nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-25796-w OPEN

Characterising proteolysis during SARS-CoV-2 infection identifies viral cleavage sites and cellular targets with therapeutic potential

Bjoern Meyer[1], Jeanne Chiaravalli[2], Stacy Gellenoncourt[3], Philip Brownridge[4], Dominic P. Bryne[5], Leonard A. Daly[4], Arturas Grauslys[6], Marius Walter[7], Fabrice Agou[2], Lisa A. Chakrabarti[3], Charles S. Craik[8], Claire E. Eyers[4], Patrick A. Eyers[5], Yann Gambin[9], Andrew R. Jones[5], Emma Sierecki[9], Eric Verdin[7], Marco Vignuzzi[1] & Edward Emmott[4]✉

SARS-CoV-2 is the causative agent behind the COVID-19 pandemic, responsible for over 170 million infections, and over 3.7 million deaths worldwide. Efforts to test, treat and vaccinate against this pathogen all benefit from an improved understanding of the basic biology of SARS-CoV-2. Both viral and cellular proteases play a crucial role in SARS-CoV-2 replication. Here, we study proteolytic cleavage of viral and cellular proteins in two cell line models of SARS-CoV-2 replication using mass spectrometry to identify protein neo-N-termini generated through protease activity. We identify previously unknown cleavage sites in multiple viral proteins, including major antigens S and N: the main targets for vaccine and antibody testing efforts. We discover significant increases in cellular cleavage events consistent with cleavage by SARS-CoV-2 main protease, and identify 14 potential high-confidence substrates of the main and papain-like proteases. We show that siRNA depletion of these cellular proteins inhibits SARS-CoV-2 replication, and that drugs targeting two of these proteins: the tyrosine kinase SRC and Ser/Thr kinase MYLK, show a dose-dependent reduction in SARS-CoV-2 titres. Overall, our study provides a powerful resource to understand proteolysis in the context of viral infection, and to inform the development of targeted strategies to inhibit SARS-CoV-2 and treat COVID-19.

[1] Viral Populations and Pathogenesis Unit, CNRS, UMR 3569, Institut Pasteur, CEDEX 15 Paris, France. [2] Chemogenomic and Biological Screening Core Facility, C2RT, Departments of Cell Biology & Infection and of Structural Biology & Chemistry, Institut Pasteur, CEDEX 15 Paris, France. [3] CIVIC Group, Virus & Immunity Unit, Institut Pasteur and CNRS, UMR 3569 Paris, France. [4] Centre for Proteome Research, Department of Biochemistry & Systems Biology, Institute of Systems, Molecular & Integrative Biology, Biosciences Building, Crown Street, University of Liverpool, Liverpool L69 7ZB, UK. [5] Department of Biochemistry & Systems Biology, Institute of Systems, Molecular & Integrative Biology, Biosciences Building, Crown Street, University of Liverpool, Liverpool L69 7ZB, UK. [6] Computational Biology Facility, LIV-SRF, Institute of Systems, Molecular & Integrative Biology, Biosciences Building, Crown Street, University of Liverpool, Liverpool L69 7ZB, UK. [7] Buck Institute for Research on Aging, Novato, CA 94945, USA. [8] Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA. [9] EMBL Australia Node for Single Molecule Sciences, and School of Medical Sciences, Botany Road, The University of New South Wales, Sydney, NSW 2052, Australia. ✉email: e.emmott@liverpool.ac.uk

SARS-CoV-2 emerged into the human population in late 2019, as the latest human coronavirus to cause severe disease following the emergence of SARS-CoV and MERS-CoV over the preceding decades[1,2]. Efforts to develop vaccines and therapeutic agents to treat COVID-19 are already yielding results; however, it is widely expected that this first generation of treatments might provide imperfect protection from disease. As such, in-depth characterisation of the virus and its interactions with the host cell can inform current and next-generation efforts to test, treat, and vaccinate against SARS-CoV-2. Past efforts in this area have included the proteome, phosphoproteome, ubiquitome, and interactome of SARS-CoV-2 viral proteins and infected cells[3–9]. Proteolytic cleavage plays a crucial role in the life cycle of SARS-CoV-2, and indeed most positive-sense RNA viruses. Inhibitors targeting both viral and cellular proteases have previously shown the ability to inhibit SARS-CoV-2 replication in cell culture models[10–13]. Here we present a first unbiased study of proteolysis during SARS-CoV-2 infection, and its implications for viral antigens, as well as cellular proteins that may represent options for antiviral intervention.

Proteolytic cleavage of the two coronavirus polyproteins generates the various viral proteins needed to form a replication complex, required for transcription and replication of the viral genome and subgenomic mRNAs. The key viral enzymes responsible are the papain-like (PLP, nsp3) and main proteases (Mpro, nsp5). Aside from cleaving viral substrates, these enzymes can also act on cellular proteins, modifying or neutralising substrate activity to benefit the virus. A recent study highlighted the ability of the viral proteases to cleave proteins involved in innate immune signaling including IRF3, NLRP12, and TAB1[14]. However, there has yet to be an unbiased study to identify novel substrates of the coronavirus proteases in the context of viral infection. The identification of such substrates can identify cellular enzymes or pathways required for efficient viral replication that may represent suitable targets for pharmaceutical repurposing and antiviral intervention for the treatment of COVID-19.

