Supplementary Information for

Distinct evolutionary trajectories of SARS-CoV-2 interacting proteins in bats and primates identify important host determinants of COVID-19

Marie Cariou, Léa Picard, Laurent Guéguen, Stéphanie Jacquet, Andrea Cimarelli, Oliver I Fregoso, Antoine Molaro, Vincent Navratil, Lucie Etienne

Lucie ETIENNE
Email: lucie.etienne@ens-lyon.fr

This PDF file includes:

- Supplementary text
- Figures S1 to S12
- Tables S1 to S3
- Legends for Dataset S1
- SI References

Other supplementary materials for this manuscript include the following:

- Datasets S1
Supplementary Information Text

Results and Discussion

ZNF318 has undergone positive selection at three sites in primates, but not in bats

ZNF318 (NP_055160, also known as TZF, testicular zinc finger protein) is a putative RNA binding Matrin-type zinc finger protein. Matrin-type zinc fingers are RNA interacting domains found most notably in matrins and U1 small nuclear ribonucleoprotein C (InterPro: IPR003604). Aside from two central matrin-type ZNF domains, ZNF318 interacts with the androgen receptor and forms homodimers through its N-terminal tail (Figure S10A) (1). Finally, despite accounting for about half the protein, the Proline-rich C-terminal domain of ZNF318 has currently no ascribed function (Figure S11A). While the precise molecular function of ZNF318 remains unknown, it was shown to interact with the HUSH chromatin silencing complex and contributes to splicing-coupled transposon, and potentially latent HIV, silencing (2, 3). Antiviral functions have also been described for some human matrin genes (4).

Our DGINN screens identified signatures of positive selection in ZNF318 from primates, but not in bat species. While we confirmed the presence of ZNF318 orthologs in 29 primate and 16 bat species, many were 5' truncated. Thus, we used 5' trimmed CDS for in-depth selection analyses (Table 1). Unlike the majority of codons that are highly conserved (mean dN/dS of ~0.3), the C-terminal portion of ZNF318 displays many non-synonymous substitutions between orthologs. Statistical analyses using CodeML identified three sites under positive selection in this domain (M7 vs M8, BEB > 0.95): residues 1481, 1756 and 1908 in the corresponding human full-length protein (NP_055160, Figure S11, Table 1). There does not seem to be strong co-evolutionary signatures between these three residues and the combination of sites found in human (“V”, “R” and “P”) arose once in the last common ancestor with chimps and bonobos (Figure S11B). Of the three sites,
non-synonymous substitutions altering the proline residue 1908 might have strongest structural impact. Interestingly, these residues show some level of variation in bats, with bias for Threonines and Alanines at orthologous positions of sites 1481 and 1908.

While the function of ZNF318 during SARS-CoV2 infection remains to be determined, our finding of positive selection in primates suggest that it contributed to host adaptation in this lineage. Moreover, our analysis identified ZNF318 Proline-Rich domain as a putative functional interface with viruses. Hence, it would be interesting to investigate if this domain carries functions relating only to viral infections, including with non-coronaviruses (Figure 2), or if it is also involved in ZNF318 other cellular functions.

**Rapid evolution of the Prim-Pol primase complex (POLA1, PRIM1, PRIM2) in primates and bats**

Initiation of DNA replication in eukaryotes is dependent on the multisubunit primase-polymerase alpha (Prim-Pol α) complex that is responsible for the *de novo* synthesis of RNA/DNA primers on both the leading and lagging strands. Primase consists of the small catalytic subunit PRIM1 (p49) and the large regulatory subunit PRIM2 (p58), while polymerase consists of the catalytic subunit POLA1 (p180) and the accessory subunit POLA2 (p70). During primer synthesis, the primase subunits generate short RNA oligos, which are then subsequently extended with DNA by Pol α to be further elongated by replicative DNA polymerases. While many viruses encode their own Prim-Pol (5), including a putative CoV primase consisting of NSP7 and NSP8, SARS-CoV-2 NSP1 interacted with all four subunits of Prim-Pol α (6).

From our initial DGINN analysis, we identified positive selection on three out of the four Prim-Pol α subunits: POLA1 (four tests for primates and bats), PRIM1 (three tests for primates and two tests for bats), and PRIM2 (five tests for primates only). POLA2 was not
identified under positive selection in primates or bats (one test for bats and zero tests for primates).

