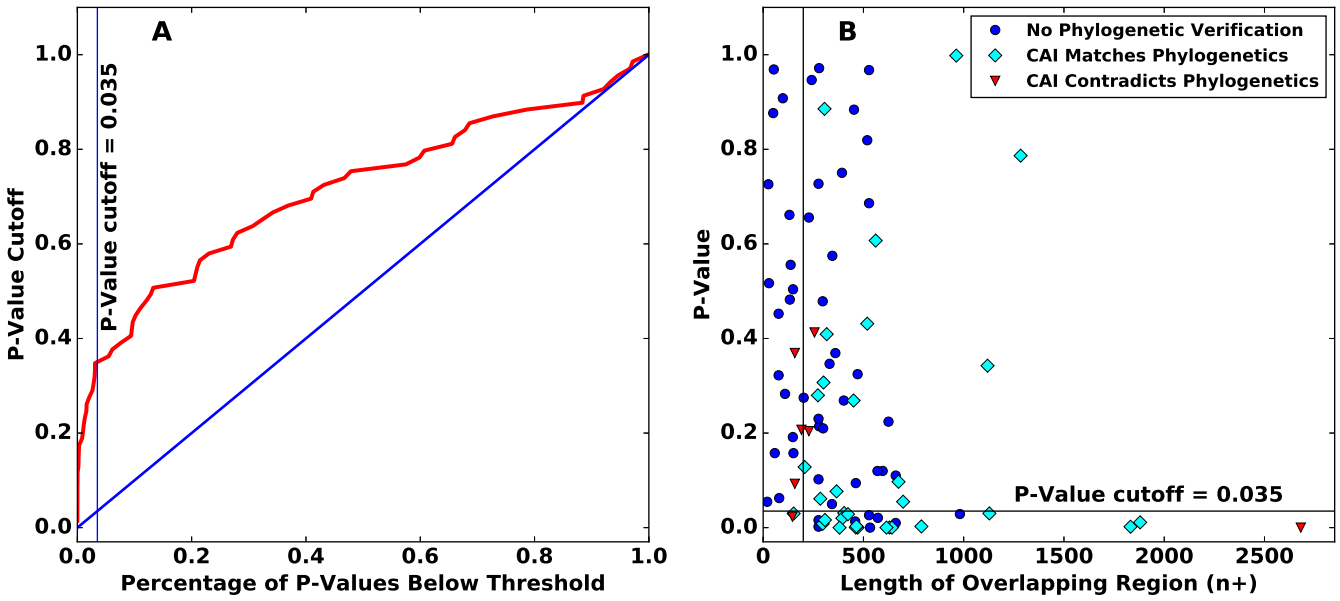

Fig. 2. Statistical classification of relative ages. **(A)** The receiver operating characteristic plot for determining which member of an overlapping gene pair has higher CAI, and is hence presumed to be ancestral. Only genes with an overlapping region of at least 200 nucleotides are plotted. **(B)** CAI classification of the 91 gene pairs for which codon usage data were available was based both on p-value and on the length of the overlapping regions. The vertical line shows the overlapping length cutoff of 200 nucleotides, the horizontal line shows the p-value cutoff; CAI classification was considered informative for the 27 bottom right points.

valuesusing the wilcox.test function in R to perform a Mann-Whitney U Test. We chose a p-value cutoff of 0.035 after analyzing a receiver operating characteristic (ROC) plot (Figure 2A). The combined effects of our length threshold and p-value cutoff are illustrated in Figure 2B. In total, 27 gene pairs were determined to have statistically-significant CAI values, 19 of which had also been classified via phylogenetic analysis.

Of the gene pairs whose ancestral vs. novel classification were obtained both by statistically significant CAI differences and by phylogenetics, there was one for which the CAI classification contradicted the phylogenetics. That exception was the p104/p130 overlap in the Providence virus. This overlap is notable because the ancestral member of the pair was acquired through horizontal gene transfer, which renders codon usage an unreliable predictor of relative gene ages [35]. We therefore used the phylogenetic classification and disregarded the CAI results. In total, we were able to classify ancestral vs. novel status for 47 overlapping gene pairs (Figure 3).

C. Non-overlapping and Pre-overlapping Controls

Both non-overlapping genes and pre-overlapping genes were used as controls. 150 non-overlapping genes were compiled from the viral species in which the 92 overlapping gene pairs were found; matching for species helps control for %GC content or other idiosyncrasies of nucleotide composition.

Of the 47 overlapping gene pairs for which we could assign relative ages, we were able to locate pre-overlapping homologs for 27 of the ancestral genes in our dataset in the literature [40] and/or NCBI (BLAST search with E-value threshold $= 10^{-6}$).

Frameshifting these control sequences was performed in two ways. First, removing one or two nucleotides immediately after the start codon and two or one nucleotides immediately before the stop codon generated +1 and +2 frameshifted non-ORF controls, respectively (Figure 4A). Any stop codons that appeared in the new reading frames were removed.

Second, for the frameshifted ORF controls, we took only in situ ORFs within each of the two alternative reading frames. If multiple ORFs terminated at the same stop codon, we used only the longest. We excluded ORFs less than 25 amino acids in length (after removal of cysteines for analysis by IUPred). For non-overlapping genes, this yielded 151 and 24 ORFs in the +1 and +2 reading frames, respectively. For the smaller set of pre-overlapping genes, it yielded only 25 and 5 ORFs in the +1 and +2 frames, respectively.