Viral proteins can also be the targets of cellular proteases, with the most prominent example for coronaviruses being the cleavage of the spike glycoprotein by the cellular proteases FURIN, TMPRSS2, and Cathepsins[10,11,15,16], but the exact cleavage sites within spike for most of these individual cellular proteases are not yet characterised. Proteolytic processing can also be observed for other coronavirus proteins, for example, signal peptide cleavage of SARS-CoV ORF7A[17] and caspase cleavage of the nucleocapsid protein[18,19]. The spike glycoprotein especially, forms the key or sole component of the vaccines currently in use. For a functional immune response, it is vital that the antigens presented to the immune system, as part of these vaccines, closely mimic those seen in natural infection. An understanding of any modifications to these antigens observed during natural infection, such as glycosylation, phosphorylation, and proteolytic cleavage, is critical to enable the rational design and validation of vaccine antigens and the selection of appropriate systems for their production. Currently a range of vaccine platforms are being explored, and certain platforms or delivery routes more likely to suffer from altered post-translational modification states than others[20].

Mass spectrometry-based proteomic approaches have already led to rapid advances in our understanding of SARS-CoV-2, with notable examples including the rapid release of the cellular interactome[6] and proximity interactome[7] for a majority of SARS-CoV-2 proteins, as well as proteomic[3,5], phosphoproteomic[4,8], and ubiquitomic analyses[9]. Larger scale-initiatives have been launched focusing on community efforts to profile the immune response to infection, and provide in-depth characterisation of viral antigens[21]. Mass spectrometry has particular advantages for investigation of proteolytic cleavage as analysis can be conducted in an unbiased manner, and identify not only the substrate, but also the precise site of proteolytic cleavage[22].

In this work we apply mass spectrometry-based methods for N-terminomics to study proteolysis and the resulting proteolytic proteoforms generated in the context of SARS-CoV-2 infection, enabling the identification of novel cleavage and processing sites within viral proteins. We discover several of these novel viral cleavage sites show altered cleavage following treatment with the cathepsin/calpain inhibitor calpeptin. We also identify cleavage sites within cellular proteins that match the coronavirus protease consensus sequences for Mpro and PLP, show temporal regulation during infection, are cleaved in vitro by recombinant Mpro and PLP, and demonstrate these proteins are required for efficient SARS-CoV-2 replication. These SARS-CoV-2 protease substrates include proteins that can be targeted with drugs in current clinical use to treat other conditions[23]. Indeed, we demonstrate potent inhibition of SARS-CoV-2 replication with two compounds that are well-established chemical inhibitors of the SARS-CoV-2 protease substrates SRC and myosin light chain kinase (MYLK).

Results

Proteomic analysis of SARS-CoV-2-infected cell lines identifies alterations to the N-terminome. To investigate proteolysis during SARS-CoV-2 infection, N-terminomic analysis at various timepoints during the course of SARS-CoV-2-infected Vero E6 and A549-Ace2 cells (Fig. 1a) was performed. Vero E6 cells are an African Green Monkey kidney cell line commonly used for the study of a range of viruses, including SARS-CoV-2 which replicates in this cell line to high titres. A549-Ace2 cells are a human lung cell line, which has been transduced to overexpress the ACE2 receptor to allow for SARS-CoV-2 entry. Cells were infected in biological triplicates at a multiplicity of infection (MOI) of 1, and harvested at 4 timepoints (0, 6, 12, and 24 h) post-infection. Mock-infected samples were collected at 0 h and 24 h post-infection. These timepoints were chosen to cover SARS-CoV-2 infection from virus entry, over replication to virus egress: RNA levels increased from 9 h post-infection (Fig. 1b), protein levels showed steady increases throughout infection (Fig. 1c), and viral titres increased at the 24 h timepoint (Fig. 1d). These features were shared in both cell lines, with the Vero E6 cells showing greater RNA and protein levels, as well as viral titres compared with the A549-Ace2 cells.

Analysis of the N-termini-enriched samples was performed by LC-MS/MS following basic reverse phase fractionation. For the purposes of this analysis, neo-N-termini were taken to be those beginning at amino acid 2 in a given protein or later. By this definition these neo-N-termini will include those with post-translational removal of methionine, signal peptide cleavage, as well as those cleaved by viral or cellular proteases. The modified N-terminomic enrichment strategy used[22] employed isobaric labelling (TMTpro) for quantification as this permitted all samples to be combined prior to enrichment, minimising sample variability. This strategy meant that only those peptides with a TMTpro-labelled N-terminus or lysine residue were quantified. As only unblocked N-termini are labelled with undecanal, this approach results in the selective retention of undecanal-tagged tryptic peptides on C18 in acidified 40% ethanol, with N-terminal and neo-N-terminal peptides enriched in the unbound fraction[22].

Quality filtering of the dataset was performed (Supplementary Fig. 1), infected and mock samples separated by PCA and 0 h Mock, 0 h infected, and 6 h infected clustered together, and away from the 12 h and 24 h infected samples (Supplementary Fig. 1a–d). With the exception of the enriched Vero E6 dataset the 24 h mock sample clustered with the 0 h mocks. The Vero E6 24 h Mock clustered away from the 0 h and infected samples,

