OLGenie: Estimating Natural Selection to Predict Functional Overlapping Genes

Chase W. Nelson [1,2]*, Zachary Ardern [3] and Xinzhu Wei [4,5]

[1]Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY

[2]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan

[3]Microbial Ecology, ZIEL—Institute for Food & Health, Technische Universität München, Freising, Germany

[4]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI

[5]Department of Integrative Biology and Statistics, University of California, Berkeley, CA

*Corresponding author: E-mail: cnelson@amnh.org.

Associate editor: Koichiro Tamura

Abstract

Purifying (negative) natural selection is a hallmark of functional biological sequences, and can be detected in protein-coding genes using the ratio of nonsynonymous to synonymous substitutions per site (d_N/d_S). However, when two genes overlap the same nucleotide sites in different frames, synonymous changes in one gene may be nonsynonymous in the other, perturbing d_N/d_S. Thus, scalable methods are needed to estimate functional constraint specifically for overlapping genes (OLGs). We propose OLGenie, which implements a modification of the Wei–Zhang method. Assessment with simulations and controls from viral genomes (58 OLGs and 176 non-OLGs) demonstrates low false-positive rates and good discriminatory ability in differentiating true OLGs from non-OLGs. We also apply OLGenie to the unresolved case of HIV-1's putative *antisense protein* gene, showing significant purifying selection. OLGenie can be used to study known OLGs and to predict new OLGs in genome annotation. Software and example data are freely available at https://github.com/chasewnelson/OLGenie (last accessed April 10, 2020).

***Key words:* *antisense protein* (*asp*) gene, d_N/d_S, gene prediction, genome annotation, human immunodeficiency virus-1, open reading frame, overlapping gene (OLG), purifying (negative) selection.**

Natural selection in protein-coding genes is commonly inferred by comparing the number of nonsynonymous (amino acid changing; d_N) and synonymous (not amino acid changing; d_S) substitutions per site, with $d_N/d_S < 1$ indicative of purifying (negative) selection. Thus, d_N/d_S can be used to predict functional genes (Gojobori et al. 1982; Nekrutenko et al. 2002). However, complications arise if synonymous changes are not neutral, in which case purifying selection may reduce d_S (i.e., increase d_N/d_S). This is usually negligible, as the effects of most synonymous variants are dwarfed by those of disadvantageous nonsynonymous variants, causing the majority of genes to exhibit $d_N/d_S < 1$ (Hughes 1999; Holmes 2009). However, this assumption does not hold for overlapping genes (OLGs). A double-stranded nucleic acid may encode up to six open reading frames (ORFs), three in the sense direction and three in the antisense direction, allowing pairs of genes to overlap the same nucleotide positions in a genome (fig. 1). In such OLGs, changes that are synonymous in one gene may be nonsynonymous in the other, making otherwise "silent" variants subject to selection. As a result, d_N/d_S methods designed for regular (non-overlapping) genes do not take into account the nonsynonymous effects (in the alternate gene) of some synonymous changes (in the reference gene). As a result, standard (non-OLG) d_N/d_S methods can fail to detect purifying selection or erroneously predict positive (Darwinian) selection when applied to OLGs (Holmes et al. 2006; Sabath et al. 2008; Sabath and Graur 2010).

OLGs are widespread in viruses (Belshaw et al. 2007; Brandes and Linial 2016; Pavesi et al. 2018), and may not be uncommon in prokaryotes (Meydan et al. 2018; Vanderhaeghen et al. 2018; Weaver et al. 2019) and eukaryotes, including humans (Makałowska et al. 2007; Sanna et al. 2008). The number of OLGs has likely been underestimated, partly because genome annotation software is biased against both short ORFs (Warren et al. 2010) and overlapping ORFs (Vanderhaeghen et al. 2018). Current methods for detecting OLGs, such as Synplot2 (Firth 2014), d_N/d_S estimators (Sabath et al. 2008; Wei and Zhang 2015), and long-ORF identifiers (Schlub et al. 2018) are subject to one or more of the following limitations: restricted to long OLGs, limited to single or pairs of sequences, unsuitable for low sequence divergence, not specific to protein-coding genes, lacking accessible implementation, or too computationally intensive for genome-scale data (Table 1). For example, those available methods that are suitable for genome-scale analysis are not able to specifically detect protein-coding OLGs. Scalable bioinformatics tools are therefore needed to predict OLG candidates for further analysis, preferably by utilizing the evolutionary information available in multiple sequences and quantifying purifying selection in a way that is comparable with that of non-OLGs. We wrote OLGenie to fill this void.