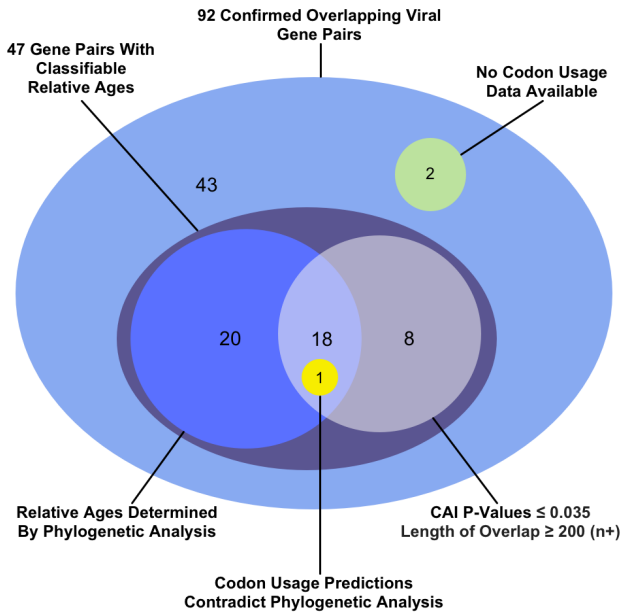

Fig. 3. How the relative ages of the genes were classified for 47 out of the 92 overlapping gene pairs for which sequence data were available. Within each shaded region, each gene pair is counted within only one of the subregions shown. Each shaded region's total is found by summing the individual subtotals within it, some of which are noted outside the shaded regions. For example, the relative ages of 39 genes were classified via phylogenetic analysis: 20 through phylogenetic analysis alone (blue circle), 18 supported via codon analysis (intersection of blue and white circle), and one contradicted by codon analysis (yellow circle).

D. Homology Groups

Treating each gene as an independent datapoint is a form of pseudoreplication, because homologous genes can share properties via a common ancestor rather than via independent evolution. This problem of phylogenetic confounding can be corrected for by using gene family as a random effect term in a linear model [52], and by counting each gene birth event only once.

We constructed a pHMMer (`http://hmmer.org/`) database including all overlapping regions, non-overlapping genes and their frameshifted controls. After an all-against-all search, sequences that were identified as homologous, using an expectation value threshold of 10^{-4}, were provisionally assigned the same homology group ID. These provisional groups were used to determine which gene birth events were unique. Two pairs were considered to come from the same gene birth event when both the ancestral and the overlapping sequence were classified as homologous. We also used published phylogenetic analysis to classify the TGBp2/TGBp3 overlap as two birth events (one occurring *Virgaviridae*, the other occurring in *Alpha-* and *Betaflexiviridae*) [27].

Some homologous pairs had such dissimilar protein sequences that ISD values were essentially independent. We therefore manually analyzed sequence similarity within each homology group using the Geneious [18] aligner with free end gaps, using Blosum62 as the cost matrix. The percent similarity using the Blosum62 matrix with similarity threshold 1 was then used as the criterion for whether a protein sequence would remain in its homology group for the ISD analysis. We used $\geq 50\%$ protein sequence similarity as the threshold to assign a link between a pair, and then used single-link clustering to assign protein sequences to 561 distinct homology groups. Pre-overlapping genes were then assigned to the homology groups of the corresponding ancestral genes.

E. ISD Prediction

We used IUPred [12] to calculate ISD values for each sequence. Following [52], before running IUPred, we excised all cysteines from each amino acid sequence, because of the uncertainty about their disulphide bond status and hence entropy [49]. Whether cysteine forms a disulphide bond depends on whether it is in an oxidizing or reducing environment. IUPred implicitly, through the selection of its training data set, assumes most cysteines are in disulphide bonds, which may or may not be accurate for our set of viral proteins. Because cysteines have large effects on ISD (in either direction) depending on disulphide status and hence introduce large inaccuracies, cysteines were dropped from consideration altogether.

IUPred assigns a score between 0 and 1 to each amino acid. To calculate the ISD of an overlapping region, IUPred was run on the complete protein (minus its cysteines), then the average score was taken across only the pertinent subset of amino acids.

F. Statistical Models

Prior to fitting linear models, sequence-level ISD values were transformed using a Box-Cox transform. The optimal value of λ for the combined ancestral, novel and artificially-frameshifted non-overlapping, non-ORF control group data was 0.41, which we rounded to 0.4 and used throughout all linear models, and for central tendency and confidence intervals in the figures. Simple means and standard errors are described in-line in the text.

We used a multiple regression approach to determine which factors predict ISD values [44]. Gene

