Lost in translation: codon optimization inactivates SARS-CoV-2 RdRp

Bing Wang[1,2], Vladimir Svetlov[3], Evgeny Nudler[3,4], and Irina Artsimovitch[1,2#]

[1]Department of Microbiology and [2]The Center for RNA Biology, The Ohio State University, Columbus, OH 43210, USA; [3]Department of Biochemistry and Molecular Pharmacology and [4]Howard Hughes Medical Institute, New York University School of Medicine, New York, NY 10016, USA.

Address correspondence to Irina Artsimovitch, artsimovitch.1@osu.edu

Keywords: SARS-CoV-2 RdRp, RNA synthesis, synonymous codons, ribosome pausing

ABSTRACT

RNA-dependent RNA polymerase (RdRp) is a primary target for antivirals. We report that Nsp12, a catalytic subunit of SARS-CoV-2 RdRp, produces an inactive enzyme when codon-optimized for bacterial expression. We also show that accessory subunits, NTPs, and translation by slow ribosomes partially rescue Nsp12. Our findings have implications for functional studies and identification of novel inhibitors of RdRp and for rational design of other biotechnologically and medically important expression systems.

RNA viruses pose an existential threat to humankind. The coronaviral (CoV) genome encodes several non-structural proteins (Nsps) required for viral replication and gene expression[1]. Among these, Nsp12 plays a central role as a catalytic subunit of RdRp. A minimally active SARS-CoV-2 RdRp consists of Nsp12, Nsp7, and two copies of Nsp8[2-4] ($7{\bullet}8_2{\bullet}12$; **Fig. 1A**). Single-particle cryogenic electron microscopy (cryoEM) structures of RdRp[2-4] provide an excellent framework for understanding the mechanism of RNA synthesis and for the identification of novel RdRp inhibitors, studies that, in turn, critically depend on the availability of robust expression systems and active RdRp preparations. Since structures obtained with RdRp produced in *Escherichia coli*[2,4] and insect cells[3] are similar, we used the *E. coli* expression platform described in ref.[2] (Extended Data **Fig. 1**) to initiate mechanistic studies of SARS-CoV-2 RdRp. Subsequent *in vitro* experimentation revealed that RdRp assembled with thus purified Nsp12, which we will refer to as 12^R, was largely inactive. After a process of elimination, we identified Nsp12 recoding as an underlying reason for this low activity.

$Nsp12^R$ contains a non-cleavable C-terminal His_{10} tag, is relatively soluble when produced in *E. coli* and is easily purified under "native" (non-denaturing) conditions. However, the $7{\bullet}8_2{\bullet}12^R$ enzyme exhibited negligible activity on a number of different templates, including the optimal hairpin scaffold[3], which could be extended only at very low salt (**Fig. 1B**). An extensive survey of purification schemes, RNA scaffolds, and reaction conditions failed to identify conditions that supported efficient primer extension, and the removal of the His tag, proposed to be detrimental[5], did not increase activity (**Fig. 1B**). In their follow-up study, Rao and colleagues reported similar results[6], arguing that further optimization would be futile.

A robust enzyme activity is essential for functional studies - to rephrase the fourth commandment of enzymology[7], *thou shalt not waste clean thinking on dead enzymes*, and indispensable for the drug design and screening. Our survey of published reports failed to reveal an obvious reason for low activity - under similar reaction conditions, some RdRps were able to completely extend the RNA primer in minutes[3,5,8], whereas others failed to do so in an hour[6,9], regardless of the expression host. Idiosyncratic but reproducible variations in activity can arise from recombinant protein misfolding; indeed, co-expression of Nsp12, a multi-domain 932-

