Biological Sciences – Medical Sciences

Potential Achilles heels of SARS-CoV-2 displayed by the base order-dependent component of RNA folding energy

Chiyu Zhang[a], Donald R. Forsdyke[b*]

[a]Shanghai Public Health Clinical Center, Fudan University, Shanghai, China

[b]Department of Biomedical and Molecular Sciences, Queen's University, Kingston, Ontario, Canada K7L3N6

*Corresponding author. forsdyke@queensu.ca

Keywords: Sequence conservation | FORS-D analysis | HIV-1 | genome level function | RNA packaging signal | RNA structure | stem-loop potential

Pages: 21 Words: 5710 References: 42

Figures: 5 Tables 0

The energetics of the folding of a single-stranded nucleic acid into a stem-loop structure depend on both the composition and order of its bases. Composition tends to reflect *genome-wide* evolutionary pressures. Order better reflects *local* pressures. Base order is likely to be conserved when encoding a function critical for survival. The base order-dependent component of the folding energy has shown that a highly conserved region in HIV-1 genomes associates with an RNA structure. This corresponds to a packaging signal that is specifically recognized by the nucleocapsid domain of the Gag polyprotein. Long viewed as a potential HIV-1 "Achilles heel," the signal can be targeted by a recently described antiviral compound (NSC 260594) or by synthetic oligonucleotides. Thus, a conserved base-order-rich region of HIV-1 may facilitate therapeutic attack. Although SARS-CoV-2 differs in many respects from HIV-1, the same technology displays regions with a high base order-dependent folding energy component, which are also highly conserved. This indicates structural invariance (SI) sustained by natural selection. While the regions are often also protein-encoding (e.g. NSP3, ORF3a), we suggest that their nucleic acid level functions, such as the ribosomal frameshifting element (FSE) that facilitates differential expression of 1a and 1ab polyproteins, can be considered potential "Achilles heels" for SARS-CoV-2 that should be susceptible to therapies like those envisaged for AIDS. The region of the FSE scored well, but higher SI scores were obtained in other regions, including those encoding NSP13 and the nucleocapsid (N) protein.

Significance

Mutation rates of microbial pathogens are generally higher than those of their hosts. While a microbe spreading from host-to-host can "assume" that it will face a succession of broadly similar challenges, those hosts cannot likewise "assume" that new microbial invaders will remain as they were in previous hosts. Thus, immune defenses may be overwhelmed. The therapeutic challenges are (1) to locate a conserved, less-variable, part of a pathogen's genome, (2) identify the corresponding function, and (3) devise its effective inhibition. While often assumed to associate with protein-encoding functions, studies of the AIDS virus

genome have identified RNA structure itself as both functional and conserved, so signifying vulnerability. Application of this technology to the SARS-CoV-2 virus has revealed similar potential "Achilles heels."

Introduction

After four decades of research, infection with HIV-1 can routinely be controlled but, because the virus adopts DNA-form latency, not cured. Since the newly emergent SARS-CoV-2 displays neither HIV-1-like latency nor its extreme variability and chronicity, curative treatments would seem more feasible (1). Although both are RNA viruses, their many differences indicate that different therapeutic strategies will be required. Yet, as with HIV-1, strategies for SARS-CoV-2 have focused, with limited success, on the functions of encoded proteins rather than those of its RNA genome or of transcripts from that genome. Intriguingly, a nucleic acid level bioinformatics technology that long ago suggested a HIV-1 vulnerability related to genome function (2, 3), has recently gained support (4-6), and the term "Achilles heel" is increasingly employed (1, 6, 7). The same technology now suggests similar vulnerabilities for SARS-CoV-2.

The technology (see Materials and Methods) requires elimination of the contribution of base composition to the energetics of the folding into a higher order (stem-loop) structure of a single-stranded nucleic acid. Just as accent or dialect affects a spoken text in its entirety, so base composition tends to reflect *genome-wide* evolutionary pressures. Just as a local arrangement of words conveys specific meaning to a text, so base order better reflects *local* evolutionary pressures. Base order is most likely to be conserved when encoding a function critical for survival. Assays of the base order-dependent component of the folding energy have shown that a highly conserved region, in otherwise rapidly mutating HIV-1 genomes, associates with an RNA structure corresponding, not to a protein-encoding function, but to an RNA packaging signal. The latter is specifically recognized by the nucleocapsid domain of the Gag polyprotein (8) and is now seen as a potential "Achilles heel" of HIV-1 that can be targeted by a recently described antiviral compound (NSC 260594) or by synthetic oligonucleotides (4).

We here report similar highly conserved structural regions of the SARS-CoV-2 genome, one or more of which should be susceptible to targeting (9, 10). We identify certain open reading frames (ORFs) that, because of their conservation, have so far attracted therapeutic interest mainly related to their functions at the protein level (11, 12), rather than at the level of the corresponding, yet highly structured, regions of the genome. The ribosomal frameshift element (FSE) that is among our results, is attracting attention (10, 13, 14). Yet our analysis, consistent with recent reports (15, 16), suggests there may be more suitable targets in other regions.

Results

Variation Due to Base Order Component. A nucleic acid folding program was applied to 200 nucleotide base "windows" that were moved along a SARS-CoV-2 sequence in 20 base steps (see Materials and Methods). This provided a set of folding energy values fluctuating from -26 kCal/mol to -78 kCal/mol, with the latter negative value indicating *stronger* folding (Fig. 1; blue line). These "folding of natural sequence" (FONS) values were further dissected into their two fundamental components, that due to base *composition* (FORS-M) and that due to base *order* (FORS-D).

The composition component (Fig. 1; red line) is generally less negative than the FONS value (blue line), the difference being due to the contribution made by the base order-dependent component (green line). Composition (FORS-M) usually makes a larger, but less variable contribution, whereas the order component (FORS-D) usually makes a smaller, but highly variable contribution. Thus, much of the fluctuation that distinguishes different parts of the genome in the FONS profile is due to the base order component. As with our previous studies of a wide variety of genomes, FORS-D values provide a sensitive indicator of regional differences in the strength of folding. In a few regions FORS-M is more negative than FONS and here FORS-D values are positive. Thus, here "nature" appears to have arranged the order of the nucleotide bases to *restrain* the impact of base composition on the strength of folding. A need for this appears particularly evident in windows centered on the following bases: 1360 (NSP2), 9200 (NSP4), 12300 (NSP8), 15880 and 15920 (NSP12), 19940 (NSP15), 29680 and 29700 (3' untranslated region).