Letter

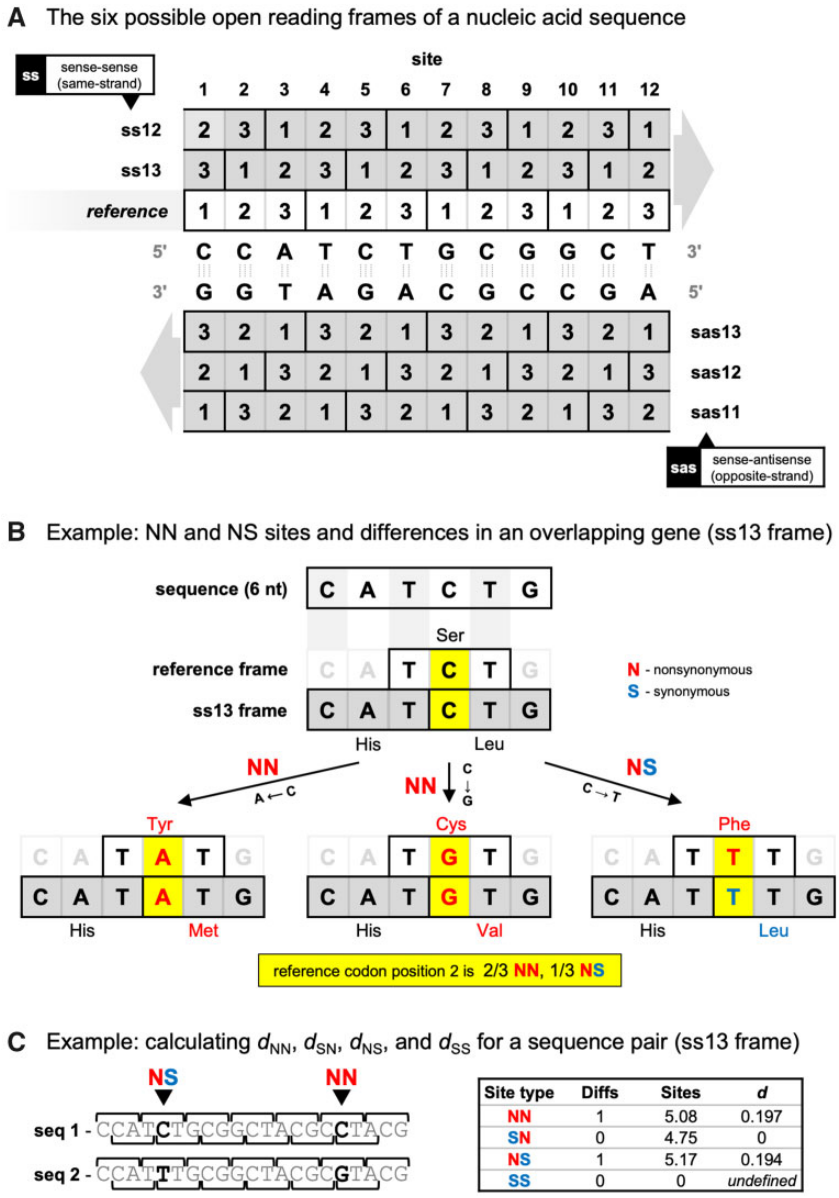

FIG. 1. Overlapping genes: reading frames and terminology. (*A*) The six possible protein-coding open reading frames (ORFs) of a double-stranded nucleic acid sequence. Codons are denoted with solid black boxes, each comprising three ordered nucleotide positions (1, 2, 3) with light gray boundaries. The reference gene frame is shown with a white background, whereas alternate gene frames are shown with a gray background. Frame relationships are indicated using the nomenclature of Wei and Zhang (2015), where "ss" indicates "sense–sense" (same-strand), "sas" indicates "sense–antisense" (opposite-strand), and the numbers indicate which codon position of the alternate gene (second number) overlaps codon position 1 of the reference gene (first number). For all alternate frames except sas13, one reference codon partially overlaps each of two alternate codons. (*B*) Example of an overlapping gene in the ss13 frame. A minimal overlapping unit of 6 nt is shown, comprising one reference gene codon and its two overlapping codons in the alternate gene. At position 2 of the reference codon (highlighted in yellow), three nucleotide changes are possible: two cause nonsynonymous changes in both genes (NN; nonsynonymous/nonsynonymous) and one causes a nonsynonymous change in the reference gene but a synonymous change in the alternate gene (NS; nonsynonymous/synonymous). No synonymous/nonsynonymous (SN) or synonymous/synonymous (SS) changes are possible at this site. Thus, this site is counted as two-thirds of an NN site and one-third of an NS site. Finally, a pair of sequences having a C/A or C/G difference at this site is counted as having 1 NN difference, whereas a pair of sequences having a C/T difference at this site is counted as having 1 NS difference. (*C*) Example calculation of d_{NN}, d_{SN}, d_{NS}, and d_{SS} for a pair of sequences with an overlapping gene in ss13. Codons are denoted with brackets above (reference gene) and below (alternate gene) each sequence. The distance *d* is calculated for each site type (NN, SN, NS, and SS) as the number of differences divided by the number of sites of that type. Because the first and last reference codons only partially overlap alternate codons, they are excluded from analysis and the numbers of sites sum to 15 (= 5 codons × 3 nt; codons 2–6). Numbers of sites are not an exact multiple of 1/3 because nucleotide 6 of sequence 2 (TTT; alternate codon TTG) does not tolerate a change to A, as this would lead to a stop codon in the alternate gene (TAG). Thus, this position is considered an SN site in sequence 1, but one-half of an NN site and one-half of an SN site in sequence 2, for a mean of 0.25 NN and 0.75 SN sites. The table shows the mean numbers of sites for the two sequences (sequence 1 = 4.33 NN, 5 SN, 5.67 NS, and 0 SS; sequence 2 = 5.83 NN, 4.5 SN, 4.67 NS, and 0 SS), used to calculate each *d* value. For a multiple sequence alignment, the mean number of differences and sites for all pairwise comparisons would be used.

Table 1. Methods with Available Implementations for Detecting Selection in Overlapping Genes.

Program[a]	Reference	Target	Implementation	Method Description	Advantages and Limitations	Available from
OLGenie	This study	Protein-coding sequence	Perl	Estimates d_N/d_S by introducing three modifications to Wei–Zhang: 1) minimal overlapping units of 6 nt, that is, 1 reference codon and 2 alternate codons; 2) the Nei–Gojobori method; and 3) only single nucleotide differences rather than all mutational pathways	Fast; applicable to multiple sequence alignments; tree-agnostic; conservative for purifying selection and high levels of divergence, but nonconservative for positive selection; loss of power for pairwise distance >0.1 and neighboring variants	https://github.com/chasewnelson/OLGenie, last accessed April 10, 2020.
"Frameshift"	Schlub et al. (2018)	Protein-coding sequence	R	Finds ORFs longer than expected by chance given nucleotide context; includes two complementary methods: "codon permutation" and "synonymous mutation"	Medium to high accessibility as an R script requiring minor modifications. Can only detect relatively long OLGs. Slow for long sequences.	https://github.com/TimSchlub/Frameshift, last accessed April 10, 2020.
"StopStatistics"	Cassan et al. (2016)	Protein-coding sequence	Python, bash	Tests for depletion of those stop codons in sas12 that would be synonymous in reference; also applicable to enrichment of start codons	Low accessibility; scripts specific to particular data sets	https://figshare.com/s/9668ef62e84488-d4787a, last accessed April 10, 2020.
FRESCo	Sealfon et al. (2015)	Constraint at synonymous sites	HYPHY batch language	Rates of nucleotide evolution across an alignment inferred using a maximum-likelihood model. Models of neutral and nonneutral evolution tested in sliding windows to infer regions with excess synonymous constraint	Suitable for short genomes/regions despite using a codon model; requires a phylogenetic tree; performs best at deep sequence coverage and increased sequence divergence	https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-015-0603-7/MediaObjects/13059_2015_603_MOESM1_ESM.zip, last accessed April 10, 2020.
Wei–Zhang method	Wei and Zhang (2015)	Protein-coding sequence	Perl	Estimates d_N/d_S in minimal-length coding regions flanked by variant-free codons (i.e., data-dependent minimal overlapping units) to determine the effects of all mutational pathways in the reference and alternate genes using the modified Nei–Gojobori method	Accurate but slow, especially for highly diverged sequences; tree-agnostic; outperforms Sabath et al. method (according to Wei and Zhang [2015]); only implemented for pairs of sequences; low accessibility and scalability	http://www.umich.edu/~zhanglab/download/Xinzhu_GBE2014/index.htm, last accessed April 10, 2020.
Synplot2	Firth (2014)	Constraint at synonymous sites	C++; Web-interface	Evolution at synonymous sites in a codon alignment compared to a null model of neutral evolution in order to infer sites with excess constraint; expected diversity at synonymous sites is set equal to diversity over the full alignment, and diversity is measured between sequential pairs around a phylogenetic tree	Medium accessibility; fast; limited use in the case of sas12; requires a phylogenetic tree; does not distinguish between coding and noncoding overlapping features	http://guinevere.otago.ac.nz/cgi-bin/aef/synplot.pl, last accessed April 10, 2020 http://www.firthlab.path.cam.ac.uk/SynPlot2.zip, last accessed April 10, 2020.
KaKi ("Multilayer")	Rubinstein et al. (2011)		C++	Maximum-likelihood codon model approach that allows variation in both the	Low accessibility (requires an old Linux distribution to install); requires a	https://www.tau.ac.il/~talp/