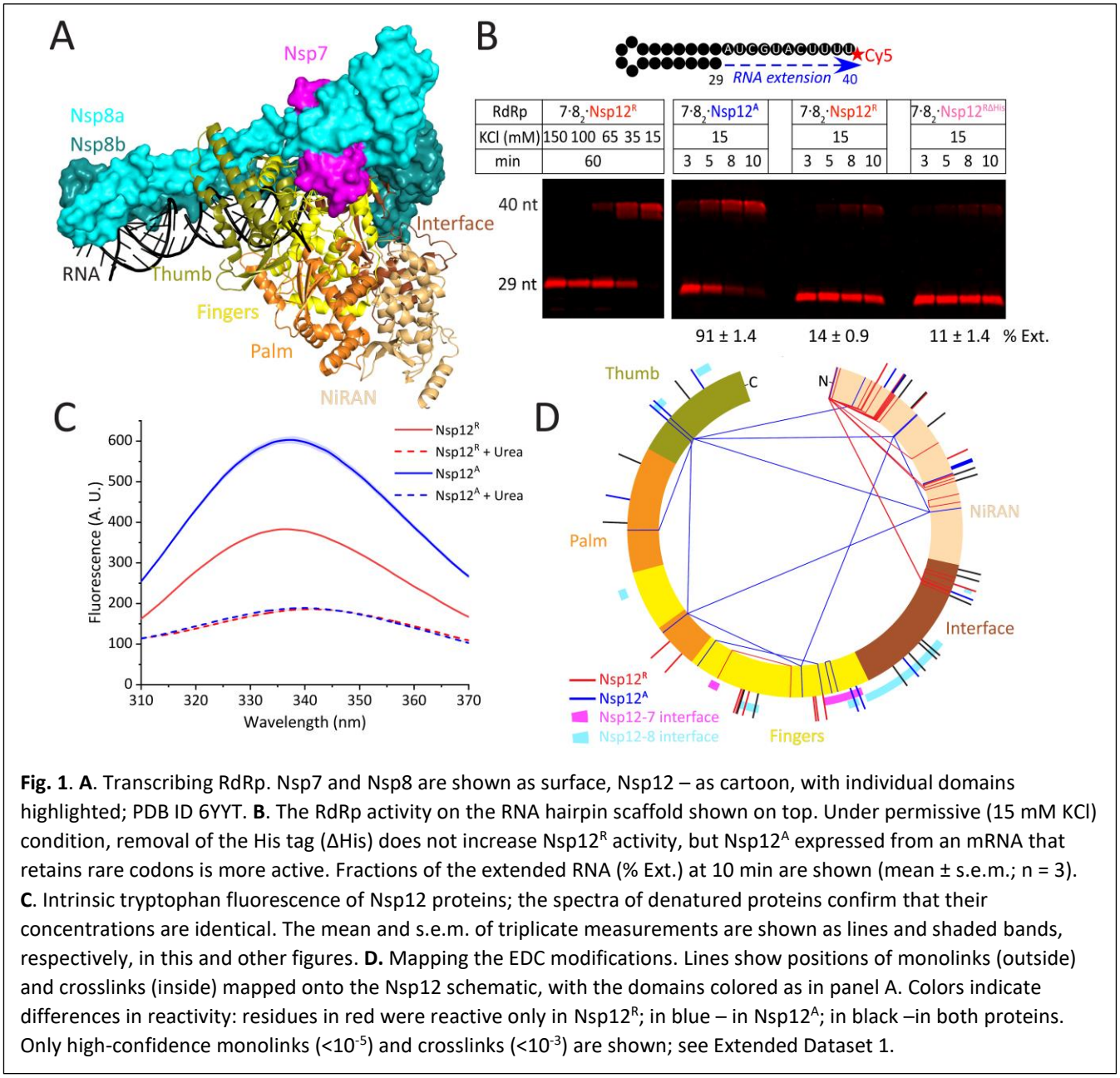

Fig. 1. **A**. Transcribing RdRp. Nsp7 and Nsp8 are shown as surface, Nsp12 – as cartoon, with individual domains highlighted; PDB ID 6YYT. **B**. The RdRp activity on the RNA hairpin scaffold shown on top. Under permissive (15 mM KCl) condition, removal of the His tag (ΔHis) does not increase Nsp12^R activity, but Nsp12^A expressed from an mRNA that retains rare codons is more active. Fractions of the extended RNA (% Ext.) at 10 min are shown (mean ± s.e.m.; n = 3). **C**. Intrinsic tryptophan fluorescence of Nsp12 proteins; the spectra of denatured proteins confirm that their concentrations are identical. The mean and s.e.m. of triplicate measurements are shown as lines and shaded bands, respectively, in this and other figures. **D.** Mapping the EDC modifications. Lines show positions of monolinks (outside) and crosslinks (inside) mapped onto the Nsp12 schematic, with the domains colored as in panel A. Colors indicate differences in reactivity: residues in red were reactive only in Nsp12^R; in blue – in Nsp12^A; in black –in both proteins. Only high-confidence monolinks (<10^{-5}) and crosslinks (<10^{-3}) are shown; see Extended Dataset 1.

residue protein, with cellular chaperons improves its activity[5,10]. The feasible source of this variability could lie in the coding mRNA itself: whereas all Nsp12s have the same amino acid sequence (ignoring the tags), their coding sequences (CDS) have been altered to match the codon usage of their respective hosts to maximize protein expression. This approach is routinely used by us and others, yet protein function can be compromised even by a single synonymous codon substitution[11], and protein expression in an *E. coli* RIL strain, which alleviates codon imbalance by supplying some rare tRNAs, may hinder proper folding[12]. The loss of ribosome pausing at rare codons is thought to uncouple the nascent peptide synthesis from its folding, giving rise to aberrantly folded proteins[11,13].