To validate the finding and to precisely characterize gene evolution in primates and bats for POLA1, PRIM1, and PRIM2, we retrieved additional sequences for these orders and generated new high-quality codon alignments. POLA1 still showed significant signatures of positive selection across both primate and bat lineages (Table 1). PRIM1 maintained signatures of positive selection in bats (Table 1). PRIM2 maintained signatures of positive selection in primates for some tests, and the additional sequences identified positive selection in bats not seen with our initial pipeline (Table 1). Together this confirmed our initial analysis that POLA1, PRIM1, and PRIM2 are rapidly evolving in primate and bat lineages, with bats having more robust positive selection than primates.

Our codon-specific positive selection analysis identified multiple residues that have rapidly evolved in primates and bats. We specifically focused on sites that were found with more than one test of selection (Table 1). To determine if these sites of positive selection were found at the interface between subunits of the Prim-Pol α complex or at surface exposed sites, we mapped PS sites onto the human Prim-Pol α crystal structure (PDB: 5EXR) (7). None of the sites were found at Prim-Pol α complex interfaces, and while many were clustered close together, all were surface exposed or found in putative unstructured regions that were not present in the crystal structures (Figure S12). Together, this indicates that the positive selection identified for this complex is not being driven by complex formation and intra-complex coevolution, or DNA replication.

POLA1 is of particular interest as it has been modeled to dock with SARS-CoV-2 NSP1 (PDB 7OPL) (8)(9). However, none of the sites identified in our positive selection analysis were at the predicted POLA1-NSP1 interface (residues 615-629; Figure S12A-B), suggesting that NSP1 is not driving positive selection on POLA1. Instead, 7/9 PS sites in bats and primates were found in unstructured regions that were not present in the crystal
structure, with four sites (V235, D232, E239, and E240) all falling in a predicted disordered region of the protein (amino acids 232-251). That this unstructured region showed strong positive selection in both primates and bats could suggest that primate and bat POLA1 rapid evolution is being driven by similar unknown pressures in these distinct families.

Altogether, our positive selection analysis has identified that the Prim-Pol α complex is under strong positive selection in bats and primates, however this is not driven by Prim-Pol α complex formation or NSP1. Thus, it remains unclear what is driving PS on the Prim-Pol α complex, whether SARS-CoV interact with the Prim-Pol α complex as a whole or individual proteins from this complex, and why SARS-CoV-2 directly interact with nuclear host DNA replication machinery.

So why does SARS-CoV-2 recruit the Prim-Pol α complex? One possibility is that Prim-Pol α may have a role in the innate immune response to viral infection, and thus viruses directly antagonize components of this complex. Prim-Pol α was identified to interact specifically with NSP1 (6), which functions to inhibit host translation and innate immunity (10, 11). Pola1 is found in both the nucleus and cytoplasm, where it generates RNA-DNA hybrids that may be important for innate immune sensing (12). Loss of function mutations in POLA1 lead to increased pathogen infection, innate immune activation, and decreased number and effectiveness of NK cells (12, 13). While this supports a role of POLA1 in innate immunity against pathogens, roles for PRIM1 and PRIM2 have yet to be investigated.