(continued)

Table 1 Continued

Program[a]	Reference	Target	Implementation	Method Description	Advantages and Limitations	Available from
		Unexpected variation at synonymous sites		synonymous and nonsynonymous substitution rates along a sequence; accounting for variability in the baseline substitution rate allows more reliable inference of positive selection	phylogenetic tree; complex input and results; focus of explicit testing is on positive selection; applicable (but not specific) to protein-coding OLGs.	multilayer.tar.gz, last accessed April 10, 2020. https://www.tau.ac.il/~talp/readme.txt, last accessed April 10, 2020.
Sabath et al. method	Sabath et al. (2008)	Protein-coding sequence	MATLAB	Maximum-likelihood framework for estimating d_N/d_S; similar to the (nonimplemented) method of Pedersen and Jensen (2001)	Slower than Wei–Zhang; not recommended for highly similar sequences (pairwise distance <0.08); similar to OLGenie in the use of 6 nt ("sextet") units; only implemented for pairs of sequences; low accessibility and scalability	http://nsmn1.uh.edu/dgraur/Software.html, last accessed April 10, 2020.
MLOGD	Firth and Brown (2006)	Protein-coding sequence	C++	Simple statistics on properties of sequence variation by codon position, and a maximum-likelihood statistic (MLOGD) taking into account nucleotide and amino acid substitution rates and codon usage	Less sensitive at detecting OLGs than Synplot2 (according to Firth [2014]); requires a minimum of ~20 independent nucleotide variants; sas12 frame generates false-positives.	http://guinevere.otago.ac.nz/aef/MLOGD/software.html, last accessed April 10, 2020.

[a]Programs in descending order by year of publication; methods lacking implementations at active URLs are not listed.

New Approaches

OLGenie is executed at the Unix/Linux command line with two inputs: 1) a multiple sequence alignment (FASTA file) of contiguous codons known or hypothesized to constitute an OLG pair; and 2) the frame relationship of the OLGs. The codon frame beginning at site 1 of the alignment is considered the "reference" gene, which overlaps one "alternate" gene. The choice of which gene to consider the reference versus the alternate is arbitrary; however, in practice, the reference gene ORF is typically longer, whereas the alternate gene ORF usually occurs entirely or partially within the reference gene, and is of unknown or more recently established functionality (Pavesi et al. 2018). The alternate gene can occur in any one of five frames: ss12, ss13, sas11, sas12, or sas13. Here, "ss" indicates "sense–sense" (same-strand), "sas" indicates "sense–antisense" (opposite-strand), and the numbers indicate which codon position of the alternate gene (second number) overlaps codon position 1 of the reference gene (first number) (fig. 1). We prefer this nomenclature because the meaning of each frame is described in its name; however, at least nine others have been employed, summarized in Table 2.

OLGenie estimates d_N and d_S in OLGs by modifying the method of Wei and Zhang (2015). Four expanded d_N and d_S measures are used: d_{NN}, d_{SN}, d_{NS}, and d_{SS}, where the first subscript refers to the reference gene and the second subscript refers to the alternate gene (NN, nonsynonymous/nonsynonymous; SN, synonymous/nonsynonymous; NS, nonsynonymous/synonymous; SS, synonymous/synonymous). For example, d_{NS} refers to the mean number of nucleotide substitutions per site that are nonsynonymous in the reference gene but synonymous in the alternate gene (NS). Given these values, d_N/d_S may be estimated for the reference gene as d_{NN}/d_{SN} or d_{NS}/d_{SS}, or for the alternate gene as d_{NN}/d_{NS} or d_{SN}/d_{SS}. In each case, the effect of mutations in one of the two OLGs is held constant (N or S), ensuring a "fair comparison" in the other gene. For example, if nonsynonymous changes observed in the reference gene are disproportionately synonymous in the alternate gene ($d_{NS} > d_{NN}$), the result will be $d_{NN}/d_{NS} < 1.0$, and purifying selection on the alternate gene can be inferred (Hughes and Hughes 2005). In practice, d_{NN}/d_{NS} rather than d_{SN}/d_{SS} is typically used to test for selection in the alternate gene, as SS sites are usually too rare to allow a reliable estimate of d_{SS}.

The original Wei–Zhang method is computationally prohibitive when many nucleotide variants are present in neighboring codons, and the size of the minimal bootstrap unit is data-dependent (Table 1). To circumvent these issues, we introduce three modifications: 1) consider each reference codon to be an independent unit of the alignment amenable to bootstrapping; 2) apply the Nei–Gojobori method to each OLG, as implemented in SNPGenie (Nei and Gojobori 1986; Nelson and Hughes 2015; Nelson et al. 2015); and 3) consider only single nucleotide differences, rather than all mutational pathways, that is, a given nucleotide change to a given codon either does (synonymous) or does not (nonsynonymous) encode the same amino acid. Modification (1) is not strictly true

Table 2. Nomenclature for Overlapping Protein-Coding Reading Frames.