The viral *nsp12* mRNA is not optimized for translation in human cells[14], whereas the 12^R codon usage matches that of highly expressed *E. coli* genes, suggesting that the 12^R CDS could be over-optimized. Consistently, we found that RdRp assembled with $Nsp12^A$, whose CDS has more rare codons (Extended Data **Fig. 2**), had a much higher activity on the hairpin scaffold (**Fig. 1B**). We also noted that $Nsp12^A$ co-purified with nucleic acids, whereas $Nsp12^R$ did not; subsequent gel shift assays revealed that $7•8_2•12^A$ readily bound the RNA hairpin, whereas $7•8_2•12^R$ did not (Extended Data **Fig. 3**).

To test whether $Nsp12^R$ is misfolded, we used several approaches. *First*, we assessed the Nsp12 thermal stability using differential scanning fluorimetry. We recorded melting temperatures (T_m) of 41.3 °C for $Nsp12^R$ and 47.3 °C for $Nsp12^A$ (Extended Data **Fig. 4A**); T_m of 43.6 °C was reported for another *E. coli*-expressed Nsp12[15]. *Second*, we compared the fluorescence spectra of Nsp12, which contains nine tryptophan residues expected to be sensitive to microenvironment [16]**.** $Nsp12^A$ and $Nsp12^R$ exhibit similar emission peaks but the $Nsp12^A$ intensity is two-fold higher (**Fig. 1C)**, suggesting that at least one Trp is more buried; the derivative spectra (Extended Data **Fig. 4B**) did not reveal any additional differences. These results show that $Nsp12^A$ and $Nsp12^R$ structures are different but cannot identify the altered regions. We next used a carboxyl- and amine-reactive reagent EDC [1-ethyl-3-(3-dimethylaminopropyl)carbodiimide) to map solvent-accessible (surface) residues and intra-protein crosslinks by mass spectrometry. We observed significant differences in the accessibility of several regions centered at residues 150 (NiRAN domain), 415 (Fingers), 600 (Palm), and 850 (Thumb) and in crosslinking, particularly of the NiRAN domain (**Fig. 1D**).