It is also possible that SARS-CoV are usurping the host primase complex (or components of this complex) to enhance genome replication. SARS-CoV NSP7 and NSP8 are proposed to function as primase important for initiation of genome replication (14, 15). However, recent cryoEM structures of NSP7-8-12 complex suggest that NSP7 and NSP8 are too far from the RdRP NSP12 active site to act as primase (16). Thus, it is possible
that the host Prim-Pol α is recruited by NSP1 to further help initiation (and/or elongation) of CoV genome replication.
Figure S1. Identification of the SARS-CoV-2 interactome with signatures of positive selection in bats. A, Overview of the key steps of the bat VIP DGINN screen workflow. Details in Table S1. B, VIP-encoding genes identified under significant positive selection by at least three methods in DGINN (embedded legend). The percentage of positively selected sites in each VIP is shown on the right panel.
Figure S2. Identification of SARS-CoV-2 interactome with signatures of positive selection in primates. A, Overview of the key steps of the bat VIP DGINN screen workflow. Details in Table S1. B, VIP-encoding genes identified under significant positive selection by at least three methods in DGINN (embedded legend). The percentage of positively selected sites in each VIP is shown on the right panel. Of note, seven genes are false positives due to erroneous sequences or alignments: EMC1 (ER membrane protein complex subunit 1), MOV10 (Mov10 RISC complex RNA helicase), POR (cytochrome p450 oxidoreductase), PITRM1 (pitrilysin metallopeptidase 1), RAB14, RAB2A, and TIMM8B (translocase of inner mitochondrial membrane 8 homolog B).
Figure S3. Comparison of primate positive selection analyses between this study and Gordon et al. A, Comparison of the omega (dN/dS) values in PAML Codeml M0 model of the primate VIP genes calculated using the automated DGINN pipeline (y axis) and from Gordon et al study (x axis). Raw data were kindly provided by Janet Young, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. In black, the bisector. In red, the linear regression. B, Comparative analysis of the number of VIPs “under positive selection” in primates. In dark grey, the “strong and weak positively selected genes with benjamini-hochberg correction” from Gordon et al study (Codeml M8 vs M8a, p-value < 0.10); in medium grey, same from “uncorrected p-values”; in light grey, the genes identified by at least three methods in the primate DGINN screen. A total of 322 genes were in common between the two studies. C, VIPs under positive selection in Gordon et al with a DGINN score below three.
Figure S4. Comparative analyses of adaptive signatures in SARS-CoV-2 interactome in primates and in bats. A, Full table (associated with Figure 1D) showing the genes identified by x,y DGINN methods in bats and primates, respectively. B, Tanglegram of genes under positive selection in bats and primates. At the top, genes with the highest DGINN score (DGINN scores: purple, 5; blue, 4; teal, 3; green, 2; brown, 1; red, 0).
Figure S5. Meta-Analysis by Information Content (MAIC) scores of the VIP genes under positive selection. A-B, MAIC rank of VIPs identified under positive selection (by at least three methods in DGINN; DGINN scores of 3-5) in bats (A) and primates (B). The ACE2 and TMPRSS2 genes are highlighted in red as references. C-D, MAIC rank for all VIPs with, or without, evidence of positive selection (DGINN scores > or = 3, or < 3, respectively) in bats (C) and primates (D).
Figure S6. Biological processes of SARS-CoV-2 VIPs with (A-B) or without (C-D) evidence of positive selection in bats (blue) and primates (red). The graphs present the top 10 biological pathways retrieved after analysis on the Reactome database of the given set of VIPs.
Figure S7. Expression pattern of SARS-CoV-2 VIP-encoding genes with evidence of positive selection. A, Comparison of the expression level of genes under positive selection or not,
computed as the mean FPKM (Fragments per kilobase million) for each set of genes, in 29 human tissues. B-C, FPKM in human lung for all positively selected genes in bats (B) and primates (C). Genes of focus in our study are highlighted in black boxes.
Figure S8. Genetic variation of mammalian TMPRSS2 at the corresponding residues under positive selection in primates. Top, sequence logo of the positively selected sites in mammals that are naturally and experimentally permissive to SARS and MERS coronaviruses. Logos generated using Geneious. Bottom, Alignment of the corresponding amino acids in the mammalian species.
Figure S9. RIPK1 interactome. A, RIPK1 was used to interrogate the Reactome database and to retrieve RIPK1 cellular interactors that were subsequently subdivided according to their involvement in the indicated pathways. B, Described RIPK1 microbial antagonists or interactors from viruses (top) and bacteria (bottom).
**Figure S10. RIPK1 3D-structure prediction and amino acid variations in primates and bats.** A, Human and Rhinolophus ferrumequinum RIPK1 sequences were used to generate 3D-structure prediction models on RaptorX (grey hRIPK1 and orange bRIPK1, respectively). Magnified views of the structural homologies between the indicated domains. B, Sequence logo of bat RIPK1 from an amino-acid alignment of 18 species sequences, using WebLogo3. The y axis corresponds to the probability score. K residues are in blue, S and P residues are in purple. Positively selected sites...
are highlighted by triangles; the ones in red are those at regulatory sites (phosphorylation or ubiquitination) in human RIPK1. C, Alignment of bat (blue) and primate (red) amino-acid sequences of RIPK1 at sites under positive selection during bat evolution.
Figure S11. ZNF318 has undergone positive selection at three sites in primates, but not in bats. A, Schematic representation of human ZFN318 protein domains. Numbering of residues is relative to the human full-length protein sequence. The three residues subjected to positive selection in primates are shown with arrows. B, Amino-acids found at rapidly evolving sites across ZFN318 primate orthologs used in this study (top). Orthologs are organized according to the accepted species phylogeny. Bottom, same as top with the corresponding residues in bats. Dashes indicate an alignment gap due to an indel.
Figure S12. Rapid evolution of the Prim-Pol primase complex (POLA1, PRIM1, PRIM2) in primates and bats. A, Schematic representation of human DNA primase complex. B, Diagrams of predicted domains for POLA1, PRIM1, and PRIM2, respectively. Sites under positive selection in primates are represented by black triangles and sites under positive selection in bats are represented by purple triangles (Table 1). Codon numbering based on Homo sapiens genes. C, POLA1 amino acid variation at the positively selected sites in primates (top) and bats (bottom). Codon numbering and coloring as in B. Representation with Geneious.
Table S1. Number of genes screened in the initial primate and bat automatic screens – associated to Figures S1-S2.