Study[a]	Frame[b]					
	5′-123123-3′ 5′-123123-3′	5′- 123 -3′ 5′-123123-3′	5′- 123 -3′ 5′-123123-3′	5′- 123 -3′ 3′-321321-5′	5′- 123 -3′ 3′-321321-5′	5′-123123-3′ 3′-321321-5′
OLGenie; Wei and Zhang (2015)	**Reference (ss11)**	**ss12**	**ss13**	**sas11**	**sas12**	**sas13**
Scherer et al. (2018)	**+1**	**+3**	**+2**	**−3**	**−2**	**−1**
Lèbre and Gascuel (2017)	**+0**	**+2**	**+1**	**−1**	**−2**	**−0**
Schlub et al. (2018)	**+0**	**+2**	**+1**	**−c2**	**−c1**	**−c0**
Sabath et al. (2008)	**0**	**2 (same-strand)**	**1 (same-strand)**	**1 (opposite-strand)**	**2 (opposite-strand)**	**0 (opposite-strand)**
Belshaw et al. (2007)	**0**	**−1**	**+1**	**rc-1**	**rc+1**	**rc0**
Firth and Brown (2006)[c]	**0**	**+2**	**+1**	**−1**	**−2**	**−3**
Rogozin et al. (2002)[c]	**−**	**−**	**−**	**C1**	**C3**	**C2**
Krakauer (2000)[c]	**−**	**+2**	**+1**	**−1**	**0**	**−2**
Smith and Waterman (1980)[c]	**0**	**2**	**1**	**5**	**3**	**4**

[a]Studies in descending order by year of publication.
[b]Black denotes the reference frame and blue denotes the alternate frame; one alternate codon position is underlined to show overlap with reference codon position 1 (e.g., position 3 for ss13).
[c]As reported by Lèbre and Gascuel (2017).

when two neighboring reference codons share sites with the same alternate codon, introducing biological nonindependence. Nevertheless, no individual site is included in more than one unit of the alignment, and the assumption of independence has proven widely effective (Nei and Kumar 2000), even though nearby codons may never evolve independently. Modification (3) is identical to the original Wei–Zhang method when a pair of sequences contains only one difference in contiguous codons. However, differences may be misclassified when $\geq$2 sites in contiguous codons differ. As a result, OLGenie tends to underestimate the denominator of d_N/d_S (d_{NS} or d_{SN}), biasing the ratio upward and yielding a conservative test of purifying selection that nevertheless has increased power over non-OLG d_N/d_S (supplementary section S1, Supplementary Material online).

Results and Discussion

Assessment with Simulated Data

To evaluate OLGenie when selection dynamics are known, we first performed simulation experiments for each frame across a range of d_N/d_S values, setting sequence divergence to that observed in our positive controls (median 0.0585; supplementary fig. S1, Supplementary Material online). Calibration plots reveal that OLGenie produces relatively accurate estimates, especially for purifying selection, improving substantially for lower sequence divergence (supplementary fig. S2, Supplementary Material online) and suffering minimally at higher transition/transversion ratios (supplementary fig. S3, Supplementary Material online). However, three biases are noteworthy: 1) except for frame sas12, d_N/d_S is always overestimated; 2) except for sas12, d_N/d_S overestimation increases when the OLG is under stronger purifying selection; and 3) for sas12, d_N/d_S is somewhat underestimated for the OLG when $d_N/d_S \geq 1$ (fig. 2 and supplementary tables S1 and S2, Supplementary Material online). Bias (1) is mainly explained by modification (3) in the previous section. Bias (2) is explained by the failure to account for unobserved changes (multiple hits), for which no known correction is applicable to OLGs (Hughes et al. 2005); this causes the disproportionate underestimation of the denominator (d_{NS} or d_{SN}) in the presence of purifying selection. Bias (3) may be due to the preponderance of "forbidden" codon combinations in sas12 (Lèbre and Gascuel 2017), which must necessarily be avoided to prevent STOP codons in the overlapping frame, leading to the overestimation of NN sites and underestimation of d_{NN}. Additionally, our observations may be partly attributable to the fact that avoided STOP codons (TAA, TAG, and TGA) are AT-rich, implicitly favoring high GC content and biasing codon usage in OLGs (supplementary fig. S4 and table S3, Supplementary Material online) (Pavesi et al. 2018). Finally, for all frames, bias and variance for a given gene are highest when the alternate gene is under purifying selection.

Our simulations also allowed us to identify the most accurate and precise ratios for estimating each frame's d_N/d_S. For ss12/ss13, sas11, and sas13, the rarest site class is SS (0–2.7% of sites), leading to high stochastic error when estimating d_{SS}. Thus, for alternate genes in these frames, the d_{NN}/d_{NS} ratio is relatively "site-rich" and preferred. Contrarily, for sas12, SS sites are usually more common (18.3%) than NS (7.4%) and SN (7.4%) sites, so that d_{NN}/d_{NS} is preferred only 52.5% of the time (51.2–53.9%; binomial 95% C.I.) (supplementary tables S4 and S5 and figs. S5 and S6, Supplementary Material online). Thus, for alternate genes in sas12, either ratio can potentially be informative, and should be selected on a case-by-case basis, according to the number of sites: d_{NN}/d_{NS} if the minimum of (NN, NS) $\gg$ minimum of (SN, SS); d_{SN}/d_{SS} if the inequality is reversed; or both if the minima are approximately equal.