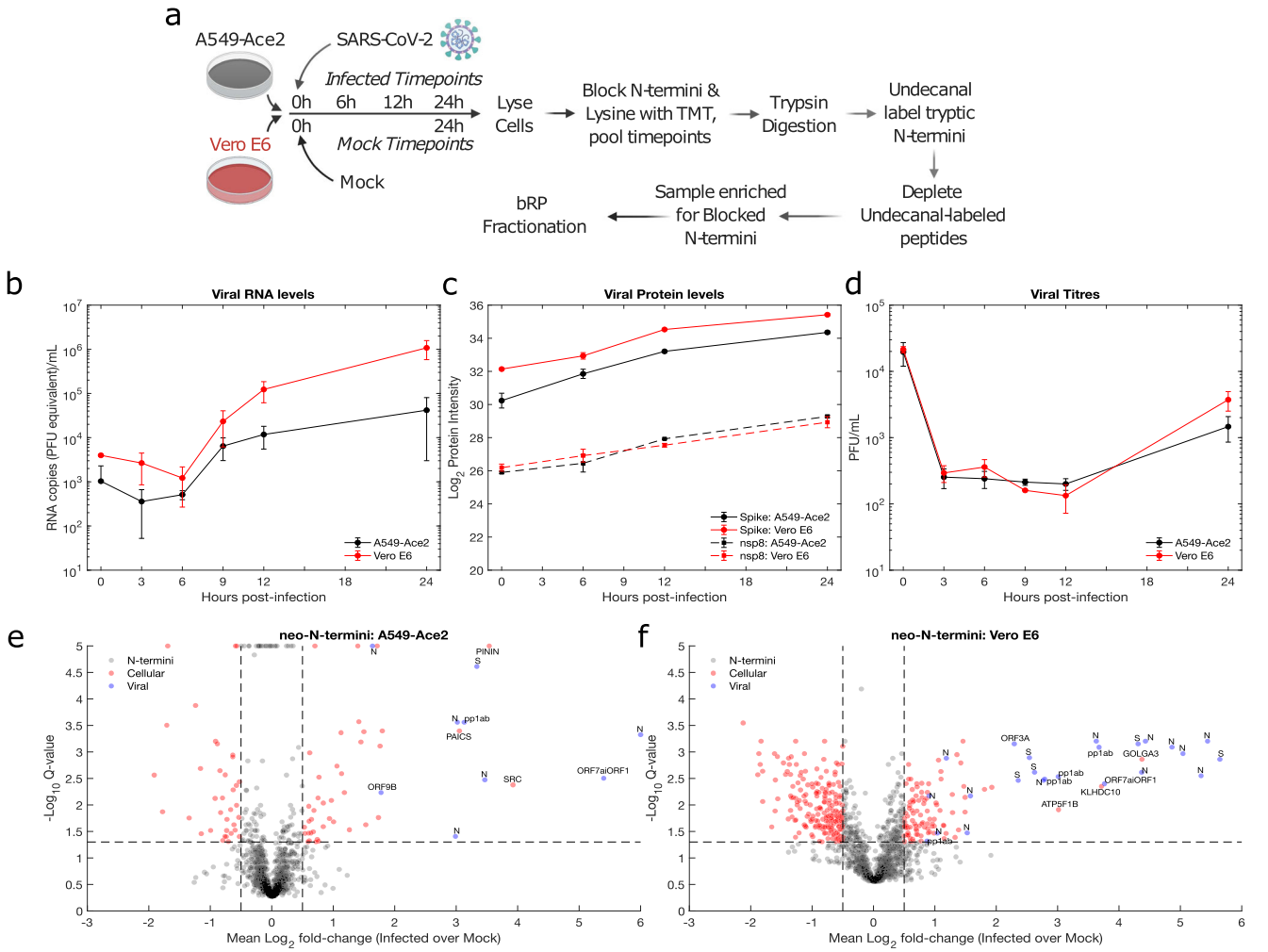


Fig. 1 N-terminomic analysis of SARS-CoV-2 infection of A549-Ace2 and Vero E6 cells. a Experimental design. **b** Viral RNA levels were determined by qRT-PCR ($n = 3$ biological replicates). **c** Protein levels were determined based on the TMTpro fractional intensity of the total protein intensity for the unenriched proteomic samples ($n = 3$ biological replicates). **d** Infectious virus production (PFU, $n = 3$ biological replicates). Error bars for **b**–**d** represent standard deviation. **e** A549-Ace2 and **f** Vero E6 neo-N-terminomic analysis reveals significant increases in peptides corresponding to viral and cellular neo-N-termini, where neo-N-termini must begin from amino acid 2 or later. *P* values were obtained by two-tailed unpaired *t*-test, correction for multiple-hypothesis testing to obtain Q-values was performed as described Storey (2002)[70]. bRP basic reverse phase fractionation, PFU plaque-forming units. The vertical and horizontal lines correspond to fold change (0.5) cut-off and Q-value (0.05) cut-offs, respectively.

which may reflect regulation due to cell confluence as this was not observed with the paired unenriched sample. Sample preparation successfully enriched for blocked N-termini consisting of acetylated, pyroglutamate N-termini, and TMTpro-labelled N-termini (Supplementary Fig. 1), and blocked N-termini were more abundant in the enriched samples. In both datasets, TMTpro-labelled N-termini represent ~50% of the blocked N-termini, with the rest split evenly between pyroglutamate and N-terminal acetylation (Supplementary Fig. 1). After filtering, over 2700 TMTpro-labelled N-termini representing neo-N-termini were identified from each cell line. While the experimental design chosen is based around minimising missing data within rather than between the two datasets, 497 neo-N-termini were common to the two cell lines (Supplementary Fig. 2), and show positive correlation (Pearson's ρ, 0.64–0.7).

When the 24 h infected and mock-infected timepoints were compared, both cellular and viral neo-N-termini in A549-Ace2 (Fig. 1e) and Vero E6 cells (Fig. 1f) were identified as showing significant alterations in their abundance. In line with expectation, N-termini from viral proteins were solely identified as showing increased abundance during infection in both cell lines. N-termini from cellular proteins showed both increased and decreased abundance during infection. We reasoned that those neo-N-termini showing increased abundance would include viral neo-N-termini, as well as those cellular proteins cleaved by the SARS-CoV-2 PLP and Mpro proteases. For this study we therefore focused specifically on viral N-termini and those cellular neo-N-termini identified as showing significantly increased abundance (*t*-test, multiple-hypothesis testing corrected Q value ≤ 0.05) during infection.