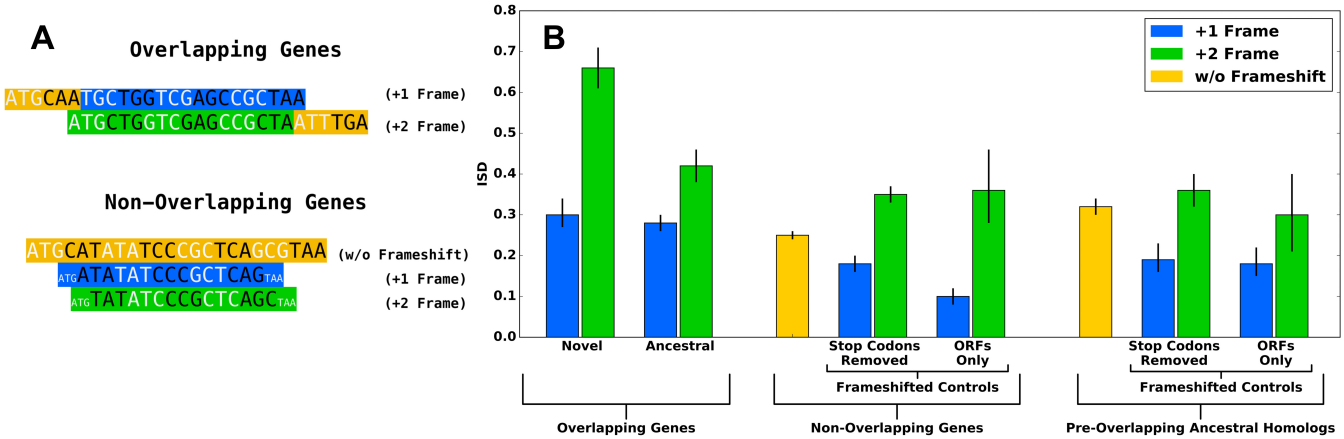

Fig. 4. ISD results support the birth-facilitation and (preadaptation version of the) conflict resolution hypotheses. (A) Data are from the overlapping sections of the 47 gene pairs whose ages could be classified, from non-overlapping genes in the species in which those 47 overlapping gene pairs were found, and from the 27 available pre-overlapping ancestral homologs. Some numbers in the main text are for the full set of 92 overlapping gene pairs and might not match exactly. (B) While frame significantly impacts disorder content (green is consistently higher than blue), it does not drive the high ISD of overlapping genes. Even controlling for frame, novel ISD $>$ ancestral (left), supporting birth facilitation. Supporting conflict resolution, ancestral $>$ non-overlapping either in-frame (yellow) or matched frameshifted control. The preadaptation version of the constraint hypothesis is supported by the fact that non-overlapping ISD $<$ pre-overlapping (compare the two yellow bars). Means and 66% confidence intervals were calculated from the Box-Cox transformed (with λ = 0.4) means and their standard errors, and are shown here following back-transformation.

designation (ancestral vs. novel vs. one or more non-genic controls) and relative reading frame (+1 vs. +2) were used as fixed effects. Homologous sequences are not independent; we accounted for this by using a linear mixed model [34], with homology group as a random effect. Species is a stand-in for a number of confounding factors, e.g. %GC content, and so was included as a second random effect. The data were normalized using a Box-Cox transformation prior to analysis. Pairwise comparisons discussed throughout the Results were performed using contrast statements applied to the linear model, using the minimum number of gene designations necessary to make the comparison in question.

We used the lmer and gls functions contained in the nlme and lme4 R packages to generate the linear mixed models. The main model used to calculate the relative effect sizes used frameshifted non-ORF non-overlapping genes as the non-genic control. In this model, frame, gene designation, species, and homology group terms were retained in the model with $p = 5\times10^{-20}$, 1×10^{-6}, 9×10^{-19}, and 2×10^{-3} respectively. All four terms were also easily retained in all other model variants exploring different non-genic controls.

We also considered overlap type (internal vs. terminal) as a fixed effect, but removed it because it did not significantly enhance our model ($p = 0.86$). To determine the statistical significance of each effect, we used the anova function in R to compare nested models.

G. Data Availability

Scripts and data tables used in this work may be accessed at: `https://github.com/MaselLab/Willis_Masel_Overlapping_Genes_Structural_Disorder_Explained`

III. Results

A. Confirming elevated ISD in our dataset

Because most verified gene overlaps in the literature, especially longer overlapping sequences, are in viruses [30], [36], [50], we focused on viral genomes, compiling a list of 92 verified overlapping gene pairs from 80 viral species. The mean predicted ISD of all overlapping regions (0.39 ± 0.02) was higher than that of the non-overlapping genes (0.25 ± 0.01), confirming previous findings that overlapping genes have elevated ISD.

B. Frame affects ISD

We artificially frameshifted 150 non-overlapping viral genes in those 80 species, and found higher ISD in the +2 reading frame (0.35 ± 0.02) than in the +1 reading frame (0.19 ± 0.01) for non-ORF controls, and an even more extreme difference for ORF

controls (0.47 ± 0.02 vs. 0.18 ± 0.01). The artifact hypothesis predicts that the +1 and +2 members of the 92 verified overlapping gene pairs will follow suit. In agreement with this, the overlapping regions of genes in the +2 reading frame had higher mean ISD (0.48 ± 0.03) than those in the +1 reading frame (0.31 ± 0.02). While this provides strong evidence that frame shapes ISD as an artifact of the genetic code, average ISD across both ways of frameshifting non-overlapping genes (0.27 ± 0.01 and 0.22 ± 0.02 for non-ORF and ORF frameshifted non-overlapping genes, respectively) is significantly lower than the ISD of all overlapping sequences (0.39 ± 0.02), showing that the artifact hypothesis cannot fully explain elevated ISD in the latter.

The Birth-Facilitation Hypothesis is Supported

We find stronger support for the birth-facilitation hypothesis. Of the 92 verified overlapping viral gene pairs, we were able to classify the relative ages of the component genes as ancestral vs. novel for 47 pairs (Table I). In agreement with the predictions of the birth-facilitation process, and controlling for frame, novel genes have higher ISD than either ancestral members of the same gene pairs or artificially-frameshifted controls (Figure 4B). We confirmed this using a linear mixed model, with frame (+1 vs. +2) as a fixed effect, gene type (novel vs. ancestral) as a fixed effect, species (to control for %GC content and other subtle sequence biases) as a random effect, and homology group (to control for phylogenetic confounding) as a random effect. Within this linear model, the prediction unique to the birth-facilitation hypothesis, namely that ISD in the overlapping regions of novel genes is higher than that in ancestral genes, is supported with $p = 0.03$. Inspection of Figure 4B suggests that elevation of novel gene ISD above ancestral is specific to the +2 frame; this is confirmed in the analysis of Figure 5B. Running separate statistical models for the two frames, the +2 frame difference is supported with $p = 0.006$.