Assessment with Biological Controls

To evaluate OLGenie's performance with real biological data, we next applied the program to 58 known OLG (positive control) and 176 non-OLG (negative control) loci from viral genomes (Pavesi et al. 2018). Strict codon alignments were generated from quality-filtered BlastN hits (Materials and Methods). OLGenie results are 73% accurate ($\alpha = 0.05$), with receiver operating characteristic curves yielding an area under the curve (AUC) of 0.70 for the full data set (supplementary table S6, Supplementary Material online). AUC

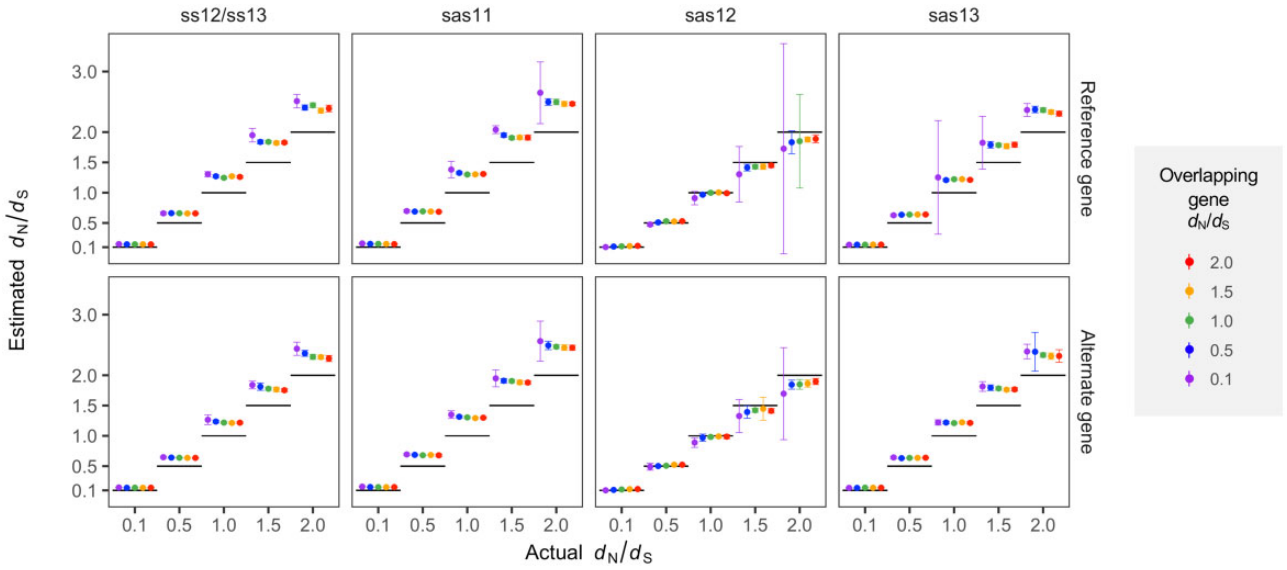

FIG. 2. Assessment of OLGenie using simulated sequences. Calibration plots show the accuracy and precision of OLGenie d_N/d_S estimates for the reference (top row; d_{NN}/d_{SN}) and alternate (bottom row; d_{NN}/d_{NS}) genes when mean pairwise distance is set to 0.0585 per site (median of biological controls). For each frame relationship, estimated d_N/d_S is shown as a function of the actual simulated value, indicated by horizontal black line segments (x axis values), and of the d_N/d_S value of the overlapping gene, indicated by color (left to right: purple = 0.1; blue = 0.5; green = 1.0; orange = 1.5; and red = 2.0). For example, all purple points in the top row refer to simulations with alternate gene $d_N/d_S = 0.1$, whereas all purple points in the bottom row refer to simulations with reference gene $d_N/d_S = 0.1$. To obtain highly accurate point estimates, each parameter combination (reference d_N/d_S, alternate d_N/d_S, frame) was simulated using 1,024 sequences of 100,000 codons (supplementary table S1, Supplementary Material online). Then, to obtain practical estimates of variance relevant to real OLG data, simulations were again carried out for each parameter combination so as to emulate our biological control data set: a sample size of 234, with sequence lengths (number of codons) and numbers of alleles (max 1,024) randomly sampled with replacement from the controls (supplementary table S2, Supplementary Material online). Error bars show SEM, estimated from replicates with defined d_N/d_S values ($\leq$234) using 10,000 bootstrap replicates (reference codon unit). A transition/transversion ratio (R) of 0.5 (equal rates) was used; similar results are obtained using $R = 2$ (supplementary fig. S3, Supplementary Material online). Full simulation results are presented in supplementary tables S1–S6 and figures S1–S6, Supplementary Material online.

increases marginally for longer sequences and drastically for lower d_N/d_S values, reaching AUC = 1.0 for $d_N/d_S \leq 0.2$ (fig. 3 and supplementary tables S7 and S8, Supplementary Material online). Results are comparable even with less strict alignment criteria (supplementary figs. S7 and S8, tables S9–S12, and section S3, Supplementary Material online). Importantly, these results may underestimate OLGenie's performance, as our data set included more negative than positive controls, and negative controls may include unannotated OLGs. For example, four negative controls of length 204–2,664 nt exhibit $d_N/d_S < 0.2$, warranting investigation (supplementary table S6, Supplementary Material online). Finally, performance would likely improve for curated alignments limited to carefully defined taxonomic groups.

Case Study: HIV-1's Putative *Antisense Protein* Gene

Lastly, we examined the unresolved case of human immunodeficiency virus-1's (HIV-1) *env*/*asp* sas12 overlap (Miller 1988; Torresilla et al. 2015), where the putative *antisense protein* (*asp*) gene has evaded detection by several bioinformatic methods, including non-OLG d_N/d_S (Cassan et al. 2016; Schlub et al. 2018). We used OLGenie to test for purifying selection in three subregions of *env*: 1) 5′ non-OLG; 2) putative *asp*-encoding; and 3) 3′ non-OLG. Three data sets were used: 1) M group from Cassan et al. (2016) (1,723 codons × 23,831 sequences; functional *asp* hypothesized); 2) non-M groups from Cassan et al. (1,723 codons × 92 sequences; no functional *asp* hypothesized); and 3) HIV-1 BLAST hits for *env* using the same methods as our control data set (1,355 codons × 4,646 sequences). We employed d_{NN}/d_{NS} for the alternate gene, as this ratio is by far the most site-rich for all *env* frames (i.e., sas12 site counts: NN = 2,127.2 and NS = 825.3, vs. SN = 190.1 and SS = 636.4; supplementary table S13, Supplementary Material online).