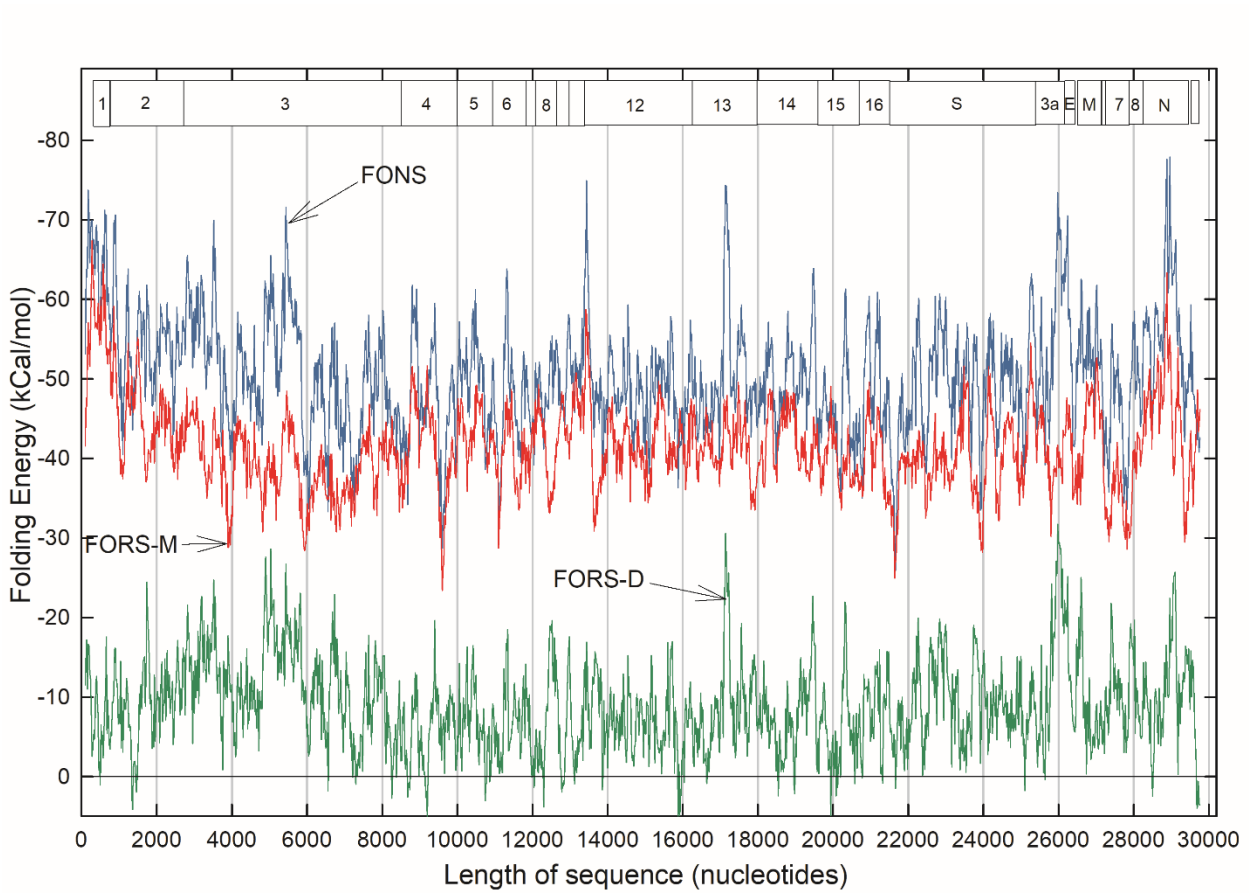

Fig. 1. Variations in folding energetics throughout the single-stranded RNA genome of a clade B SARS-CoV-2 sequence, deemed "prototypic" since it appeared in China early in the COVID-19 epidemic (Wuhan-Hu1; NC_045512.2). Data points for 1484 200 base windows (beginning at base 1 and moved in *20 base steps*) were connected rectilinearly. FONS values (blue); FORS-M values (red); FORS-D values (green). Boxes indicate the locations of the nonstructural proteins (NSP 1-16) and structural proteins (S, etc.).

Inverse Relationship between Folding and Substitutions. The degree of conservation in sequential windows of members of a set of sequences from China was evaluated as the number of substitutable base positions in that set (polymorphism), relative to the prototype sequence. This was compared with corresponding FORS-D values (Fig. 2), the profile of which differed a little from that of Fig. 1 due to different step values (see Materials and Methods).