These non-ORF frameshifted control sequences do not take into account the fact that ORFs vary in their propensity to appear and disappear, and that this can shape the material available for de novo gene birth, including ISD values [33]. However, we did not observe this affecting ISD in our data set. In a linear model comparing ORF vs non-ORF non-overlapping control sequences, and one comparing ORF vs. non-ORF pre-overlapping controls sequences, there was no significant difference between the two controls ($p = 0.7$ in both cases, with frame included as a fixed effect, and gene as a random effect). This justifies a focus on the larger non-ORF control dataset, as well as excluding this non-adaptive force as a driver of the birth facilitation hypothesis.

The Preadaptation Version of the Conflict-Resolution Hypothesis is Supported

In agreement with the conflict resolution hypothesis, ancestral overlapping sequences, not just novel ones, have higher ISD than non-overlapping genes (0.35 ± 0.02 vs. 0.25 ± 0.01 ; $p = 2 \times 10^{-5}$). This seems to be because ISD was already high at the time of de novo gene birth; today's ancestral overlapping genes have indistiguishable ISD from their pre-overlapping homologs ($p = 0.6$), while those pre-overlapping homologs have higher ISD than non-overlapping genes (p= 2×10^{-3}) (Figure 4B).

It is possible that the overlapping gene pairs for which we are able to identify pre-overlapping homologs are significantly younger than other overlapping gene pairs, and that there might therefore simply not have been enough time for the ancestral members of the pair to evolve higher ISD. Contradicting this, these 27 ancestral overlapping genes have indistinguishable ISD from the other 20 ancestral overlapping gene (p= 0.2, controlling for frame and homology group).

The Artifact Hypothesis is Rejected

Non-overlapping gene ISD is not statistically different from the mean of the +1 and +2 artificially-frameshifted control versions of the same non-overlapping nucleotide sequences ($p = 0.88$; contrast statement applied to a linear model with fixed effect of actual non-overlapping gene sequence vs. +1 artificially-frameshifted version vs +2 artificially-frameshifted version, and gene identity as a random effect). In other words, despite the enormous effect of +1 vs. +2 reading frame, we find no support for the artifact hypothesis in explaining the elevated ISD of overlapping regions. In each overlapping gene pair, there is always exactly one gene in each of the two reading frames, such that the large effects of each of the two frames cancel each other out when all overlapping genes are considered together. It is nevertheless important to control for the large effect of frame while testing and quantifying other hypotheses.

The Relative Magnitude of Each Hypothesis

We calculated the degree to which birth facilitation elevates ISD using a contrast statement, as half the