The sas12 d_N/d_S ratio is significantly <1 in all three data sets for the 5′ non-OLG ($d_N/d_S \leq 0.66$; $P = 2.04 \times 10^{-7}$) and *asp* ($d_N/d_S \leq 0.58$; $P = 2.75 \times 10^{-5}$) subregions of *env*. The lowest ratio for each data set always occurs in *asp*, reaching very high significance in the BLAST data set ($d_N/d_S = 0.29$; $P = 5.04 \times 10^{-25}$). As a benchmark, our ss12/ss13 controls suggest a false-positive rate of 0% for $d_N/d_S \leq 0.4$ when employing $P \leq 1.04 \times 10^{-6}$ (based on 28 OLGs and 27 non-OLGs). The 3′ non-OLG region is also significant for the Cassan non-M groups ($d_N/d_S = 0.78$, $P = 0.00921$); however, the expected false-positive rate is high (~22–28%) and the other two data sets are not significant in this region ($d_N/d_S \geq 0.74$; $P \geq 0.107$) (supplementary table S14, Supplementary Material online).

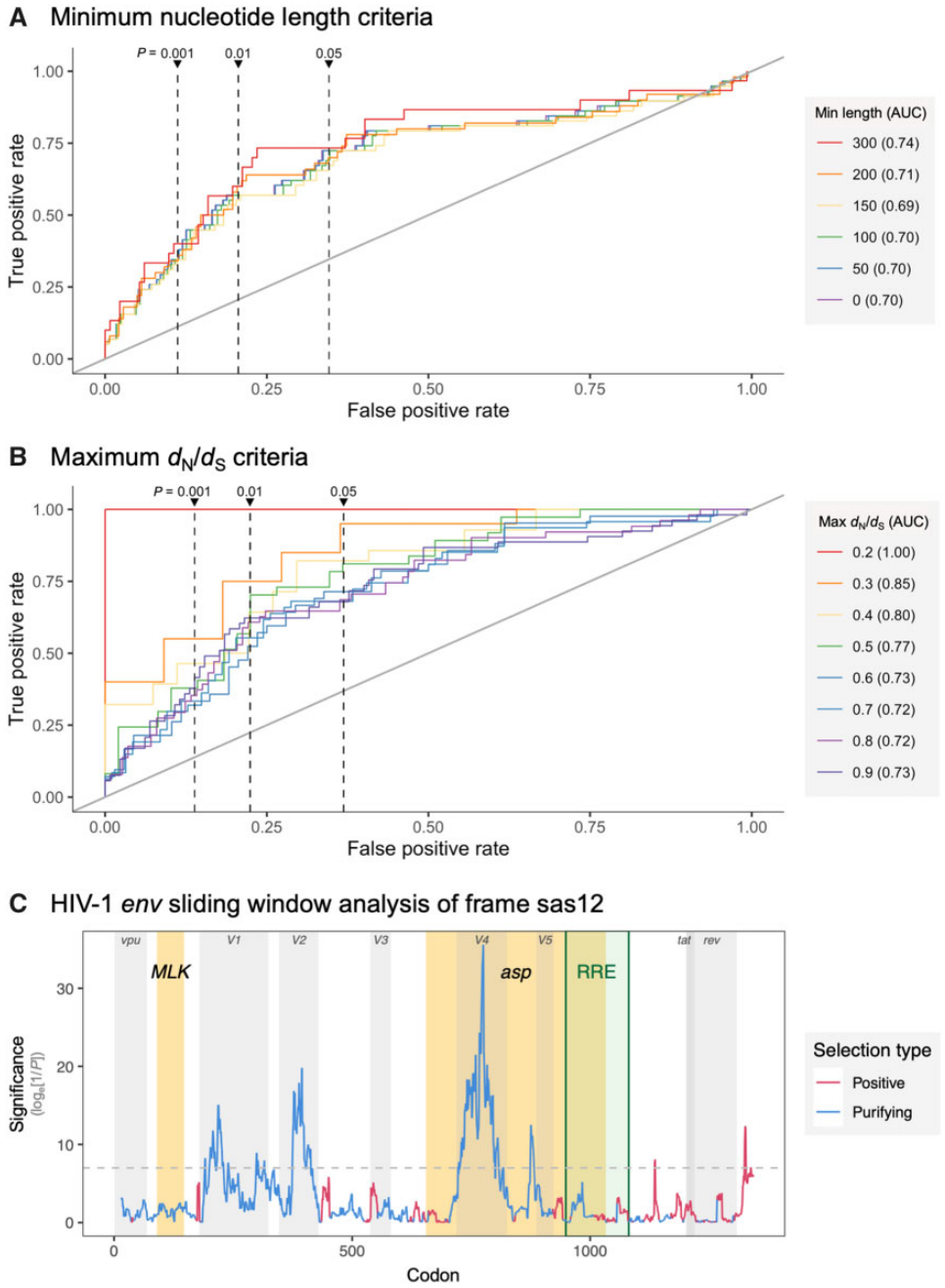

Fig. 3. Assessment of OLGenie using biological controls. (*A* and *B*) Receiver operating characteristic (ROC) curves for overlapping (alternate) gene prediction at varying *P* value cut-offs. The *y* axis shows the true-positive rate (sensitivity) and the *x* axis shows the false-positive rate (1−specificity). Curves show subsets of the data corresponding to differing minimum length (*A*) and maximum d_N/d_S (*B*) criteria, following the approach of Schlub et al. (2018), with red indicating the strictest criteria. The full data set is represented by purple in (*A*) (overlaps blue). Area under the curve (AUC) is reported in parentheses in the key (supplementary tables S6–S8, Supplementary Material online), and the ROC expected using random classification (AUC = 0.5) is shown as a diagonal gray line. Vertical dashed lines show mean false-positive rates for *P* value cut-offs of 0.001, 0.01, and 0.05 (left to right). The site-rich d_{NN}/d_{NS} ratio was used to analyze 234 controls (81 ss12 and 153 ss13): 58 positive (16 ss12 and 42 ss13) and 176 negative (65 ss12 and 111 ss13). Of these, 162 (30 positive, 132 negative) had length ≥ 300 nt, and 14 (10 positive, 4 negative) had $d_N/d_S \leq 0.2$. (*C*) The HIV-1 *env* gene was analyzed in sas12 with the site-rich ratio d_{NN}/d_{NS} using 25-codon sliding windows (step size = 1 codon), limiting to codons with ≥6 defined (nongap) sequences. The hypothesized *asp* gene is located at codons 655–1,033 (supplementary table S15, Supplementary Material online). The *y* axis shows significance, calculated as the natural logarithm of the inverse *P* value, as suggested by Firth (2014), using *Z* tests of the null hypothesis that $d_{NN} = d_{NS}$ (1,000 bootstrap replicates per window; reference codon unit). The horizontal dashed gray line shows the multiple comparisons *P* value threshold (0.000924) suggested by Meydan et al. (2019) and described in supplementary section S5, Supplementary Material online, that is, a threshold of 0.05/(CDS length/window size). Results for other frames are shown in supplementary figure S9, Supplementary Material online. Positive selection (red) refers to $d_N/d_S > 1$; purifying selection (blue) refers to $d_N/d_S < 1$. Sequence features are described in supplementary table S15, Supplementary Material online and shown here as shaded rectangles: yellow for hypothesized sas12 genes, green for the highly structured RNA Rev response element (RRE), and gray otherwise.