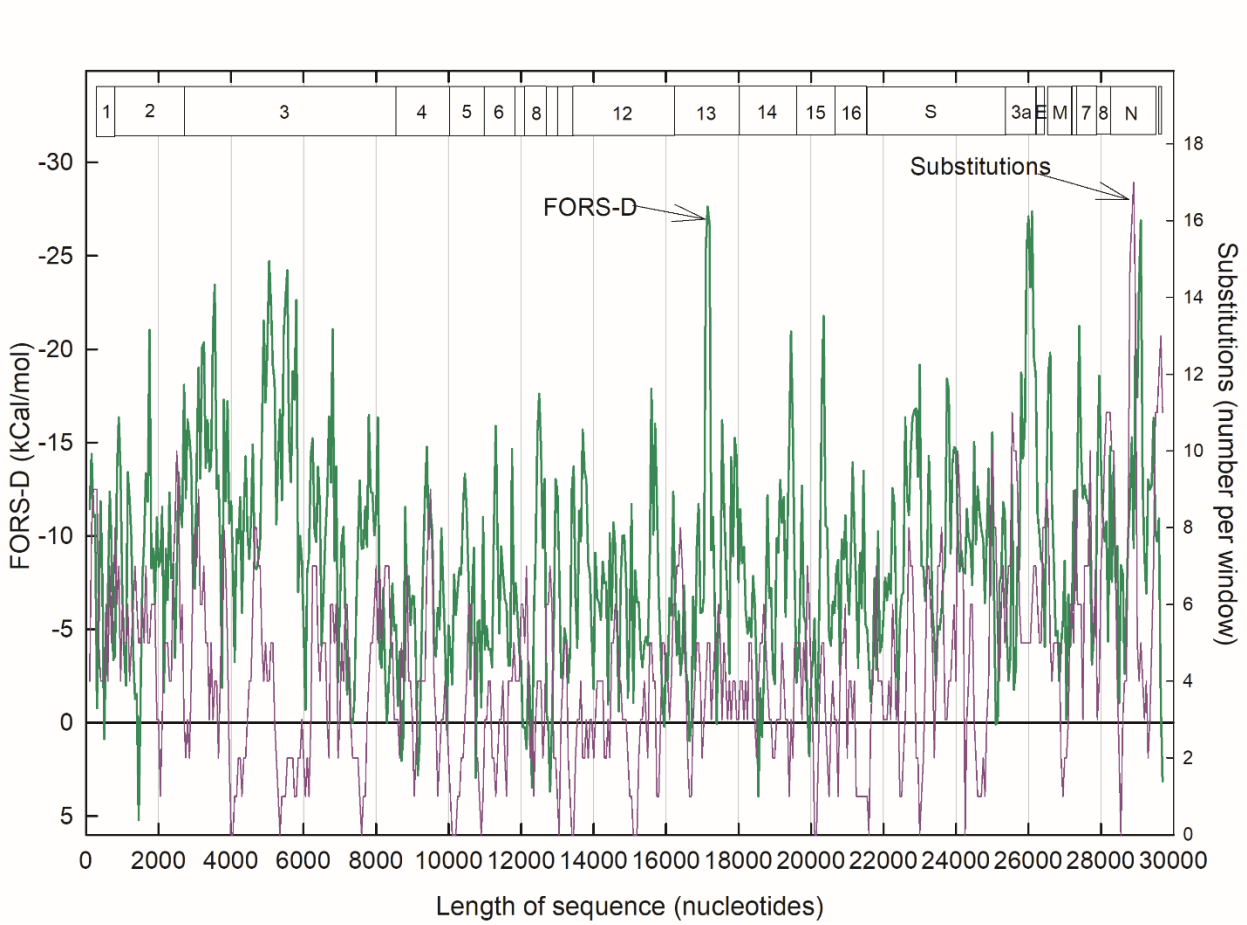

Fig. 2. Comparison of base order-dependent stem-loop potential values (FORS-D; green) for the Wuhan-HU1 SARS-CoV-2 prototype and base substitutions values (purple). The latter were found in the same sequence windows among a set of Chinese SARS-CoV-2 isolates. Data points for 594 200 base sequence windows (beginning at base 1 and moved in *50 base steps*) were connected rectilinearly. (There were some ambiguities in substitutions in the sequences included in the first two windows, making the first two data points less reliable).

As with some previous studies (17), there tended to be a reciprocal relationship between the base order contribution to folding energy (FORS-D) and substitutions. When one was high the other was low, and vice-versa. Negative FORS-D values were generally greatest in the region of ORF NSP3, and in narrow regions corresponding to the NSP13, ORF3a and the nucleocapsid (N) proteins. Zero substitutions were observed for windows centered on the following base nucleotide values, some isolated and some grouped (in brackets): [4000, 4050], 5350, 7600,

[10100, 10150, 10200], 10900, 13050, [13400, 13450; close to ribosomal FSE], [15100, 15150, 15200], [20100, 20150], 21600, 23000, 24250, 28550. Substitutions were higher in the last part of the sequence, being particularly evident in the nucleocapsid (N) protein ORF. While the zero-substitution windows centered on bases 5350, 7600 and 21600 appeared isolated, each had many neighboring windows with few substitutions. At some positions, *low* substitution values were accompanied by *high* negative RNA folding energy values, indicating conservation of structural function that could be at the genome level or in RNA transcripts.

Structural Invariance Index Identifies Potential Targets. For a more focused view of the reciprocal separation of high *negative* folding values and corresponding numbers of substitutions (ranging from zero to high *positive* values) the two were added to provide the "structural invariance" (SI) index (Fig. 3).

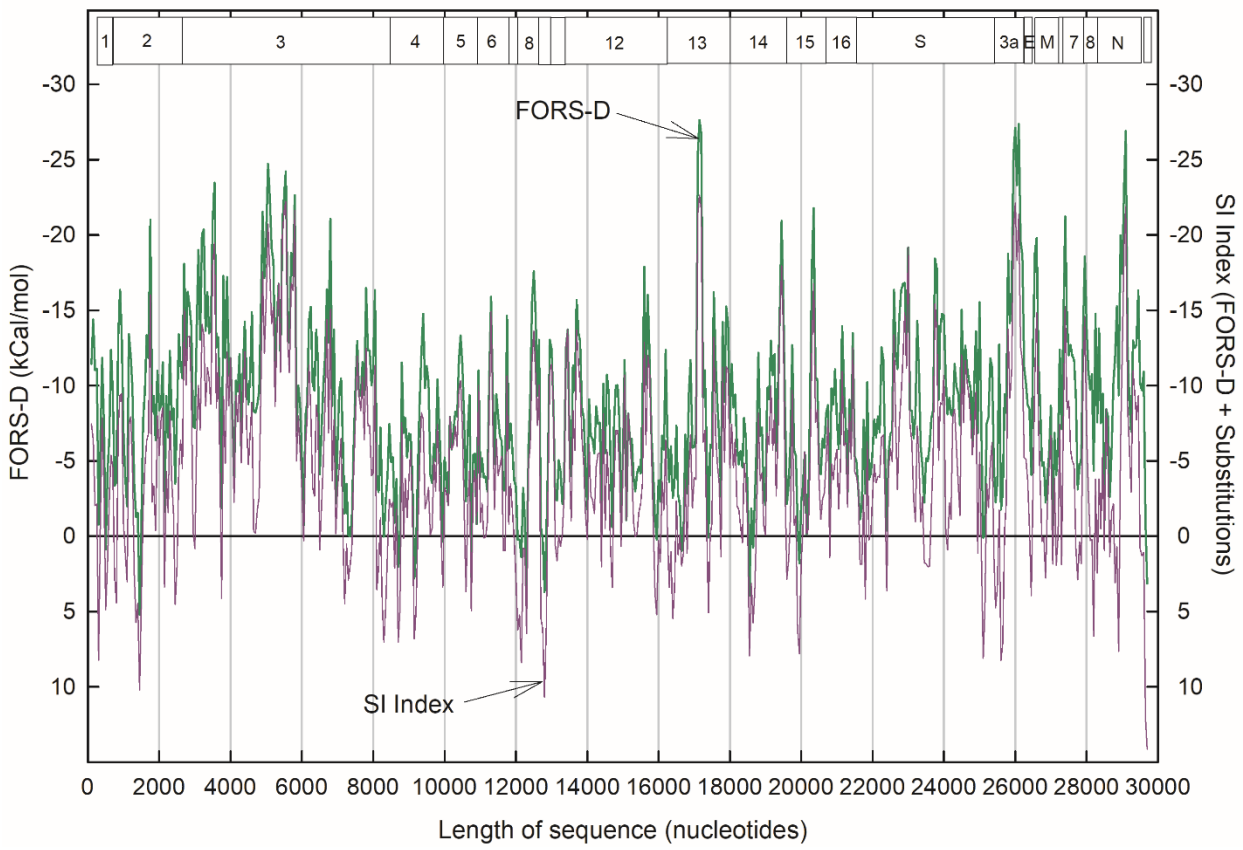

Fig. 3. Data of figure 2 plotted with FORS-D values (green) accompanied by the structural invariance (SI) index (SI = FORS-D values + substitutions; purple).