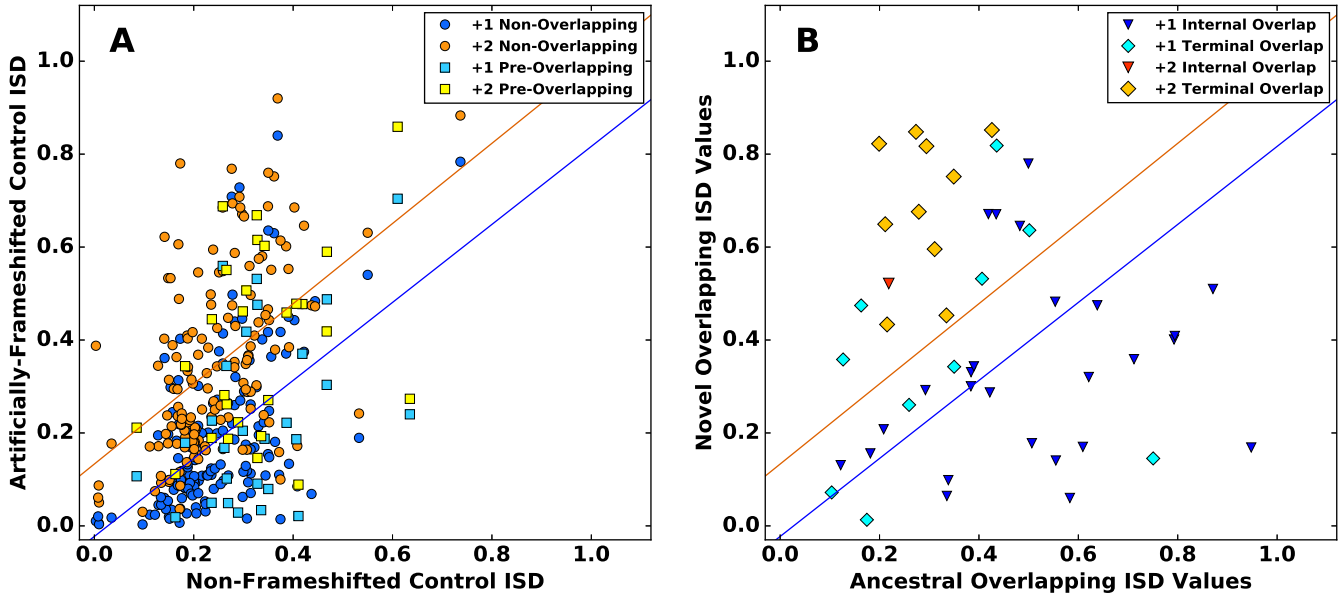

Fig. 5. **(A)** The ISD of a non-overlapping gene predicts the ISD of the non-ORF frameshifted version of that sequence. Model I regression lines are statistically indistiguishable in our set of 150 nonoverlapping viral genes vs. our set of 27 preoverlapping genes, which are therefore pooled ($R^2 = 0.30$ and $R^2 = 0.24$ for the +1 and +2 groups, respectively). **(B)** The overlapping ISD of the 47 gene pairs with classifiable relative ages. Each datapoint represents one overlapping pair. The regression lines from (A) are superimposed to illustrate the elevation of novel gene ISD born into the already intrinsically high-ISD +2 frame, which destroys the correlation ($R^2 = 6.6 \times 10^{-5}$ for +2 in contrast to 0.26 for +1).

difference between novel and ancestral genes, because exactly half of the genes are novel, and hence elevated above the "normal" ISD level of ancestral genes. By this calculation, birth facilitation accounts for $32\% \pm 13\%$ of the estimated total difference in ISD between overlapping and non-overlapping genes.

Note that frameshifted versions of high-ISD proteins have higher ISD than frameshifted versions of low-ISD proteins (Figure 5A). The two members of an overlapping pair share the same %GC content, and random sequences with higher %GC have substantially higher ISD [1], so this could be responsible for the trait correlation. A facilitate-birth bias toward high ISD in newborn genes might do so in part via high %GC in the overlapping region at the time of birth, causing overlapping sequences to be biased not just toward high-ISD novel genes, but also toward high-ISD ancestral genes. Our 32% estimate attributes all of the ISD elevation in ancestral overlapping genes to constraint, but given the trait correlation shown in Figure 5A, some of this might also be due to birth facilitation, making 32% an underestimate. Note that novel genes born into the +2 frame have high ISD above and beyond the intrinsic correlation (Figure 5B), and are thus responsible for the statistical difference between ancestral vs. novel when frame is controlled for (Figure 4B).

C. Mutation Bias is Responsible for More Births in the Low-ISD +1 Frame

Given the strong influence of frame combined with support for the facilitate-birth hypothesis, we hypothesized that novel genes would be born more often into the +2 frame (Figure 4A, green) because the intrinsically higher ISD of the +2 reading frame would facilitate high ISD in the novel gene and hence birth. Our dataset contained 41 phylogenetically independent overlapping pairs. Surprisingly, we found the opposite of our prediction: 31 of the novel genes were in the +1 frame of their ancestral counterparts, while only 10 were in the +2 frame ($p = 10^{-3}$, cumulative binomial distribution with trial success probability 0.5).

This unexpected result is stronger for "internal overlaps", in which one gene is completely contained within its overlapping partner (23 +1 events vs. 1 +2 event, $p = 3 \times 10^{-6}$), and is not found for "terminal overlaps", in which the $5'$ end of the downstream gene overlaps with the $3'$ end of the upstream member of the pair (9 +1 events vs. 9 +2 events). (This double-counts a +1 event for which there were three homologous gene pairs, two of which were internal overlaps, and one of which was a terminal overlap.) Following [2],

we interpret the restriction of this finding to internal overlaps as evidence that the cause of the bias applies to complete de novo gene birth, but not to the addition of a sequence to an existing gene.

The unexpected prevalence of +1 gene births, despite birth facilitation favoring +2, can be explained by mutation bias. One artifact of the genetic code is that +1 frameshifts yield more start codons and fewer stop codons, and hence fewer and shorter ORFs [2]. In our control set of 150 non-overlapping viral genes, we confirm that stop codons are more prevalent in the +2 frame (1 per 11 codons) than the +1 frame (1 per 14), decreasing the mean ORF length, and that start codons are more prevalent in the +1 frame (1 per 27 codons) than the +2 frame (1 per 111). Similar results were found in the pre-overlapping ancestral homologs, with more start codons in the +1 frame (1 per 33) than the +2 frame (1 per 169), and fewer stop codons in the +1 frame (1 per 20) than the +2 frame (1 per 13).

This is reflected in the relative numbers and sizes of our frameshifted ORF controls. Prior to implementing a minimum length requirement (see Materials and Methods), we found 465 ORFs in the +1 frame of our non-overlapping genes, with a mean and maximum length of 24 and 149 amino acids, respectively, while only 92 ORFs were found in the +2 frame with a mean and maximum length of 19 and 92 amino acids. The same pattern was found in the pre-overlapping ancestral homologs, with 65 ORFs found in the +1 frame with a mean and maximum length of 36 and 315 amino acids, vs. 13 ORFs in the +2 frame with a mean and maximum length of 36 and 173 amino acids.

IV. Discussion

There is growing interest in the topic of de novo gene birth, but identifying de novo genes is plagued with high rates of both false positives and false negatives [24], with phylostratigraphy tools being particularly controversial due to homology detection biases [28]. The overlapping viral genes that we study are unlikely either to be non-genes, and must have arisen via de novo gene birth, and so circumvent many of these difficulties. [8] have disputed that young genes have high ISD, in an analysis that was prone to false positives [52]; our findings here provide an independent line of evidence, free from the danger of homology detection bias, that younger genes have higher ISD.