Novel proteolytic processing of SARS-CoV-2 proteins is observed during infection. The 30 kb SARS-CoV-2 genome encodes a large number of proteins including two long polyproteins formed through ribosomal frameshifting, the structural proteins S, E, M, and N and a range of accessory proteins (Fig. 2). Coronavirus proteins, in line with those of other positive-sense RNA viruses are known to undergo post-translational modifications, including proteolytic cleavage in some cases. Across all datasets we identified the S, M, and N structural proteins, with the exception of E, which has also not been observed in other proteomics datasets due to both short length and sequence

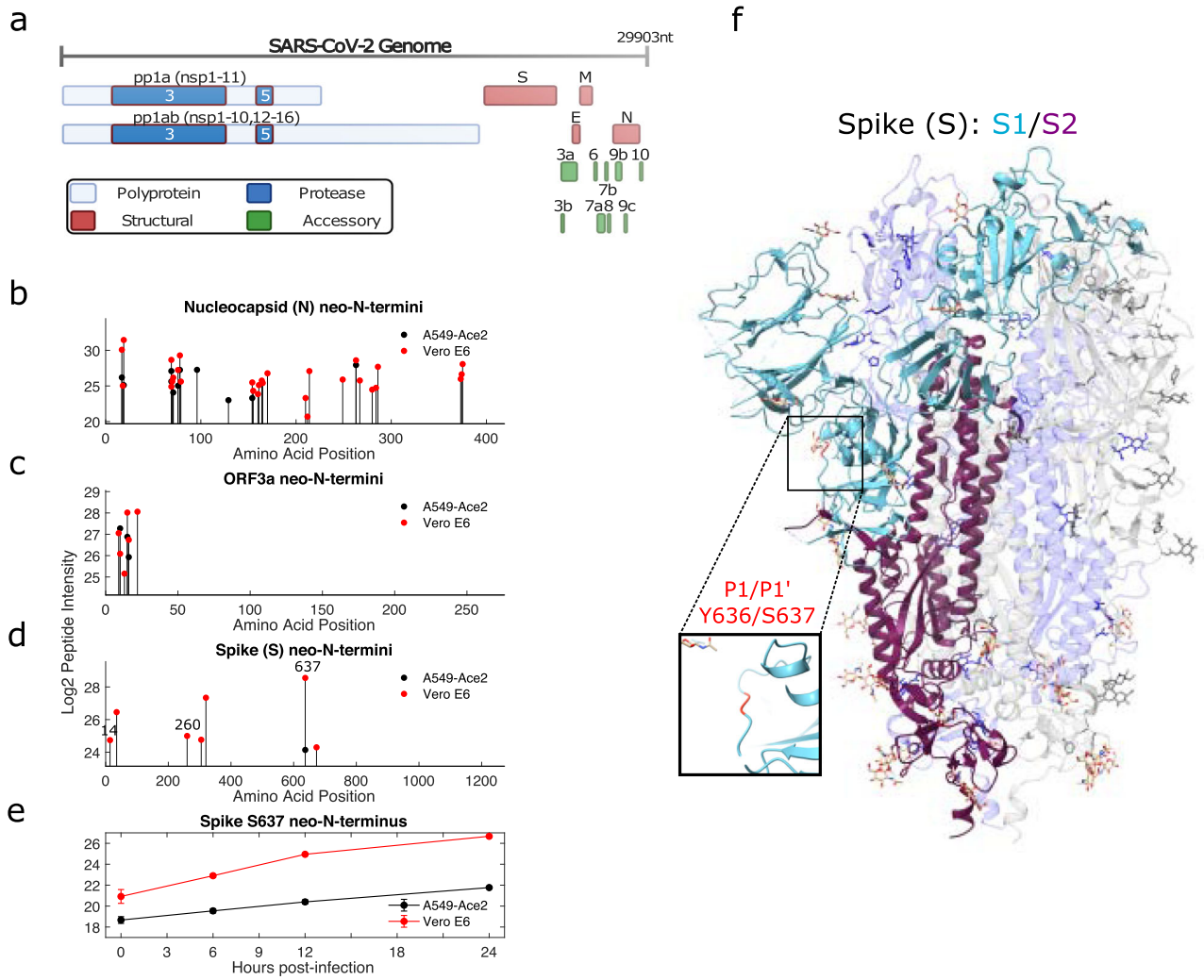


Fig. 2 Proteolysis of viral proteins during SARS-CoV-2 infection. a Schematic of the SARS-CoV-2 genome and proteome, with the nsp3 (PLP) and nsp5 (MPro) highlighted. Proteolytic processing of SARS-CoV-2 proteins during infection of A549-Ace2 and Vero E6 cells includes **b** extensive cleavage of the nucleocapsid protein, **c** N-terminal processing of the ORF3a putative viroporin, and **d** a novel cleavage site between Y636 and S637 in spike, N-terminal of the FURIN cleavage site. **e** The abundance of the S637 spike neo-N-terminus increases over the infection timecourse ($n = 3$, error bars show standard deviation from 3 biological replicates). **f** This cleavage site is present on a flexible region, C-terminal of the RBD (PDB: 6X6P).

composition[3,5]. We identified the ORF3a, ORF6, ORF8, and ORF9b accessory proteins, and all domains of the polyprotein aside from nsp6, 7, and 11.