Here, despite having some substitutions, ORFs NSP13 and NSP3 were preeminent (scoring -22.6 and -22.2 SI units, respectively). The next two best scores were within ORF3a (scoring -22.1 SI units) and within the N protein ORF (scoring -21.9 units). Whereas the scores for the NSP13, ORF3a and N locations were narrowly based, the entire NSP3 location tended towards high scores. Close downstream to the high-scoring N region (window centered on base 29100) is a region with the highest number of substitutions (17; see Fig. 2) centered on base 28900 (scoring +7.6 SI units).

Windows 13400 and 13450 in the FSE region scored -12.5 and -13.7 units, respectively. However, there were many more windows with higher negative scores. Indeed, SI values more negative than -15.0 units were observed for windows centered on the following base nucleotide values, some isolated and some grouped (in brackets): 1750, 2700, [3500, 3550], 4900, [5000, 5050, 5100], 5200, [5450, 5500, 5550], 5800, 6800, [17100, 17150, 17200], 19450, 20350, 23000, 23750, [25950, 26000, 26050, 26100], 27400, [29050, 29100].

SI Indices for Italy and New York (USA). The SI profile for China (Fig. 3) was, in broad outline, confirmed with corresponding data later downloaded from Italy and New York, USA (Fig. 4). The high *negative* SI indices found in the NSP3, NSP13, ORF3a and N regions, were evident with sequences for all three locations. Other regions, notably the S region, were also corroborated. The high *positive* SI values, indicating regions likely to have poorly conserved structures, are also corroborated at some locations. A high *intraspecies* mutation rate for the N protein ORF (Fig. 2) is also seen when *interspecies* comparisons are made with other coronavirus species, with implications for early speciation mechanisms (18).

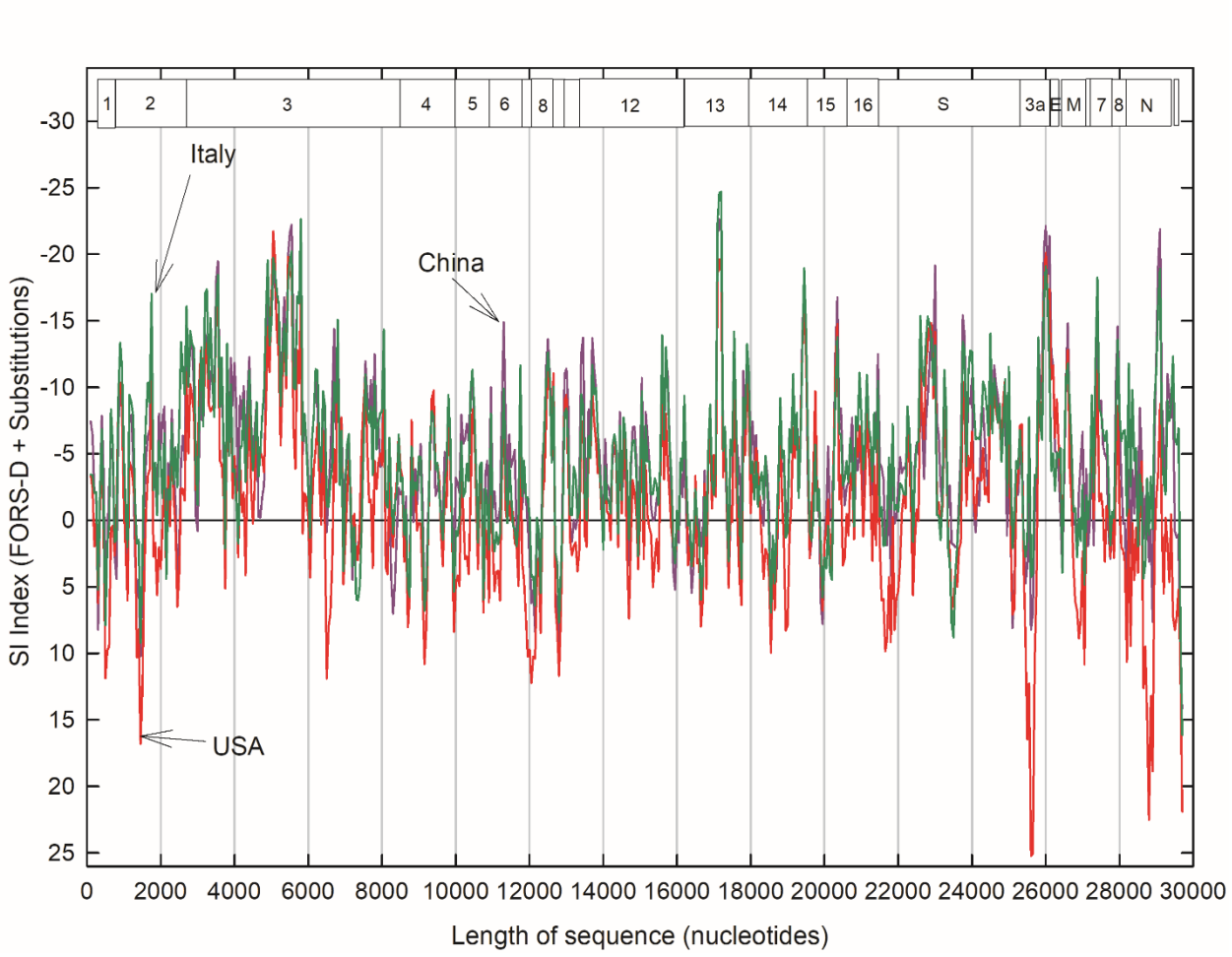

Fig. 4. Comparison of SI indices for China (381 isolates) as in Fig. 3 (purple), with SI indices for Italy (430 isolates; green) and New York, USA (932 isolates; red).