|                                    | Primates | Bats |
|------------------------------------|----------|------|
| **Initial dataset (Gordon et al., 2020 + ACE2 + TMPRSS2)** | 334      | 334  |
| **Phylogenetic analyses**          |          |      |
| Failed runs                        | 5        | 4    |
| Duplicated genes retrieved by DGINN | 39       | 5    |
| Complete alignments and phylogenies| 368      | 335  |
| **Positive selection analyses**    |          |      |
| Results from 0 to 4 (out of 5) DGINN methods | 43       | 11   |
| Complete results (5/5 PS methods)  | 325      | 324  |
### Table S2. SARS-CoV-2 interacting proteins with no other known virus interactors

*HGNC names of the SARS-COV-2 VIPs with no other known virus interactors (*, except with other coronaviruses)*

| HGNC names         |
|--------------------|
| ACE2*              |
| GOLGA7*            |
| PRIM1*             |
| EIF4E2*            |
| GORASP1*           |
| MARK1*             |
| TRIM59*            |
| CISD3              |
| GGH                |
| INHBE              |
| PUSL1              |
| DPH5               |
| GCC2               |
| HS6ST2             |
| NUP58              |
| PLAT               |
| PTBP2              |
| TMEM39B            |
| TMPRSS2*           |
Table S3. Results from the comprehensive positive selection analyses of the genes of interest. Associated to Table 1. For each gene, are presented the results of the comprehensive phylogenetic and positive selection analyses: BUSTED, MEME, FUBAR, aBSREL from HYPHY/Datamonkey.com, M1vsM2, M7vsM8, M8a vs M8 from Bpp, and M1vsM2, M7vsM8, M8a vs M8 from PAML Codeml. The genes identified under positive selection are highlighted in grey. The sites considered under positive selection after the analyses are in “PSS aln” and “PSS in human ref”, corresponding to the site number in the codon alignment and the corresponding amino acid site in the human reference sequence. Alignments, trees, and interactive table are available at: https://virhostnet.prabi.fr/virhostevol/.

Legend details: Size, length of the codon alignment; n. sp., number of species included in the alignment; PS?, if the gene is under positive selection: Y, yes, N, no; p value, supporting a model under positive selection; PSS, positive selection sites; the cutoff for each method is given in the table; omega (PS), corresponds to the omega value in the positive selection class (dN/dS>1). ZNF318 and the proteins from the Primase complex are in Supplementary Information, and in Figures S11 and S12, respectively. *, for bpp M8 PSS analyses there were dozens of sites under positive selection due to the low omega value in the class w>1. For aBSREL, the branch identified under positive selection is given by the DGINN nomenclature (three letters from the genus and three letters from the species, na, not available.

| Sequence alignment | Gene | PS? | Length | n. sp. | PS?, if gene is under positive selection | p value | PSS, positive selection sites | Omega (PS) |
|--------------------|------|-----|--------|-------|----------------------------------------|--------|----------------------------|------------|
| PYCOI              | hts  | N   | 0.22   | 375, 376, 377, 379, 380, 381 | N       | 0.986                               | 0.010     |
| PYCOI primate     | 1503 | N   | 0.87   | 182, 183, 184, 185, 186, 187 | N       | 0.486                               | 0.129     |
| PSOLAI             | hts  | Y   | 0.03   | 105, 106, 107, 108, 109, 110 | Y       | 0.785                               | 0.582     |
| PSOLAI primate    | 1505 | N   | 0.87   | 182, 183, 184, 185, 186, 187 | N       | 0.486                               | 0.129     |
| PSOLAI             | hts  | N   | 0.22   | 375, 376, 377, 379, 380, 381 | N       | 0.986                               | 0.010     |
| PSOLAI primate 2 | 1503 | N   | 0.87   | 182, 183, 184, 185, 186, 187 | N       | 0.486                               | 0.129     |