We first sought to characterise neo-N-termini from viral proteins to understand potential patterns of cleavage that might generate functional proteolytic proteoforms of the viral proteins. Our search database included recently identified noncanonical translation products[24], and neo-N-termini and N-termini were identified from eight viral proteins including the polyprotein (Fig. 2b–d; Supplementary Fig. 3). Of these the nucleocapsid (N), ORF3a accessory protein and spike were most prominent. More cleavage sites were observed from infected Vero E6 cells than A549-Ace2 cells, which is in line with expectation given the higher levels of viral protein expression, and superior infectivity of this cell line compared to the A549-Ace2 cell line, permitting detection of less abundant cleavage products.

The coronavirus N protein is highly expressed during infection, and also represents a major antigen detected by the host immune response. Prior studies have identified cleavage of the SARS-CoV N protein by cellular proteases[18,19], and our data identified multiple neo-N-termini consistent with proteolytic cleavage from both infected A549-Ace2 and Vero E6 cells (Fig. 2b). neo-N-termini common to both datasets include amino acids 17, 19, 69, 71, 76, 78, 154, and 263. Many of these cleavage sites were spaced closely together (e.g. 17/19, 69/71), consistent with a degree of further exoproteolytic processing. Some of these cleavage sites have subsequently been identified as autolysis products following extended incubations with N in vitro[25].

The ORF3a putative viroporin also shows N-terminal processing, possibly reflecting signal peptide cleavage (Fig. 2c). In a recent study, cryoEM of ORF3a in lipid nanodiscs did not resolve the first 39 N-terminal suggesting this region is unstructured[26]. We observed N-terminal processing sites in the first 22 residues of the protein, with neo-N-termini beginning at amino acids, 10, 13, and 16 identified in both datasets, giving a possible explanation for the lack of N-terminal amino acids in cryoEM experiments.

Proteolytic cleavage of the spike glycoprotein is of major interest as it plays an important role in virus entry, with different distributions of cellular proteases between cell types resulting in the usage of different entry pathways, as well as potentially changing availability of surface epitopes for antibody recognition. Key proteases include FURIN, TMPRSS2, and cathepsins, although in the latter two cases the actual cleavage sites targeted

by these enzymes to process spike into S1 and S2 remain unclear. Consistent with previous observations[3,5], we do not detect a neo-N-terminus deriving from the FURIN cleavage site as the trypsin digestion we employed would not be expected to yield peptides of suitable length for analysis. However, while beneficial for replication, FURIN cleavage is not essential and other cleavage events within spike can compensate[15,16]. We detect neo-N-terminal peptides from S637 in both datasets (Fig. 2d). In line with the pattern of viral gene expression observed in the unenriched datasets this neo-N-terminus showed consistent increases in abundance throughout the experimental timecourse (Fig. 2e). S637 is located on a flexible loop near the FURIN cleavage site (Fig. 2f), suggesting it is accessible for protease cleavage[27]. A mass spectrum for the S637 neo-N-terminus from the A549-Ace2 dataset is shown in Supplementary Fig. 4a, the same peptide was observed with both 2+ and 3+ charge states in the Vero E6 dataset, and with a higher Andromeda score (124.37 vs. 104.82). Intriguingly, S637 was identified as a phosphorylation site in Davidson et al.[3]. As phosphorylation can inhibit proteolytic cleavage when close the cleavage site, this suggests potential post-translational regulation of this cleavage event.

Further neo-N-termini from spike were identified in the Vero E6 dataset alone, including a neo-N-terminus beginning at Q14 Supplementary Fig. 4b. This is slightly C-terminal of the predicted signal peptide, which covers the first 12 amino acids. This peptide featured N-terminal pyroglutamic acid formed by cyclisation of the N-terminal glutamine residue. The peptide does not follow an R or K residue in the spike amino acid sequence and thus represents non-tryptic cleavage. The absence of TMTpro labelling at the N-terminus suggests that this N-terminus was blocked prior to tryptic digestion, with this modified N-terminus preventing TMTpro modification. Artifactual cyclisation of N-terminal glutamine or glutamic acid residues typically results from extended trypsin digestion and acidic conditions[28]. However, the order of labelling and digestion steps in our protocol, and non-tryptic nature of this peptide suggests that this N-terminal pyroglutamic acid residue is an accurate reflection of the state of this neo-Nterminus in the original biological sample. While this is to our knowledge, the first observation of this N-terminal modification and signal peptide cleavage site for spike during infection, cleavage and cyclisation of Q14 has been observed with recombinantly produced SARS-CoV-2 spike in HEK 293[29] and CHO cells[30]. Three further N-terminal pyroglutamic acid residues were identified in SARS-CoV-2 proteins (N, ORF7A) within the Vero E6 dataset and can be found in Supplementary data 2.

We detected viral neo-N-termini and N-termini in M, ORF7a, ORF9b, and pp1ab. Due to conservation with SARS-CoV ORF7a, the first 15 residues of SARS-CoV-2 ORF7a are expected to function as a signal peptide, which is posttranslationally cleaved[17,31]. neo-N-termini were identified in both datasets consistent with this hypothesis. Due to inclusion of the ORF7a iORF1 proposed N-terminal truncation of ORF7a, which lacks the first two amino acids in ORF7a in the SARS-CoV-2 sequences used for data analysis, the start position of this neo-N-terminal peptide is given as 14[24]. However, this would be position 16 in ORF7a, consistent with removal of the signal peptide (MKIIL-FLALITLATC, in Uniprot P0DTC7), and conserved with that in SARS-CoV ORF7a.