Discussion

Mutation rates of microbial pathogens are generally higher than those of their hosts. While a microbe spreading from host-to-host can "anticipate" that it will face a succession of broadly similar challenges, in the short-term those hosts cannot likewise "anticipate" that new microbial invaders will remain as they were in previous hosts. Thus, host immune defenses may be overwhelmed. (In the long-term there is a different scenario related to innate immunity; see below). Therapeutic challenges are, first, to locate a conserved, less-variable, part of a pathogen's genome that it will have inherited sequentially from a multiplicity of past generations, and so is likely to carry through to a multiplicity of future generations. Second, is to identify the

corresponding primary function, be it at the genome, RNA transcript, or protein level. Third, from this knowledge (that may be incomplete; i.e. function not fully clarified), devise effective pathogen inhibition without imposing deleterious side-effects on the host.

Viral vulnerability is often assumed to associate with protein level functions (11, 12). However, studies of the AIDS virus have identified genome structure itself as both functional and conserved, so signifying vulnerability (2, 3; see Fig. 5 in Materials and Methods). A genomic packaging signal for HIV-1, which is specifically recognized by the nucleocapsid domain of its Gag polyprotein, has long been recognized as a potential "Achilles heel" (2, 3, 7) so inviting therapeutic exploration (4-6). Through targeting of specific RNA conformations, Gag not only influences the assembly of HIV-1 genomic RNA into virus particles, but also regulates HIV-1 mRNA translation (19). RNA conformational flexibility, facilitated by fine-tuned intermediate-strength binding, should permit easier switching between regulatory options. Indeed, small changes in target RNA structure can impede this (6). Thus, unlike most other regions of the HIV-1 genome, mutations here would likely lead to negative Darwinian selection at an early stage – hence the high conservation. Application of the same bioinformatic technology to the SARS-CoV-2 virus genome has now revealed similar "Achilles heels."

Lacking the chronicity of HIV-1 infection, the genome of SARS-CoV-2 should have been shaped less by adaptations to counter long-term host immune defenses. It cannot hide within its host genome in latent DNA form. Yet, the larger SARS-CoV-2 genome contains many more genes than HIV, which require differential expression according to the stage of infection. Even more complex regulatory controls can be envisaged, likely requiring conserved genome conformations at appropriate locations. Be they synonymous or non-synonymous, mutations in these structured regions could result in negative selection of the viruses in which they occurred – hence high conservation. The ribosome FSE located close to base 13468 (Figs. 2, 3) would seem to exemplify this (13), and a potential targeting agent is now available (10). However, we have here identified other structurally important regions with higher degrees of conservation (Figs 3, 4), that might, either singly or collectively, be better candidates for targeting.

Our approach depends on eliminating contributions of base composition when determining folding energy. Base composition is a distinctive characteristic of entire genomes or large genome sectors that reflects their underlying oligomer ("k-mer") content. The slow genome-wide accumulation of mutations in oligomer composition (easiest documented as changes in base composition; GC%) can serve to initiate divergence into new species. By preventing that accumulation, potentially diverging organisms can stay within the confines of their species (20-22). The presence or absence of synonymous mutations, which affect structure rather than amino acid composition (16, 23), can have an important role in this process. The primary role of constancy in the base composition-related character is to prevent recombination with allied species (*interspecies* recombination) while facilitating *intraspecies* recombination, which can correct mutations so retaining species individuality (3, 7). Such recombination is initiated by "kissing" interactions between complementary unpaired bases at the tips of stem-loop structures (20).

Thus, we are here concerned with *localized* intraspecies mutations that affect fitness, so making members of a species carrying those mutations liable to natural selection. The mutations facilitate *within-species* evolution rather than divergence into new species. And when that evolution has run its course, some of the polymorphic bases will have become less mutable, so will be deemed "conserved." Indeed, mutations of ORF NSP3 are high when the sequences of *different* coronavirus species are compared (24), yet when, from *intraspecies* comparisons, mutations (in the form of base substitutions) are scored, they are very low (Figs. 2-4). Our technology (see Materials and Methods) removes the base composition-dependent component of mutational changes (that relates more to *interspecies* evolution) and focuses on the base order-dependent component (that relates more to *intraspecies* evolution). It best reflects localized functions, be they encoding protein or determining the folding of linear, single-stranded, nucleic acid sequences into higher order structures.

Is conservation necessarily a good indicator of likely therapeutic success? We sought regions that were both high in stem-loop potential and bereft of mutations, following the premise that conserved functions would be best targeted therapeutically, assuming the availability of pathogen-specific therapeutic agents that would not cross-react with hosts. Interference with

structural nucleic acid level functions would seem less likely to produce unforeseen host side-effects than with protein-level functions. Yet, the possibility remains. Indeed, a conserved function in a pathogen could *owe* that conservation to the pathogen strategy of, whenever possible, mutating to resemble its host. This would make it less vulnerable to host immune defenses. To prevent autoimmunity, the generation of immune cell repertoires involves the negative selection of self-reacting cells so creating "holes" in the repertoire that pathogens can exploit by progressive mutation towards host-self, testing mutational effectiveness a step at a time. This would make it advantageous for the host, during the process of repertoire generation, not only to *negatively* select immune cells of specificity towards "self," but also to *positively* select immune cells of specificity towards "near-self." A high level of anti-near-self immune cell clones would constitute a barrier limiting the extent of pathogen mutation towards self. The existence of such positive selection is now generally accepted, with the implication that some pathogen functions might have approached so close to host-self that targeting them therapeutically would result in cross-reactivity (25).

This unlikely caveat aside, we deem conservation a reliable indicator that a certain pathogen function is likely to be a suitable target for therapy. Attacking a short segment of a pathogen nucleic acid sequence is unlikely to ensnare a similar segment of its host's nucleic acid. In any case, to militate against this, the pathogen specificity of a potential therapeutic agent can be screened against the prototypic human genomic sequence (assuming it is unlikely that patient genomes will significantly depart from this).

Concluding Comments

While prospects for the development of prophylactic vaccines against infection with SARS-CoV-2 are promising, methods to boost post-infection host immune defenses and to directly target SARS-CoV-2 are urgently needed. These require a better understanding both of viral interactions with host innate and acquired immune systems (25), and of viral vulnerabilities. The latter enquiry proceeds in three steps: Find specific "Achilles heels." Design therapies to exploit them. Prove their clinical effectiveness. We have here been concerned with the first step. Although the bioinformatic technology related to this has long been available (2), its claim to

reveal viral "Achilles heels," as promulgated in successive textbook editions (7), has only recently gained support (4-6). This may be because the importance of removing redundant information and analyzing solely the contribution of base order to folding energy, was not fully appreciated. Thus, we have here repeated and expanded on past clarifications (26) of the conceptual basis of a technology that has contributed to the understanding of a many biological problems other than viral infections (7). Meanwhile, it is pleasing to note that, with minimal evidence on targeting, good progress is being made with the second step (9, 10). It is hoped that by targeting one or more of the conserved regions in the SARS-CoV-2 genome here identified, rapid cures will be achieved.