To test whether our results are an artifact of other sequence features, including the highly structured RNA Rev response element (RRE; supplementary table S15, Supplementary Material online; Fernandes et al. 2012), we also used OLGenie to perform sliding window analyses. Results show that purifying selection in the sas12 frame of *env* is most significant in regions of *asp* not overlapping the RRE (fig. 3C). The strongest evidence is observed in variable region 4, suggesting that accepted nonsynonymous changes in this region are disproportionately synonymous in *asp*. Significance is also attained in the correct frame for the two known ss12 OLGs, *vpu* and *rev* (supplementary figs. S9 and S10, Supplementary Material online). Thus, OLGenie specifically detects protein-coding function in all three data sets. Contrarily, Synplot2 shows the strongest evidence for synonymous constraint in the RRE, likely due to RNA structure rather than protein-coding function, and fails to detect *vpu* in the BLAST data set (supplementary fig. S11, Supplementary Material online). It should be noted that these OLGenie results concern the sas12 frame, for which the d_{NN}/d_{NS} ratio is not always conservative (fig. 2), and that our biological controls were limited to the ss12 and ss13 frames. Nevertheless, our results provide evidence that purifying selection acts on the sas12 protein-coding frame of *env*, particularly in the *asp* region. This finding is corroborated by recent laboratory evidence demonstrating expression of ASP in multiple infected cell lines, where it localizes to both the host cell membrane and viral envelope upon activation of HIV-1 expression (Affram et al. 2019). This suggests ASP as a potential drug target, for which our sliding window results may be useful for identifying functionally constrained residues, that is, regions with low and highly significant d_N/d_S (fig. 3C and supplementary figs. S9 and S10 and supplementary data, Supplementary Material online).

Conclusions

OLGenie provides a simple, accessible, and scalable method for estimating d_N/d_S in OLGs. It utilizes a well-understood measure of natural selection that is specific to protein-coding genes, making it possible to directly compare functional constraint between OLGs and non-OLGs. Moreover, although its estimates of constraint are conservative, its discriminatory ability exceeds that of other methods (Schlub et al. 2018). Power is greatest at relatively low levels of sequence divergence, and may be increased in the future by incorporating mutational pathways or comparing conservative versus radical nonsynonymous changes. Even so, not all functional genes exhibit detectable selection, so that some OLGs are likely to be missed by any selection-based method. Nevertheless, because candidate OLGs are usually subject to costly downstream laboratory analyses, minimizing the false-positive rate is paramount. To this end, OLGenie achieves a false-positive rate of 0% for several subsets of our control data, for example, regions with $d_N/d_S < 0.4$ and $P \leq 1.04 \times 10^{-6}$. OLGenie can therefore be used to predict OLG candidates with high confidence, allowing researchers to begin studying evolutionary evidence for OLGs at the genomic scale.

Materials and Methods

OLGenie is written in Perl with no dependencies, and is freely available at https://github.com/chasewnelson/OLGenie (last accessed April 10, 2020). Estimates of d are obtained by calculating $d_{NN} = m_{NN}/L_{NN}$, $d_{SN} = m_{SN}/L_{SN}$, $d_{NS} = m_{NS}/L_{NS}$, or $d_{SS} = m_{SS}/L_{SS}$, where m is the mean number of differences and L is the mean number of sites between all allele pairs at each reference codon. Simulation scripts were modified from Wei and Zhang (2015). Biological control gene coordinates were obtained from Pavesi et al. (2018) and used to retrieve nucleotide sequences from the latest NCBI genome. Homologous sequences were obtained using BlastN (Altschul et al. 1990); excluded if they contained in-frame STOP codons or were <70% of query length (Hughes et al. 2005); translated using R Biostrings (Pagès et al. 2019); aligned using MAFFT v.7.150b (Katoh and Standley 2013); codon-aligned using PAL2NAL v14 (Suyama et al. 2006); and filtered to exclude redundant alleles. Only codon positions with ≥6 defined (nongap) sequences were used for estimating d_N/d_S (Jordan and Goldman 2012). Statistical analyses were carried out in R v3.5.2 (R Core Team 2018). Significant deviations from $d_N - d_S = 0$ were detected using Z tests after estimating the SE using 10,000 and 1,000 bootstrap replicates for genes and sliding windows, respectively (reference codon unit). Complete methods, results, and data are available in the Supplementary Material online and Zenodo at https://doi.org/10.5281/zenodo.3575391 (last accessed April 10, 2020).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online, with additional data available at Zenodo, https://doi.org/10.5281/zenodo.3575391 (last accessed April 10, 2020).

Acknowledgments

This work was supported by a Gerstner Scholars Fellowship from the Gerstner Family Foundation at the American Museum of Natural History and a Postdoctoral Research Fellowship from Academia Sinica to C.W.N., and by funding from the Bavarian State Government and National Philanthropic Trust to Z.A. The authors thank Wen-Hsiung Li and two anonymous reviewers for feedback on earlier drafts of this article; Ming-Hsueh Lin for dramatic improvements to our visuals; and Reed A. Cartwright, Dan Graur, Robert S. Harbert, Jim Hussey, Chen-Hao Kuo, Michael Lynch, Sergios Orestis-Kolokotronis, Apurva Narechania, Siegfried Scherer, Sally Warring, Jeff Witmer, Meredith Yeager, Jianzhi (George) Zhang, Martine Zilversmit, and the Sackler Institute for Comparative Genomics workgroup for discussion.