The native N-terminus of ORF9b was also identified, and several sites mapping to the replicase polyprotein, including a conserved neo-N-terminus consistent with predicted nsp10-nsp12 cleavage by Mpro. A neo-N-terminus consistent with nsp15-nsp16 cleavage by Mpro was identified in A549-Ace2 cells, and several internal neo-N-termini deriving from nsp1, −2, and −3 were also observed, though not common to both datasets. All the viral neo-N-termini and N-termini identified in this study can be found in Supplementary data 1 (A549-Ace2) and 2 (Vero E6), respectively. Supplementary data 3 includes all viral peptides identified in this study in both enriched and unenriched datasets.

neo-N-termini and current Variants of Concern (VOCs). Currently there is extensive interest in mutations present in emerging VOCs and Variants of Interest (VOIs), and using our understanding of virus biology to predict how such mutations could alter protein function or antibody evasion.

To investigate if the viral neo-N-termini we identified may be under selective pressure, mutations characteristic of current VOC/VOIs were identified from covariants.org[32] on the 24 May 2021. These included B.1.1.7, B.1.351, B.1.427/9, B.1.525, B.1.526, P.1, B.1.617.1, and B.1.617.2. Compared to the a background proteome of all viral peptides identified in the analysis, neo-N-termini were not overrepresented near characteristic variant mutations (KS test, $p = 0.146$) Supplementary Fig. 6. However, not all neo-N-termini may be under selective pressure, and in some cases mutations could be lethal to the virus so would not be observed in variants.

While our data do not suggest global enrichment of neo-N-termini proximal to characteristic variant mutations, nine of our neo-N-termini were within five amino acids of a characteristic variant mutation (Supplementary data 6), and 17 were within 10 amino acids (Supplementary data 7). Those neo-N-termini within five amino acids of variant mutations are of particular interest as the variant mutations could readily alter proteolytic cleavage. These variant mutations include multiple sites in S (13, 18, 19, 677), N (12, 80, 205, 377), and ORF3a[25]. Of particular interest is the S13I mutation in spike in strain B.1.427/B.1.429, immediately preceding the pyroglutamate-modified Q14 neo-N-terminus. Recent research has indicated that this is tolerated in spite of blocking signal peptide cleavage to generate the Q14 neo-N-terminus, with signal peptide cleavage instead generating a neo-Nterminus at V16 resulting in loss of antibody binding and structural rearrangements within spike due to loss of disulphide bond formation between C15 and C136[33]. As further variants emerge, the incorporation of post-translational modification data from studies such as this can support efforts to predict phenotypes from genetic data on emerging variants.

Novel SARS-CoV-2 cleavage sites are sensitive to calpeptin and a mutation proximal to the 637 cleavage site results in a higher fraction of cleaved spike in purified pseudovirus and enhanced cell entry. The N-terminomics experiments above successfully identified multiple previously uncharacterised proteolytic cleavage sites within viral proteins. However, we have limited information on the identity of the causal proteases behind these cleavage events. To address this, we performed a further N-terminomics experiment comparing the relative abundance of these cleavage sites and viral proteins following treatment with specific protease inhibitors Fig. 3a. The first, camostat mesylate is currently in clinical trials to treat COVID-19 disease and acts on TMPRSS2 and trypsin. It should be noted that the cell lines our study focuses on are not believed to express significant levels of TMPRSS2, with virus entry being cathepsin-dependent. As such camostat mesylate was included as a control. The second, calpeptin inhibits cathepsin and calpain cleavage. The experiment was performed in Vero E6 cells as the majority of viral cleavage sites identified in the first dataset were found in this cell line. Inhibitors were added at 12 h post-infection with the aim of reducing proteolytic cleavage rather than inhibiting virus replication per se by permitting viral replication to proceed unimpeeded for the first 12 h. Rising protein and RNA levels at this