Alignments, trees, and interactive table are available at: https://virhostnet.prabi.fr/virhostevol/.
Dataset S1 (separate file). Outputs of the GO term enrichment analyses performed by GOrilla to identify GO terms potentially enriched in positively selected or not positively selected genes, either compared to the interactome or to all human genes derived from ENSEMBL for both taxonomical groups. Headers are: Ontology, Ontology domain (Molecular function/Cellular Compartment/Biological process); GO Term, Go id; Description, Go description; P-value, P-value of the enrichment test; FDR q-value, FDR of the enrichment test. Tabs are as follow:

S1A. GO_334SC2VIP_vs_all : Enrichment test comparing 334 genes from the interactome to all human genes
S1B. GO_PS_vs_334SC2VIP_bats : Enrichment test comparing genes under positive selection in bats to the 334 genes from the interactome
S1C. GO_PS_vs_all_bats : Enrichment test comparing genes under positive selection in bats to all human genes
S1D. GO_nonPS_vs_all_bats : Enrichment test comparing genes NOT under positive selection in bats to all human genes
S1E. GO_PS_vs_334SC2VIP_primates : Enrichment test comparing genes under positive selection in bats to the 334 genes from the interactome
S1F. GO_PS_vs_all_primates : Enrichment test comparing genes under positive selection in primates to all human genes
S1G. GO_nonPS_vs_all_primates : Enrichment test comparing genes NOT under positive selection in primates to all human genes
SI References

1. R.-H. Tao, et al., Opposite effects of alternative TZF spliced variants on androgen receptor. *Biochemical and Biophysical Research Communications* **341**, 515–521 (2006).

2. G. Chougui, et al., HIV-2/SIV viral protein X counteracts HUSH repressor complex. *Nat Microbiol* **3**, 891–897 (2018).

3. C. H. Douse, et al., TASOR is a pseudo-PARP that directs HUSH complex assembly and epigenetic transposon control. *Nat Commun* **11**, 4940 (2020).

4. V. S. Yedavalli, K.-T. Jeang, Matrin 3 is a co-factor for HIV-1 Rev in regulating post-transcriptional viral gene expression. *Retrovirology* **8**, 61 (2011).

5. T. A. Guilliam, B. A. Keen, N. C. Brissett, A. J. Doherty, Primase-polymerases are a functionally diverse superfamily of replication and repair enzymes. *Nucleic Acids Res* **43**, 6651–6664 (2015).

6. D. E. Gordon, et al., A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).

7. A. G. Baranovskyi, et al., Mechanism of Concerted RNA-DNA Primer Synthesis by the Human Primosome. *J Biol Chem* **291**, 10006–10020 (2016).

8. M. L. Kilkenny, et al., Structural basis for the interaction of SARS-CoV-2 virulence factor nsp1 with DNA polymerase α–primase. *Protein Science* **31**, 333–344 (2022).

9. A. Chaudhuri, Comparative analysis of non structural protein 1 of SAR-CoV2 with SARS-CoV1 and MERS-CoV: An in silico study. *J Mol Struct* **1243**, 130854 (2021).

10. K. Narayanan, et al., Severe Acute Respiratory Syndrome Coronavirus nsp1 Suppresses Host Gene Expression, Including That of Type I Interferon, in Infected Cells. *Journal of Virology* **82**, 4471–4479 (2008).

11. K. Schubert, et al., SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat Struct Mol Biol* **27**, 959–966 (2020).

12. P. Starokadomskyy, et al., DNA polymerase-α regulates type I interferon activation through cytosolic RNA:DNA synthesis. *Nat Immunol* **17**, 495–504 (2016).

13. P. Starokadomskyy, et al., NK cell defects in X-linked pigmentary reticulate disorder. *JCI Insight* **4**, e125688.

14. E. Konkolova, M. Klima, R. Nencka, E. Boura, Structural analysis of the putative SARS-CoV-2 primase complex. *J Struct Biol* **211**, 107548 (2020).

15. Y. Zhai, et al., Insights into SARS-CoV transcription and replication from the structure of the nsp7–nsp8 hexadecamer. *Nat Struct Mol Biol* **12**, 980–986 (2005).

16. R. N. Kirchdoerfer, A. B. Ward, Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat Commun* **10**, 2342 (2019).