The study of overlapping genes has of course its own statistical traps. In particular, the preponderance of novel genes in the +1 frame demonstrates the need to control for the strong effects of frame when testing hypotheses. Ancestral genes are more frequently in the high-ISD +2 frame, while the depressed ISD of the +1 frame lowers the ISD of the novel. As a result, when frame is not considered, ancestral and novel overlapping sequences encode very similar levels of disorder (0.41 ± 0.03 vs. 0.42 ± 0.04, respectively), making it easy to miss the evidence for the facilitate-birth hypothesis.

More broadly, our results are consistent with a major role for mutational availability in shaping adaptive evolution. Rare adaptive changes happen at a rate given by the product of mutation and the probability of fixation, with the latter proportional to the selection coefficient [23]. This means that differences in the beneficial mutation rate are just as important as differences in the selection coefficient in determining which path adaptive evolution takes [54]. The influence of mutational bias has previously been observed for beneficial mutations to single amino acids in the laboratory [41], [46] and in the wild [45]. Here we demonstrate it for more radical mutations, namely the de novo birth of entire protein-coding genes.

Accession Number	Organism	Ancestral Gene	Novel Gene	Overlap Length (n+)	Novel Frame
NC_001401	Adeno-Associated Virus 2	VP2	AAP	615	+1
NC_004285	Aedes Albopictus Densovirus	NS1	NS2	1119	+1
NC_001467	African Cassava Mosaic Virus	AL1	AC4	423	+1
NC_009896	Akabane Virus[a]	N	NSs	276	+1
NC_001749	Apple Stem Grooving Virus	MP	Polyprotein	963	+1
NC_001719	Arctic Ground Squirrel Hepatitis Virus[b]	P	L	1284	+1
NC_003481	Barley Stripe Mosaic Virus[cde]	TGBp2	TGBp3	191	+1
NC_003680	Barley Yellow Dwarf Virus[fg]	P5	MP	465	+1
NC_005041	Blattella Germanica Densovirus	NS-1	ORF4	789	+1
NC_001927	Bunyamwera Virus[a]	N	NSs	306	+1
NC_001658	Cassava Common Mosaic Virus[de]	TGBp2	TGBp3	152	+1
NC_001427	Chicken Anemia Virus	VP2	Apoptin	366	+1
NC_003688	Cucurbit Aphid-Born Yellowing Virus[fgh]	CP	P5	572	+1
NC_005899	Dendrolimus Punctatus Tetravirus[f]	p71	p17	381	+1
NC_016561	Hepatitis B[b]	P	L	1128	+1
NC_003608	Hibiscus Chlorotic Ringspot Virus[f]	Coat	p25	675	+1
NC_003608	Hibiscus Chlorotic Ringspot Virus	Replicase	p23	630	+1
NC_004730	Indian Peanut Clump Virus	P14	P17	158	+1
KR732417	Influenza A Virus H5N1	PB1	PB1-F2	273	+1
NC_009025	Israel Acute Paralysis Virus Of Bees	ORF2	ORFx	285	+1
NC_003627	Maize Chlorotic Mottle Virus	Coat	p31	451	+1
NC_001498	Measles Virus[i]	P	C	561	+1
NC_005339	Mossman Virus[i]	P	C	459	+1
NC_008311	Murine Norovirus	VP1	VF1	642	+1
NC_001633	Mushroom Bacilliform Virus	ORF1	Vpg-protease	533	+1
NC_001718	Porcine Parvovirus	Capsid	SAT	207	+1
NC_001747	Potato Leafroll Virus	P0	P1	661	+1
NC_003725	Potato Mop-Top Virus[cd]	TGBp2	TGBp2	146	+1
NC_003768	Rice Dwarf Virus	Pns12	OP-ORF	276	+1
NC_003771	Rice Ragged Stunt Virus	P4b	Replicase	981	+1
NC_004718	SARS Coronavirus	Nucleocapsid	Protein I	297	+1
NC_003809	Spinach Latent Virus	Replicase	2b	308	+1
NC_003448	Striped Jack Nervous Necrosis Virus	Protein A	B2	228	+1
NC_001366	Theiler's Virus	L	L*	471	+1
NC_002199	Tupaia Paramyxovirus[i]	P	C	462	+1
NC_003743	Turnip Yellows Virus[fgh]	CP	ORF5	528	+1
NC_001409	Apple Chlorotic Leaf Spot Virus	CP	MP	317	+2
NC_001719	Arctic Ground Squirrel Hepatitis Virus	P	Capsid Precursor	158	+2
NC_001719	Arctic Ground Squirrel Hepatitis Virus	P	X	256	+2
NC_003532	Cymbidium Ringspot Virus	MP	p19	519	+2
NC_003093	Indian Citrus Ringspot Virus	CP	NABP	301	+2
NC_004178	Infectious Bursal Disease Virus[j]	VP2	VP5	404	+2
NC_001915	Infectious Pancreatic Necrosis Virus[j]	VP2	VP5	395	+2
NC_001990	Nudaurelia Capensis Beta Virus[f]	CP	Replicase	1832	+2
NC_014126	Providence Virus	p104	p130	2681	+2
NC_004366	Tobacco Bushy Top Virus	MP	RNP	698	+2
NC_004063	Turnip Yellows Mosaic Virus	Replicase	MP	1880	+2

[a] N/NSs overlaps share $\geq$ 50% sequence similarity
[b] P/L overlap predicted homologous in HMMer run
[c] TGBp2 genes share $\geq$ 50% protein sequence similarity
[d] TGBp2 genes predicted homologous Morozov and Solovyev (2003)
[e] TGBp3 genes predicted homologous Morozov and Solovyev (2003)
[f] Ancestral genes predicted homologous in HMMer run
[g] Novel genes predicted homologous in HMMer run
[h] Novel genes predicted homologous in HMMer run
[i] Novel genes predicted homologous in HMMer run
[j] Ancestral VP2 genes share $\geq$ 50% protein sequence similarity

TABLE I

Gene pairs for which the relative ages could be determined. Genes are phylogenetically independent except as noted in the footnotes.