Materials and Methods

Sequences. These were obtained from the NCBI (Bethesda) and GISAID EpiCoV (Munich) databases. The Wuhan-Hu1 sequence (GenBank NC_045512.2), deemed taxonomically prototypic (27, 28), was compared (regarding base substitutions and folding potential) with 381 Chinese isolates, 430 Italian isolates, and 932 isolates from New York, USA. Our window starting point was base 1 of the 29,903 base prototype sequence. We refer to windows by their centers. The center of the first 200 base window would be 100.

Substitution frequencies. In previous HIV-1 studies the base differences between just two individual sequences sufficed for the tabulation of a statistically significant set of base substitution frequencies (2). The lower mutation rates of SARS-CoV-2 strains (11) required the tabulation of changeable positions relative to the prototype (i.e. polymorphisms), among a set of individual sequences in a geographical region. In early September, 709 downloaded Chinese sequences were filtered to remove duplicates and those with ambiguities, so yielding 381 unique sequences. These were scored for substitutable positions (whether one base could be exchanged by one of the three alternative bases). This yielded a substitution value for the 200 bases in each sequence window (i.e. values ranging from zero to 200). The values were in the low range indicating less likelihood of back mutations (site saturation) in the period since divergence from the presumed prototype. Indeed, in a study of several thousand genomes 73% of sites were

conserved (i.e. only 27% were polymorphic; 16). To gain support for our results (Fig. 4), in late September our study was extended with sets from Italy (430 from 642 downloaded) and New York, USA (932 out of 1483 downloaded). As expected, the FORS-D profiles of two early-arising members of these sets, that differed from the prototype in two (Italy) and nine (USA) bases, scarcely varied from the profile of the Chinese prototype shown in Figure 1b (i.e. overlapping profiles). In contrast, profiles differ greatly when different coronavirus species are compared (16).

Base order-dependent component of folding energy. The energetics of the folding of a single-stranded nucleic acid into a stem-loop structure depend on both the composition and order of its bases. Base composition is a distinctive characteristic of a genome or large genome sector. A localized sequence (e.g. a 200 base window), which is rich in the strongly-pairing bases G and C, will tend to have a stable structure simply by virtue of its base composition, rather than of its unique base order. This high GC% value can obscure the contribution of the base order-dependent component of the folding energy, which provides a sensitive indicator of *local intraspecies* pressures for the conservation of function within a population (i.e. a mutated organism is eliminated by natural selection so no longer can be assayed for function in the population). In contrast, *interspecies* mutations tend to influence the *genome-wide* oligonucleotide (k-mer) pressure, of which base composition (GC%) is an indicator. This pressure can act to generate and/or sustain members of emerging species by preventing recombination with parental forms (3, 7, 20-22). Elimination of this base composition-dependent component facilitates focus on local folding.

Early studies of RNA virus structure by Le and Maizel (29) were primarily concerned with the statistical significance of RNA folding, rather than with distinguishing the relative contributions of base composition and base order. However, with a pipeline between the various programs that were offered by the Wisconsin Genetics Computer Group, the base composition and base order-dependent components were separated and individually assessed ("folding of randomized sequence difference" analysis; FORS-D analysis). Departing from Le and Maizel, FORS-D values were not divided to yield Z-scores, but were simply plotted with statistical error

bars (2, 3, 7, 17, 30). The limits of the latter were generally close to the corresponding FORS-D values and, for clarity, are omitted here.

A window of 200 bases is moved along a natural sequence in 20 or 50 base steps. A folding program (31) is applied to the sequence in each window to obtain "folding of natural sequence" (FONS) values for each window, to which both base composition and base order will have contributed. The four bases in each sequence window are then shuffled to destroy their order while retaining base composition, and folding energy is again determined. This shuffle-and-fold "Monte Carlo" procedure is repeated ten times and the average (mean) folding value is taken as the "folding of randomized sequence mean" (FORS-M) value for that window. This reflects the contribution of base composition alone. The base order-dependent component is then derived by subtraction from the FONS value. This is the “folding of randomized sequence difference"(FORS-D) value. The sign of the difference value depends on the direction of the subtraction. (FORS-D was given positive values in the early 1990s, but the direction of subtraction was changed in subsequent work.) Fluctuations seen in FONS profiles of genomes are mostly due to changes in the FORS-D component, whereas the FORS-M component, while making a major contribution to folding energetics, is relatively constant (Fig. 1).

The approach was employed by others who, rather than shuffling the *four* bases, favored retaining some base order information. Accordingly, they shuffled groups of bases (e.g. the *sixteen* dinucleotides). Following disparagement of the conceptual basis of four base shuffling, which was duly clarified (26), the validity of single base level shuffling is now generally accepted and is being applied routinely to various viral genomes (15, 16, 32). The Monte Carlo procedure can also be simplified to decrease FORS-M computational time (33) using support vector machine-based technology (34). Our original software ("Bodslp"), written by Professor Jian-sheng Wu (35), retains the Monte Carlo approach and was further developed by Professor Shungao Xu as "Random Fold-Scan" for Windows-based systems (36).

In addition to assisting the study of infectious viruses and protozoa (37), FORS-D analysis proved fruitful when applied to topics such as speciation (3, 20, 38), the origin of introns (17, 30, 39), relating structure to recombination breakpoints and deletions (40, 41), and relying on a

single sequence (rather than alignments) for the determination of positive Darwinian selection (42). However, for a given sequence window, output can follow only from the base order *in that window*. Lost are higher order structures that might occur naturally through long-range interactions (14). Furthermore, if the artificial demarcation of a window happens to cut between the two limbs of a natural stem-loop structure, then lost are what might have been contributed to the folding energetics had a larger window, or a different section point, been chosen. Variations in step size will generate differences in windows and hence variations in results (e.g. compare FORS-D profiles with 20 base and 50 base steps; Figs. 1, 2). Such variations might be less when window margins correspond to natural section points, such as those demarcating an RNA transcript. There is also a kinetic aspect, particularly apparent with transcripts, due to the probability that the pattern of early 5' folding will influence later folding.