Author Contributions

C.W.N. conceived and wrote OLGenie, obtained control data, processed data, performed OLGenie analyses, created figures, modified the simulation software, and drafted the article. Z.A. obtained and processed HIV-1 example data and performed Synplot2 analyses. X.W. wrote the simulation software and

advised on statistical and sequence analyses. C.W.N. and X.W. developed statistical methods. All authors designed the study, analyzed and interpreted data, and revised the article.

References

Affram Y, Zapata JC, Gholizadeh Z, Tolbert WD, Zhou W, Iglesias-Ussel MD, Pazgier M, Ray K, Latinovic OS, Romerio F. 2019. The HIV-1 antisense protein ASP is a transmembrane protein of the cell surface and an integral protein of the viral envelope. *J Virol*. 93(21):e00574–e00619.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215(3):403–410.

Belshaw R, Pybus OG, Rambaut A. 2007. The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res*. 17(10):1496–1504.

Brandes N, Linial M. 2016. Gene overlapping and size constraints in the viral world. *Biol Direct*. 11(1):26.

Cassan E, Arigon-Chifolleau A-M, Mesnard J-M, Gross A, Gascuel O. 2016. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. *Proc Natl Acad Sci U S A*. 113(41):11537–11542.

Fernandes J, Jayaraman B, Frankel A. 2012. The HIV-1 Rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNA Biol*. 9(1):6–11.

Firth AE. 2014. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res*. 42(20):12425–12439.

Firth AE, Brown CM. 2006. Detecting overlapping coding sequences in viral genomes. *BMC Bioinformatics* 7(1):75.

Gojobori T, Li W-H, Graur D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *J Mol Evol*. 18(5):360–369.

Holmes EC. 2009. The evolution and emergence of RNA viruses. New York: Oxford University Press.

Holmes EC, Lipman DJ, Zamarin D, Yewdell JW. 2006. Comment on "Large-scale sequence analysis of avian influenza isolates". *Science* 313(5793):1573b.

Hughes AL. 1999. Adaptive evolution of genes and genomes. New York: Oxford University Press.

Hughes AL, Ekollu V, Friedman R, Rose JR. 2005. Gene family content-based phylogeny of prokaryotes: the effect of criteria for inferring homology. *Syst Biol*. 54(2):268–276.

Hughes AL, Hughes M. 2005. Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. *Virus Res*. 113(2):81–88.

Jordan G, Goldman N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*. 29(4):1125–1139.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Krakauer DC. 2000. Stability and evolution of overlapping genes. *Evolution* 54(3):731–739.

Lèbre S, Gascuel O. 2017. The combinatorics of overlapping genes. *J Theor Biol*. 415:90–101.

Makałowska I, Lin C-F, Hernandez K. 2007. Birth and death of gene overlaps in vertebrates. *BMC Evol Biol*. 7(1):193.

Meydan S, Marks J, Klepacki D, Sharma V, Baranov PV, Firth AE, Margus T, Kefi A, Vázquez-Laslop N, Mankin AS. 2019. Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Mol Cell*. 74(3):481–493.

Meydan S, Vázquez-Laslop N, Mankin AS. 2018. Genes within genes in bacterial genomes. *Microbiol Spectrum*. 6:RWR-0020-2018.

Miller RH. 1988. Human inmunodeficiency virus may encode a novel protein on the genomic DNA plus strand. *Science* 239(4846):1420–1422.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 3(5):418–426.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Nekrutenko A, Makova KD, Li W-H. 2002. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res*. 12(1):198–202.

Nelson CW, Hughes AL. 2015. Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. *Infect Genet Evol*. 30:1–7.

Nelson CW, Moncla LH, Hughes AL. 2015. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics* 31(22):3709–3711.

Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2019. Biostrings: efficient manipulation of biological strings. R package version 2.50.2. Available from: https://www.bioconductor.org/packages//2.7/bioc/html/Biostrings.html. Accessed April 10, 2020.

Pavesi A, Vianelli A, Chirico N, Bao Y, Blinkova O, Belshaw R, Firth A, Karlin D. 2018. Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS One* 13(10):e0202513.

Pedersen A-M, Jensen JL. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol Biol Evol*. 18(5):763–776.

R Core Team. 2018. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: https://www.R-project.org/. Accessed April 10, 2020.

Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV. 2002. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet*. 18(5):228–232.

Rubinstein ND, Doron-Faigenboim A, Mayrose I, Pupko T. 2011. Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. *Mol Biol Evol*. 28(12):3297–3308.

Sabath N, Graur D. 2010. Detection of functional overlapping genes: simulation and case studies. *J Mol Evol*. 71(4):308–316.

Sabath N, Landan G, Graur D. 2008. A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS One* 3(12):e3996.

Sanna CR, Li W-H, Zhang L. 2008. Overlapping genes in the human and mouse genomes. *BMC Genomics* 9(1):169.

Scherer S, Neuhaus K, Bossert M, Mir K, Keim D, Simon S. 2018. Finding new overlapping genes and their theory (FOG theory). In: Bossert M, editor. Information- and communication theory in molecular biology. Cham (Switzerland): Springer. p. 137–159.

Schlub TE, Buchmann JP, Holmes EC. 2018. A simple method to detect candidate overlapping genes in viruses using single genome sequences. *Mol Biol Evol*. 35(10):2572–2581.

Sealfon RS, Lin MF, Jungreis I, Wolf MY, Kellis M, Sabeti PC. 2015. FRESCo: finding regions of excess synonymous constraint in diverse viruses. *Genome Biol*. 16(1):38.

Smith TF, Waterman MS. 1980. Protein constraints induced by multiframe encoding. *Math Biosci*. 49(1–2):17–26.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 34(Web Server):W609–W612.

Torresilla C, Mesnard J-M, Barbeau B. 2015. Reviving an old HIV-1 gene: the HIV-1 antisense protein. *Curr HIV Res*. 13(2):117–124.

Vanderhaeghen S, Zehentner B, Scherer S, Neuhaus K, Ardern Z. 2018. The novel EHEC gene *asa* overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase. *Sci Rep*. 8(1):17875.

Warren AS, Archuleta J, Feng W-C, Setubal JC. 2010. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics*. 11(1):131.

Weaver J, Mohammad F, Buskirk AR, Storz G. 2019. Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio*. 10(2):e02819–e02918.

Wei X, Zhang J. 2015. A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biol Evol*. 7(1):381–390.