REFERENCES

[1] Annamária F Ángyán, András Perczel, and Zoltán Gáspári. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? *FEBS Letters*, 586(16):2468–2472, 2012.

[2] Robert Belshaw, Oliver G Pybus, and Andrew Rambaut. The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research*, 17(10):1496–1504, 2007.

[3] Erich Bornberg-Bauer and M Mar Alba. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology*, 23(3):459–466, 2013.

[4] Celeste J Brown, Sachiko Takayama, Andrew M Campen, Pam Vise, Thomas W Marshall, Christopher J Oldfield, Christopher J Williams, and A Keith Dunker. Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular Evolution*, 55(1):104–110, 2002.

[5] Marija Buljan, Adam Frankish, and Alex Bateman. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biology*, 11(7):R74, 2010.

[6] José A Campillo-Balderas, Antonio Lazcano, and Arturo Becerra. Viral genome size distribution does not correlate with the antiquity of the host lineages. *Frontiers in Ecology and Evolution*, 3:143, 2015.

[7] Joseph J Carter, Matthew D Daugherty, Xiaojie Qi, Anjali Bheda-Malge, Gregory C Wipf, Kristin Robinson, Ann Roman, Harmit S Malik, and Denise A Galloway. Identification of an overprinting gene in merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. *Proceedings of the National Academy of Sciences*, 110(31):12744–12749, 2013.

[8] Anne-Ruxandra Carvunis, Thomas Rolland, Ilan Wapinski, Michael A Calderwood, Muhammed A Yildirim, Nicolas Simonis, Benoit Charloteaux, César A Hidalgo, Justin Barbette, Balaji Santhanam, et al. Proto-genes and de novo gene birth. *Nature*, 487(7407):370, 2012.

[9] Nicola Chirico, Alberto Vianelli, and Robert Belshaw. Why genes overlap in viruses. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1701):3809–3817, 2010.

[10] Wen-Yu Chung, Samir Wadhawan, Radek Szklarczyk, Sergei Kosakovsky Pond, and Anton Nekrutenko. A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Computational Biology*, 3(5):e91, 2007.

[11] John M Coffin, Stephen H Hughes, and Harold E Varmus. Principles of retroviral vector design. 1997.

[12] Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa, and István Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology*, 347(4):827–839, 2005.

[13] Julian Echave, Stephanie J Spielman, and Claus O Wilke. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2):109, 2016.

[14] Diana Ekman and Arne Elofsson. Identifying and quantifying orphan protein sequences in fungi. *Journal of Molecular Biology*, 396(2):396–405, 2010.

[15] François Ferron, Sonia Longhi, Bruno Canard, and David Karlin. A practical overview of protein disorder prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 65(1):1–14, 2006.

[16] Scott G Foy, Benjamin A Wilson, Matthew HJ Cordes, and Joanna Masel. Progressively more subtle aggregation avoidance strategies mark a long-term direction to protein evolution. *bioRxiv*, 176867, 2017.

[17] Dan Graur. *Molecular and Genome Evolution*. Sinauer Associates Inc., first edition, 2016.

[18] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.

[19] Paul K Keese and Adrian Gibbs. Origins of genes: "big bang" or continuous creation? *Proceedings of the National Academy of Sciences*, 89(20):9489–9493, 1992.

[20] Dae-Soo Kim, Chi-Young Cho, Jae-Won Huh, Heui-Soo Kim, and Hwan-Gue Cho. Evog: a database for evolutionary analysis of overlapping genes. *Nucleic Acids Research*, 37(suppl_1):D698–D702, 2008.

[21] Erika Kovacs, Peter Tompa, Karoly Liliom, and Lajos Kalmar. Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proceedings of the National Academy of Sciences*, 107(12):5429–5434, 2010.

[22] Zhirong Liu and Yongqi Huang. Advantages of proteins being disordered. *Protein Science*, 23(5):539–550, 2014.

[23] David M McCandlish and Arlin Stoltzfus. Modeling evolution using the probability of fixation: history and implications. *The Quarterly review of biology*, 89(3):225–252, 2014.

[24] Aoife McLysaght and Laurence D Hurst. Open questions in the study of de novo genes: what, how and why. *Nature Reviews Genetics*, 17(9):567, 2016.

[25] Masashi Mizokami, Etsuro Orito, Ken-ichi Ohba, Kazuho Ikeo, Johnson YN Lau, and Takashi Gojobori. Constrained evolution with respect to gene overlap of hepatitis B virus. *Journal of molecular evolution*, 44(1):S83–S90, 1997.

[26] Andrew D Moore and Erich Bornberg-Bauer. The dynamics and evolutionary potential of domain loss and emergence. *Molecular Biology and Evolution*, 29(2):787–796, 2011.

[27] Sergey Yu Morozov and Andrey G Solovyev. Triple gene block: modular design of a multifunctional machine for plant virus movement. *Journal of General Virology*, 84(6):1351–1366, 2003.

[28] Bryan A Moyers and Jianzhi Zhang. Further simulations and analyses demonstrate open problems of phylostratigraphy. *Genome biology and evolution*, 9(6):1519–1527, 2017.