Validation of Structural Invariance Index. Plots of structural invariance (SI) facilitate the identification of conserved RNA structures. Structure and invariance can be united numerically by addition (SI = FORS-D + Substitutions). Structure is reflected in the degree of *negativity* of FORS-D values, and invariance is reflected in limitation of recorded *positive* substitutions at individual base sites (polymorphisms). Thus, the SI would be -20 units in the case of a FORS-D value of -20 kcal/mol in a window where there had been zero recorded substitutions (i.e. mutant forms had been eliminated by natural selection or genetic drift). If there had been 10 recorded substitutions the SI would be -10 units.

High negative SI values indicate structural conservation among a set of genomes. The validity of this approach is supported by prior work with HIV-1. Figure 5 shows application of the index to previously reported data on HIV-1 (2), where a high negative SI index value corresponds to regions recognized as likely to offer Achilles heel-like vulnerability (7). The region around the window centered on 500 nucleotide bases is the focus of recent work (4-6).

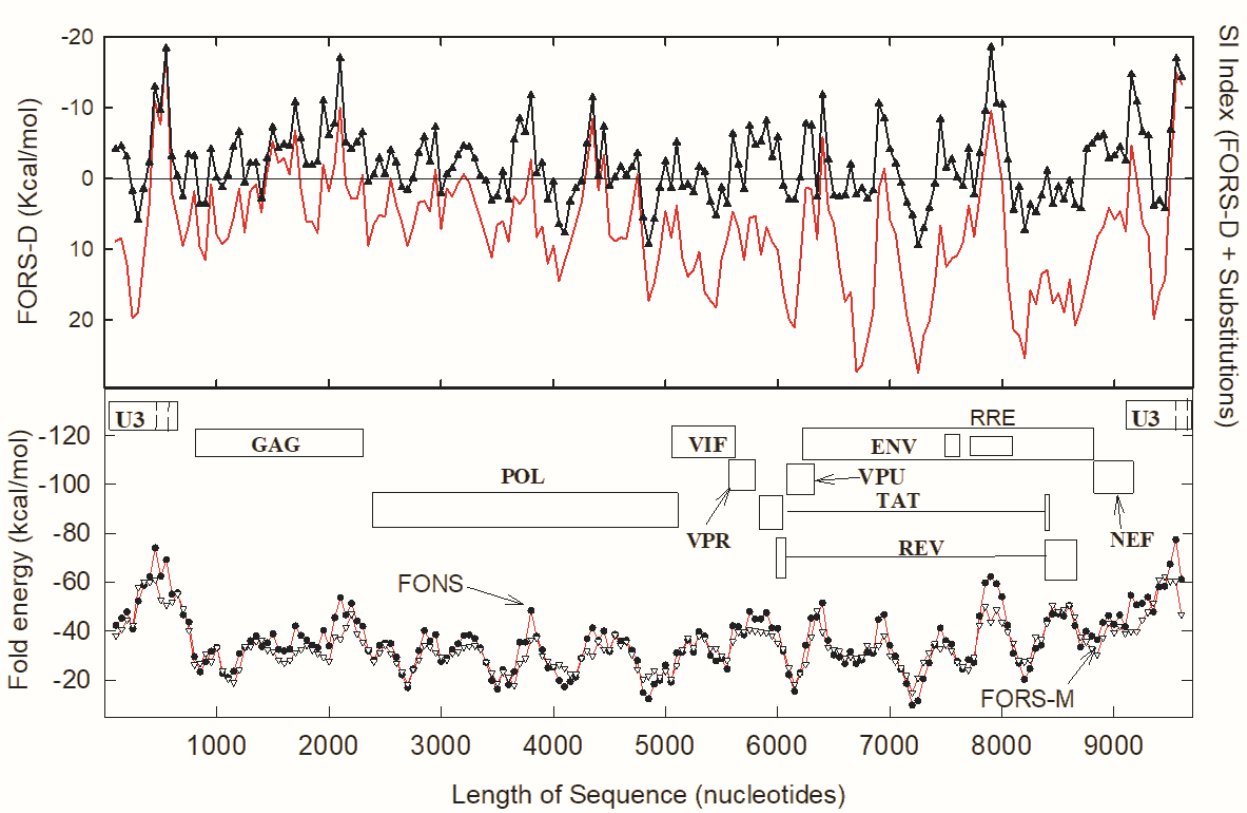

Fig. 5. Structural invariance (SI) of HIV-1 (subtype SF2) as compared with reference subtype (HXB2). The region recently recognized as a potential Achilles heel (bases 380-640; 4-6), has a high negative FORS-D value (black triangles) similar to those of the RRE (Rev response element) and the 3' untranslated region (UTR). The SI index (continuous red line) indicates highest sequence conservation in the regions of the RNA packaging signal and the 3'UTR.

ACKNOWLEDGEMENTS. We thank Prof. Shungao Xu at Jiangsu University for software, and Ms. Le Cao and Ms. Yingying Ma at Shanghai Public Health Clinical Center, Fudan University, for their technical support. Queen's University hosts Forsdyke's webpages.

1. J. W. Rausch, A. A. Capoferria, M. G. Katusiime, S. C. Patro, M. F. Kearney, Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proc. Natl. Acad. Sci. USA* **117**, 24614–24616 (2020).

2. D. R. Forsdyke, Reciprocal relationship between stem-loop potential and substitution density in retroviral quasispecies under positive Darwinian selection. *J. Mol. Evol.* **41**, 1022–1037 (1995).
3. D. R. Forsdyke, Implications of HIV RNA structure for recombination, speciation, and the neutralism-selectionism controversy. *Microbes. Infect.* **16**, 96–103 (2014).
4. C. K. Ingemarsdotter, J. Zeng, Z. Long, A. M. L. Lever, J. C. Kenyon, An RNA-binding compound that stabilizes the HIV-1 gRNA packaging signal structure and specifically blocks HIV-1 RNA encapsidation. *Retrovirology* **15**, 25 (2018).
5. A. Rein, The heart of the HIV RNA packaging signal? *Proc. Natl. Acad. Sci. USA* **117**, 19621–19623 (2020).
6. P. Ding *et al*., Identification of the initial nucleocapsid recognition element in the HIV-1 RNA packaging signal. *Proc. Natl. Acad. Sci. U.S.A*. **117**, 17737–17746 (2020).
7. D. R. Forsdyke, *Evolutionary Bioinformatics*, (3rd ed, Springer, New York, 2016), pp. 241–244.
8. S. Sarni, *et al*., HIV-1 Gag protein with or without p6 specifically dimerizes on the viral RNA packaging signal. *J. Biol. Chem*., 295:14391-14401 (2020).
9. I. G. Medeiros *et al*., A small interfering RNA (siRNA) database for SARS-CoV-2. *bioRxiv*: doi:10.1101/2020.09.30.321596 (13 October 2020).
10. H. S. Haniff *et al*., Targeting the SARS-CoV-2 RNA genome with small molecule binders and ribonuclease targeting chimera (RIBOTAC) degraders. *ACS Cent. Sci.* doi.org/10.1021/acscentsci.0c00984 (2020).
11. B. Robson, The use of knowledge management tools in viroinformatics. Example study of a highly conserved sequence motif in Nsp3 of SARS-CoV-2 as a therapeutic target. *Comput. Biol*. **125**, 103963 (2020).
12. E. Issa, G, Merhi, B. Panossian, T. Salloum, S. Tokajian, SARS-CoV-2 and ORF3a: Nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems* **5**, e00266 (2020).