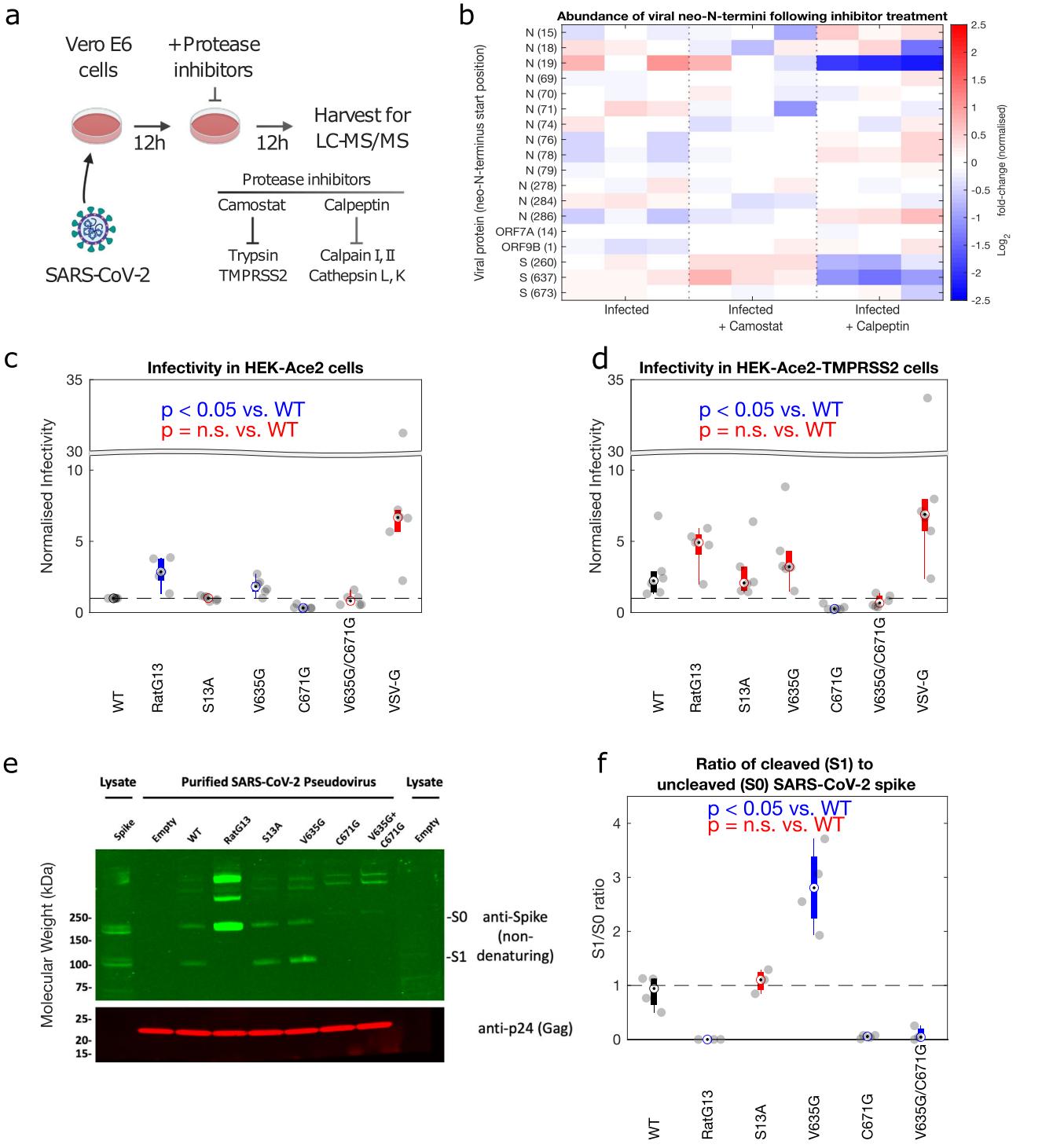


Fig. 3 Several viral neo-N-termini show sensitivity to specific protease inhibitors and a spike 637-proximal mutation alters viral entry in TMPRSS2-ve cells. a Experimental design for N-terminomics of SARS-CoV-2 infection in the presence of protease inhibitors. **b** Abundance of viral neo-N-termini in infected cells ± inhibitors. Data normalised to total levels of the relevant viral protein. $n = 3$ biologically independent samples. Pseudovirus entry assay conducted in **c** HEK-Ace2 and **d** HEK-Ace2-TMPRSS2 cells. The infectivity of lentivectors (LV) pseudotyped with the different spike mutants was normalised to that of WT in the HEK-Ace2 cell line. $n = 6$ (RatG13 $n = 5$) biologically independent samples. **e** Western blotting of the pseudovirus stocks used in **c** and **d** confirms spike expression and incorporation into lentiviral particles. **f** Densitometry analysis of spike western blotting data, examining the ratio between uncleaved (S0) and cleaved (S1) portions of the spike protein present in purified pseudotyped lentivirus stocks ($n \geq 3$ biological replicates). Boxplot minima/maxima represent the furthest non-outlier datapoints, centre the median, and bounds of box the interquartile range. Outliers are defined as datapoints >1.5 times the interquartile range from the bottom or top of the box. Unpaired Welch's *t*-tests, which do not assume equal variance were used for statistical analyses.

timepoint Fig. 1b, c indicate that adding inhibitors at this timepoint avoids inhibition of viral entry, where both cathepsin- and TMPRSS2-based entry have been previously recorded, depending on the model system used. Samples were then harvested at 24 h post-infection Fig. 3a.

Quality control of this dataset against showed tight clustering of the relevent samples, with infected samples clustering away from the mock-infected cells (Supplementary Fig. 5a, b). This dataset identified fewer quantifiable N-termini (Supplementary Fig. 5c–f), but these included the key cleavage sites in S and N.

Analysis of viral protein levels showed that when added at this late time post-infection viral protein levels were simiilar but did show some differences with the untreated infected samples Supplementary Fig. 5g. ORF9B showed significantly reduced abundance in protease inhibitor-treated infected cells compared to infected but untreated cells (Camostat: *t*-test, $p = 0.0031$; Calpeptin: *t*-test, $p = 0.0012$). ORF3A showed significantly increased abundance compared to untreated infected cells in the camostat-treated cells alone (*t*-test, $p = 0.0016$). Neither N nor S showed significant changes in their abundance with either inhibitor treatment.

Neo-N-termini corresponding to viral proteins were normalised to the abundance of the viral protein from which the neo-N-terminus was derived and can be seen in Fig. 3b. The largest changes were observed following calpeptin treatment, resulting in significantly reduced abundance of neo-N-termini beginning at N[19],S(260), and S(637). Both neo-N-termini from S (260, 637) showed increased abundance following treatment with camostat compared to untreated infected cells (*t*-test, $p = 0.0074$ and 0.0288, respectively). Two neo-N-termini within N (78, 286) show increased abundance in the calpeptin-treated infected cells (*t*-test, $p = 0.0045$, 0.0120). Reduced abundance of neo-N-termini following calpeptin inhibibion (e.g. N[19], S(260, 637)) is consistent with cleavage of these sites by cathepsin which is known to cleave S[10]. The enhanced cleavage of these sites following addition of camostat in Vero E6 cells suggests some increased diversion of S and N down a cathepsin-dependent cleavage pathway under conditions, which inhibit TMPRSS2 and trypsin. While this cell line lacks TMPRSS2 expression, clearly there is some alteration of proteolytic activity following camostat treatment, suggesting inhibition of other proteases in this system, which may have a compensatory or complementary function.

The same data lacking normalisation to total viral protein levels can be seen in Supplementary Fig. 5h. As expected given the lack of significant changes to total N and S protein levels, all sites highlighted in the previous paragraph maintained their direction of change relative to mock, and remained *t*-test significant in this unnormalised dataset ($p < 0.05$).