[29] Y. Nakamura, T. Gojobori, and T. Ikemura. Codon usage tabulated from the international dna sequence databases: status for the year 2000. *Nucleic Acids Research*, 28:292, 2000.

[30] Tomohiro Nakayama, Satoshi Asai, Yasuo Takahashi, Oto Maekawa, and Yasuji Kasama. Overlapping of genes in the human genome. *International Journal of Biomedical Science: IJBS*, 3(1):14, 2007.

[31] Anton Nekrutenko, Samir Wadhawan, Paula Goetting-Minesky, and Kateryna D Makova. Oscillating evolution of a mammalian locus with overlapping reading frames: an xlαs/alex relay. *PLoS Genetics*, 1(2):e18, 2005.

[32] Rafik Neme and Diethard Tautz. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, 14(1):117, 2013.

[33] Lou Nielly-Thibault and Christian R Landry. Differences between the de novo proteome and its non-functional precursor can result from neutral constraints on its birth process, not necessarily from natural selection alone. *bioRxiv*, page 289330, 2018.

[34] Ann L Oberg and Douglas W Mahoney. Linear mixed effects models. In *Topics in biostatistics*, pages 213–234. Springer, 2007.

[35] Angelo Pavesi, Gkikas Magiorkinis, and David G Karlin. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the gene nursery of deltaretroviruses. *PLoS Computational Biology*, 9(8):e1003162, 2013.

[36] Corinne Rancurel, Mahvash Khosravi, A Keith Dunker, Pedro R Romero, and David Karlin. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *Journal of Virology*, 83(20):10719–10736, 2009.

[37] Sebastien Ribrioux, Adrian Brüngger, Birgit Baumgarten, Klaus Seuwen, and Markus R John. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genomics*, 9(1):122, 2008.

[38] Niv Sabath, Dan Graur, and Giddy Landan. Same-strand overlapping genes in bacteria: compositional determinants of phase bias. *Biology Direct*, 3(1):36, 2008.

[39] Niv Sabath, Giddy Landan, and Dan Graur. A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS One*, 3(12):e3996, 2008.

[40] Niv Sabath, Andreas Wagner, and David Karlin. Evolution of viral proteins originated de novo by overprinting. *Molecular Biology and Evolution*, mss179, 2012.

[41] Andrew M Sackman, Lindsey W McGee, Anneliese J Morrison, Jessica Pierce, Jeremy Anisman, Hunter Hamilton, Stephanie Sanderbeck, Cayla Newman, and Darin R Rokyta. Mutation-driven parallel evolution during viral adaptation. *Molecular Biology and Evolution*, 34(12):3243–3253, 2017.

[42] Aditi Shukla and Rolf Hilgenfeld. Acquisition of new protein domains by coronaviruses: analysis of overlapping genes coding for proteins N and 9b in SARS coronavirus. *Virus Genes*, 50(1):29–38, 2015.

[43] Etienne Simon-Loriere, Edward C Holmes, and Israel Pagán. The effect of gene overlapping on the rate of RNA virus evolution. *Molecular Biology and Evolution*, 30(8):1916–1928, 2013.

[44] R Sokal and J Rohlf. *Biometry*, chapter 14.6, 16.2, 16.5. W. H. Freeman, third edition, 1994.

[45] Arlin Stoltzfus and David M McCandlish. Mutational biases influence parallel adaptation. *Molecular biology and evolution*, 34(9):2163–2172, 2017.

[46] Arlin Stoltzfus and Lev Y Yampolsky. Climbing mount probable: mutation as a cause of nonrandomness in evolution. *Journal of Heredity*, 100(5):637–647, 2009.

[47] Nobuhiko Tokuriki, Christopher J Oldfield, Vladimir N Uversky, Igor N Berezovsky, and Dan S Tawfik. Do viral proteins possess unique biophysical features? *Trends in Biochemical Sciences*, 34(2):53–59, 2009.

[48] Vyacheslav Tretyachenko, Jiří Vymětal, Lucie Bednárová, Vladimír Kopecký, Kateřina Hofbauerová, Helena Jindrová, Martin Hubálek, Radko Souček, Jan Konvalinka, Jiří Vondrášek, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Scientific Reports*, 7(1):15449, 2017.

[49] Vladimir N Uversky and A Keith Dunker. Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1804(6):1231–1264, 2010.

[50] Vamsi Veeramachaneni, Wojciech Makalowski, Michal Galdzicki, Raman Sood, and Izabela Makalowska. Mammalian overlapping genes: the comparative perspective. *Genome Research*, 14(2):280–286, 2004.

[51] Robert G Webster, William J Bean, Owen T Gorman, Thomas M Chambers, and Yoshihiro Kawaoka. Evolution and ecology of influenza A viruses. *Microbiological Reviews*, 56(1):152–179, 1992.

[52] Benjamin A Wilson, Scott G Foy, Rafik Neme, and Joanna Masel. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature Ecology & Evolution*, 1(6):0146, 2017.

[53] Bin Xue, David Blocquel, Johnny Habchi, Alexey V Uversky, Lukasz Kurgan, Vladimir N Uversky, and Sonia Longhi. Structural disorder in viral proteins. *Chemical Reviews*, 114(13):6880–6911, 2014.

[54] Lev Y Yampolsky and Arlin Stoltzfus. Bias in the introduction of variation as an orienting factor in evolution. *Evolution & Development*, 3(2):73–83, 2001.

[55] Tong Zhou, Wanjun Gu, Jianmin Ma, Xiao Sun, and Zuhong Lu. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. *Biosystems*, 81(1):77–86, 2005.