13. J. A. Kelly *et al*., Structural and functional conservation of the programmed −1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *bioRxiv*: doi:10.1101/2020.03.13.991083 (15 June 2020).
14. O. Ziv et al., The short- and long-range RNA-RNA Interactome of SARS-CoV-2. *bioRxiv*: doi:10.1101/2020.07.19.211110 (7 October 2020).
15. Andrews et al., An *in silico* map of the SARS-CoV-2 RNA Structurome. *bioRxiv*: doi.org/10.1101/2020.04.17.045161 (18 April 2020).
16. P. Simmonds. Pervasive RNA secondary structure in the genomes of SARS-CoV-2 and other coronaviruses – an endeavour to understand its biological purpose. *Mbio*. (in press) *bioRxiv*: doi.org/10.1101/2020.06.17.155200 (17 June 2020).
17. D. R. Forsdyke, Conservation of stem-loop potential in introns of snake venom phospholipase A2 genes: an application of FORS-D analysis. *Mol. Biol. Evol*. **12**, 1157–1165 (1995).
18. M. Dilucca, S. Forcelloni, A. G. Georgakilas, A. Giansanti, A. Pavlopoulou, Codon usage and phenotypic divergences of SARS-CoV-2 genes. *Viruses* **12**, 1–21 (2020).
19. E. C. Anderson, A. M. L. Lever, Human immunodeficiency virus type 1 Gag polyprotein modulates Its own translation. *J. Virol*. **80**, 10478–10486 (2006).
20. D. R. Forsdyke, Different biological species "broadcast" their DNAs at different (G+C)% "wavelengths". *J. Theoret. Biol*. **178**, 405–417 (1996).
21. D. R. Forsdyke, Success of alignment-free oligonucleotide (k-mer) analysis confirms relative importance of genomes not genes in speciation. *Biol. J. Linn. Soc.* **128**, 239–250 (2019)**.**
22. D. R. Forsdyke, Hybrid sterility can only be primary when acting as a reproductive barrier for sympatric speciation. *Biol. J. Linn. Soc.* **128**, 779-788 (2019)**.**
23. H. Wang, L. Pipes, R. Nielsen, Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. *bioRxiv*: doi.org/10.1101/2020.04.20.052019 (12 October 2020).
24. J.-M. Claverie, A putative role of de-mono-ADP-ribosylation of STAT1 by the SARS-CoV-2 Nsp3 protein in the cytokine storm syndrome of COVID-19. *Viruses* **12**, 646 (2020).

25. D. R. Forsdyke, Two signal half century: from negative selection of self-reactivity to positive selection of near-self reactivity. *Scand. J. Immunol.* **89**, e12746 (2019).
26. D. R. Forsdyke, Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues*. J. Theor. Biol*. **248**, 745–753 (2007).
27. S. Kumar *et al*., An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic. *bioRxiv*: doi:10.1101/2020.09.24.311845 (24 September 2020).
28. L. Pipes, H. Wang, J. P. Huelsenbeck, R. Nielsen, Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *bioRxiv*: doi:10.1101/2020.06.19.160630 (7 Oct 2020).
29. Le, S.-Y., Maizel, J. V., 1989. A method for assessing the statistical significance of RNA folding. *J. Theor. Biol.* **138**, 495-510.
30. D. R. Forsdyke, Introns first. *Biol. Theor.* **7**, 196–203 (2013).
31. M. Zuker, Computer prediction of RNA secondary structure. *Meth. Enzym*. **180**, 262–289 (1989).
32. R. J. Andrews, J. Roche, W. N. Moss, ScanFold: an approach for genome-wide discovery of local RNA structural elements – applications to Zika virus and HIV. *PeerJ* **6**, e6136 (2018).
33. J.-H. Chen, S.-Y. Le, B. Shapiro, K. M. Currey, J. V. Maizel, A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci*. **6**, 7–18 (1990).
34. S. Washietl, I. L. Hofacker, P. F. Stadler, Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **102**, 2454–2459 (2005).
35. C.-Y. Zhang *et al*., Evaluation of FORS-D analysis: a comparison with the statistically significant stem-loop potential. *Biochem. Genet*. **46**, 29–40 (2008).
36. S. G. Xu, J. F. Wei, C. Y. Zhang, A FORS-D analysis software “Random_fold_scan” and the influence of different shuffle approaches on FORS-D analysis. *J. Jiangsu Univ. (Med. Ed.)* **17**, 461–466,470 (2007).
37. H. Y. Xue, D. R. Forsdyke, Low complexity segments in *Plasmodium falciparum* proteins are primarily nucleic acid level adaptations. *Mol. Biochem. Parasitol*. **128**, 21–32 (2003).

38. C. Zhang, S. Xu, J.-F. Wei, D. R. Forsdyke, Microsatellites that violate Chargaff's second parity rule have base order-dependent asymmetries in the folding energies of complementary DNA strands and may not drive speciation. *J. Theor. Biol.* **254**, 168–177 (2008).
39. D. R. Forsdyke, A stem-loop "kissing" model for the initiation of recombination and the origin of introns. *Mol. Biol. Evol*. **12**, 949–958 (1995).
40. C.-Y. Zhang, J.-F. Wei, S.-H. He*,* Local base order influences the origin of *ccr5* deletions mediated by DNA slip replication. *Biochem. Genet.* **43**, 229–237 (2005).
41. C.-Y. Zhang, J.-F. Wei, S.-H. He, The key role for local base order in the generation of multiple forms of China HIV-1 B'/C intersubtype recombinants. *BMC Evol. Biol.* **5**, 53 *(*2005).
42. D. R. Forsdyke, Positive Darwinian selection. Does the comparative method rule? *J. Biol. Sys.* **15**, 95–108 (2007).