Although an overall abundance of underrepresented codons could slow translation, in many cases the ribosome must pause at one or more rare codons to promote folding of a "problematic" region. There are more differences between the 2.8 kb mRNAs encoding $Nsp12^A$ and $Nsp12^R$ than there are similarities (Extended Data **Fig. 2**). The produced proteins also differ in their N- and C-termini (Extended Data **Fig. 1**) but, based on available structural data, an extra N-terminal glycine would not be expected to account for dramatic differences in EDC reactivity (**Fig. 1D** and Extended Dataset 1). Comparative analysis identified two regions that contained rare codon clusters in the viral and 12^A mRNAs, but not in 12^R (**Fig. 2A** and Extended Data **Fig. 5**). We constructed chimeric proteins in which these 12^A segments were replaced with corresponding segments from 12^R, generating proteins with identical amino acid sequences (**Fig. 2A**). We found that while toggling of codons [143-346] between active and inactive Nsp12 variants did not alter the RdRp activity, a chimeric protein with codons [350-435] derived from the 12^R CDS was defective (**Fig. 2B**). Together with the EDC modification patterns (**Fig. 1D**), these data suggest that controlled translation of the 350-435 region is important for Nsp12 folding and that changes in contacts with Nsp7 (Extended Data **Fig. 6**), which are critical for RdRp activity[3], could be partially responsible for the low activity of $Nsp12^R$.

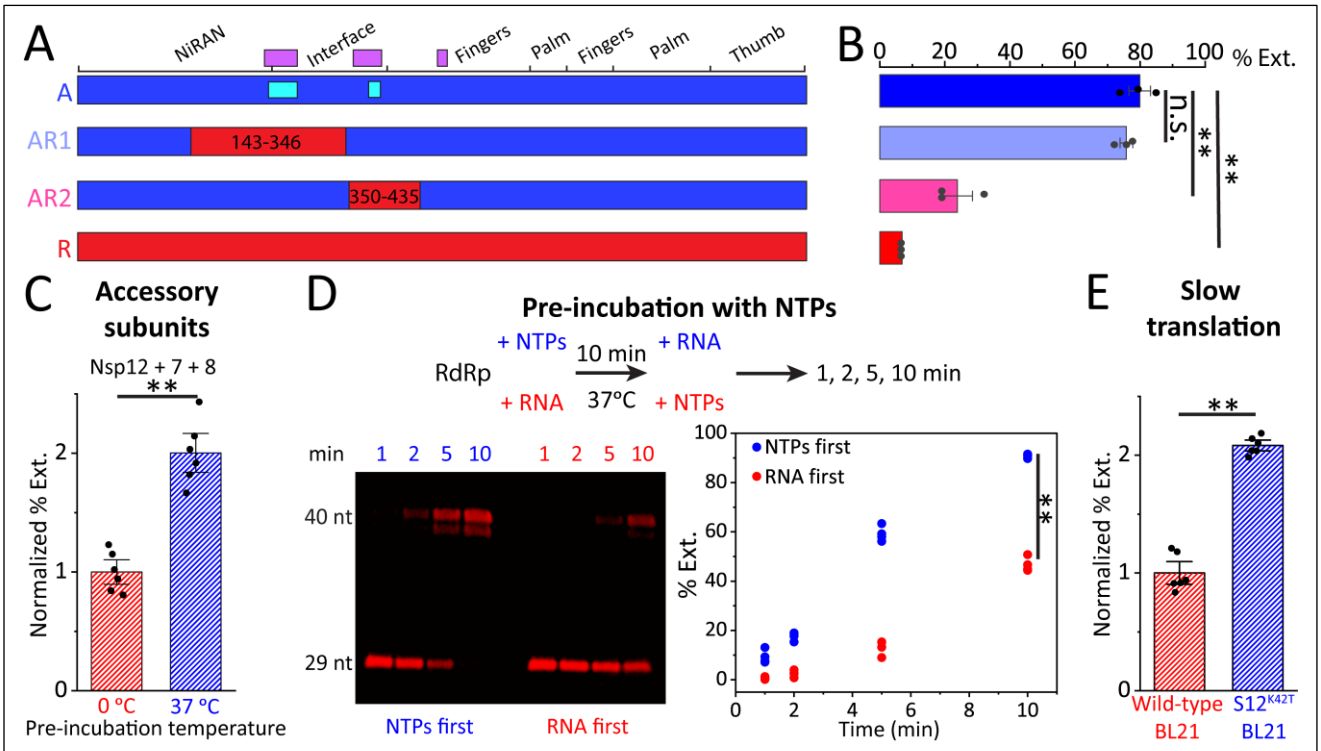

Fig. 2. Determinants of the Nsp12 activity. **A**. Translational context around residue 400 is critical for the correct folding of Nsp12. SARS CoV-2 genomic *nsp12* RNA (with domain boundaries shown on top) contains clusters of rare codons (purple bars); only 12^{A} CDS has rare codons (cyan bars) at the corresponding positions. **B**. A chimeric $Nsp^{12\text{-}AR2}$ protein is defective in RNA synthesis. **C-E**. Reactivation of $Nsp12^{R}$ via (**C**) formation of RdRp at 37 °C; (**D**) preincubation with NTPs; and (**E**) translation by slow ribosomes. RNA extension is shown as mean ± S.E.M. and p value was calculated by unpaired two-tailed t-test. n.s., not significant; **, $p < 0.01$.

Analysis of Nsp7/Nsp12 interactions by Trp fluorescence reveals that Nsp7 binds to both Nsp12 subunits and may favor a similar, $Nsp12^{A}$-like state (Extended Data **Fig. 6**). In support of the "scaffolding" function of the accessory subunits, we found that preincubation of $Nsp12^{R}$ with Nsp7 and Nsp8 led to an increased activity (**Fig. 2C**).

The patterns of EDC modification (**Fig. 1D**) suggest the NiRAN domain may be folded differently in $Nsp12^{A}$ and $Nsp12^{R}$ proteins. While the effect of this domain on RNA synthesis has not been studied, it is known to bind nucleotides[4]. We compared RNA synthesis under "standard" conditions, in which RdRp is prebound to the RNA scaffold prior to the addition of NTP substrates, to the "NTP-primed" reaction in which the order of addition was reversed (**Fig. 2D**). Our results show that preincubation with NTPs strongly potentiates $Nsp12^{R}$ activity, an effect that could be mediated by the NiRAN domain; we are working to elucidate the mechanism of NTP-dependent activation.

We next tested if slowing translation would promote Nsp12 folding. We constructed a BL21 strain with a K42T mutant of the ribosomal protein S12 (M&M) and compared the $Nsp12^{R}$ protein purified from this "slow" BL21 variant to the protein purified from the wild-type BL21. We found that $Nsp12^{R}$ purified from the mutant BL21 was more active (**Fig. 2E**), consistent with the favorable effect of attenuated translation.

Our results demonstrate that over-optimized $Nsp12^{R}$ mRNA produces a soluble but misfolded protein. Despite dramatic differences in their activities, all structures of Nsp12 transcription complexes are similar[2,4], reflecting the purifying selection of cryoEM analysis in which 95% of inactive conformers could be ignored because only active complexes are reported. However, using inactive enzymes would compromise ensemble experiments; *e.g.*, a comparison of SARS and SARS CoV-2 RdRp[15] would generate opposite conclusions if $Nsp12^{R}$ or $Nsp12^{A}$ were used. Even more critically, the inactive RdRp should not be used to screen potential inhibitors.

While recoding is routinely used to optimize heterologous protein expression, the existence of rare codons in mRNAs encoding essential proteins indicates their important roles; for example, native non-optimal codons in intrinsically disordered regions (IDR) are essential for function of circadian clock oscillators[11]. IDRs often serve as platforms for protein-protein interactions[17], but can become trapped in unproductive states in the absence of their partners. Our analysis supports this scenario: an unstructured region that binds Nsp7 displays significant differential sensitivity to EDC (**Fig. 1D**) and Nsp7 favors an "active-like" Nsp12 conformation (Extended Data **Fig. 6**). When added to the misfolded Nsp12, Nsp7/8 only modestly increase its activity (**Fig. 2C**). However, as all Nsps are produced as a giant precursor in the virus-infected cells[1], the accessory subunits could aid Nsp12 folding co-translationally, as apparently happens during their co-expression in *E. coli*[5]. Likewise, co-expression of *E. coli* RNA polymerase subunits suppresses assembly defects conferred by deletions in the catalytic subunits[18]. By contrast, single-subunit RdRps may be more dependent on the ribosome kinetics for proper folding.

Our results show that SARS-CoV-2 RdRp is highly active when expressed from an under-optimized mRNA and are of paramount importance for the ongoing search for antivirals against Covid-19. But more broadly, they show that the ubiquitous first step in designing a perfect heterologous expression system, the codon frequency optimization, may backfire spectacularly. The importance of co-translational folding, particularly for large and dynamic proteins that contain essential mobile regions, emphasizes a need for integration of diverse approaches, from ribosome profiling to machine learning, during the rational design of coding sequences.

References

1 Snijder, E. J., Decroly, E. & Ziebuhr, J. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. *Adv Virus Res* **96**, 59-126, doi:10.1016/bs.aivir.2016.08.008 (2016).

2 Gao, Y. *et al.* Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779-782, doi:10.1126/science.abb7498 (2020).

3 Hillen, H. S. *et al.* Structure of replicating SARS-CoV-2 polymerase. *Nature* **584**, 154-156, doi:10.1038/s41586-020-2368-8 (2020).

4 Chen, J. *et al.* Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-Transcription Complex. *Cell* **182**, 1560-1573 e1513, doi:10.1016/j.cell.2020.07.033 (2020).

5 Dangerfield, T. L., Huang, N. Z. & Johnson, K. A. Remdesivir Is Effective in Combating COVID-19 because It Is a Better Substrate than ATP for the Viral RNA-Dependent RNA Polymerase. *iScience* **23**, 101849, doi:10.1016/j.isci.2020.101849 (2020).

6 Wang, Q. *et al.* Structural Basis for RNA Replication by the SARS-CoV-2 Polymerase. *Cell* **182**, 417-428 e413, doi:10.1016/j.cell.2020.05.034 (2020).

7 Kornberg, A. Ten commandments: lessons from the enzymology of DNA replication. *J Bacteriol* **182**, 3613-3618, doi:10.1128/jb.182.13.3613-3618.2000 (2000).

8 Shannon, A. *et al.* Rapid incorporation of Favipiravir by the fast and permissive viral RNA polymerase complex results in SARS-CoV-2 lethal mutagenesis. *Nat Commun* **11**, 4682, doi:10.1038/s41467-020-18463-z (2020).

9 Kirchdoerfer, R. N. & Ward, A. B. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat Commun* **10**, 2342, doi:10.1038/s41467-019-10280-3 (2019).

10 Shannon, A. *et al.* Remdesivir and SARS-CoV-2: Structural requirements at both nsp12 RdRp and nsp14 Exonuclease active-sites. *Antiviral Res* **178**, 104793, doi:10.1016/j.antiviral.2020.104793 (2020).

11 Liu, Y. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun Signal* **18**, 145, doi:10.1186/s12964-020-00642-6 (2020).

12 Zhang, G., Hubalewska, M. & Ignatova, Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol* **16**, 274-280, doi:10.1038/nsmb.1554 (2009).

13 Liutkute, M., Samatova, E. & Rodnina, M. V. Cotranslational Folding of Proteins on the Ribosome. *Biomolecules* **10**, doi:10.3390/biom10010097 (2020).

14 Finkel, Y. *et al.* The coding capacity of SARS-CoV-2. *Nature*, doi:10.1038/s41586-020-2739-1 (2020).

15 Peng, Q. *et al.* Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2. *Cell Rep* **31**, 107774, doi:10.1016/j.celrep.2020.107774 (2020).

16 Chavez-Abiega, S., Goedhart, J. & Bruggeman, F. J. Physical biology of GPCR signalling dynamics inferred from fluorescence spectroscopy and imaging. *Curr Opin Struct Biol* **55**, 204-211, doi:10.1016/j.sbi.2019.05.007 (2019).

17 Bugge, K. *et al.* Interactions by Disorder - A Matter of Context. *Front Mol Biosci* **7**, 110, doi:10.3389/fmolb.2020.00110 (2020).

18 Artsimovitch, I., Svetlov, V., Murakami, K. S. & Landick, R. Co-overexpression of Escherichia coli RNA polymerase subunits allows isolation and analysis of mutant enzymes lacking lineage-specific sequence insertions. *J Biol Chem* **278**, 12344-12355, doi:10.1074/jbc.M211214200 (2003).