As a majority of these cleavage sites within viral proteins are novel, we lack a functional understanding of their role in viral infection. We sought to examine the importance of several of the novel cleavage sites found within the spike glycoprotein by mutating residues proximal to the cleavage sites and assessing their functions in a pseudovirus entry assay, utilising pseudotyped lentiviral vectors. Given our observed cleavage at Q14 we generated a S13A mutation, altering the P1 residue in this cleavage site following the nomenclature of Schechter and Berger[34]. For the cleavage sites at 637 and 671, given that cathepsin is known to cleave spike[10], and indeed the site at 637 showed sensitivity to calpeptin, a calpain and cathepsin inhibitor, we sought to modify cleavage through mutagenesis of the P2 residue within this cleavage site as the P2 site is considered important for cathepsin cleavage[35]. This approach generated V635G and C671G mutants. RatG13 spike, a mutant lacking the FURIN cleavage site due to a deletion of four residues (Δ681–684, ΔPPRA) was included as an additional control[36].

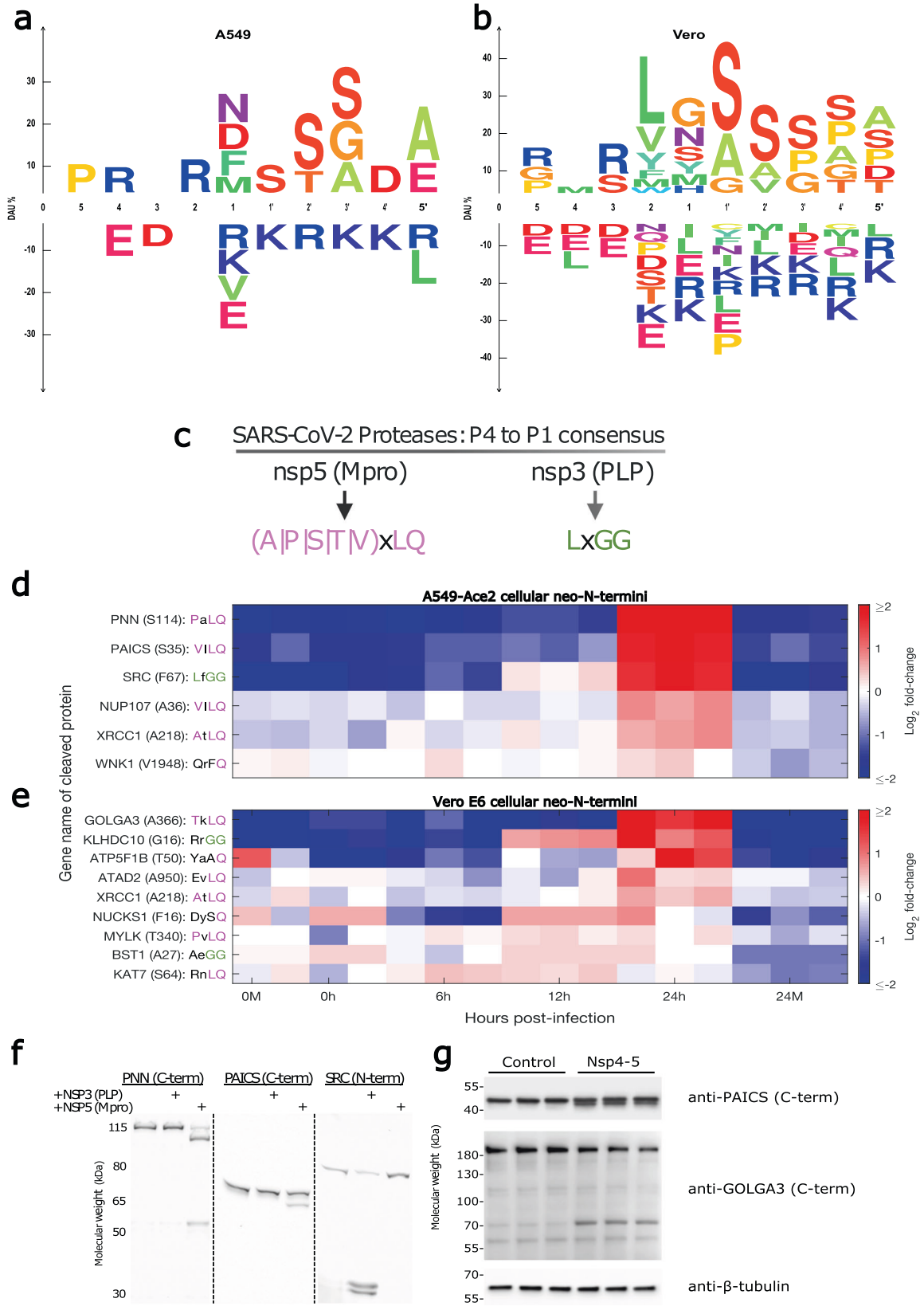
In HEK-ACE2 cells, both the V635G and RatG13 mutants showed significantly increased cell entry compared to wild-type (*t*-test $p < 0.01$), while the C671G mutant showed significantly decreased cell entry (Unpaired Welch's *t*-test, $p < 0.0001$) Fig. 3c. Entry for both the S13A and the double V635G/C671G mutant was not significantly different to wild-type Fig. 3c. Representative FACS plots and the gating strategy applied can be found in Supplementary Fig. 7.

The same pattern was observed in HEK-ACE2-TMPRSS2 cells, although the enhanced cell entry seen for the the V635G and RatG13 mutants did not reach significance Fig. 3d. This may reflect reduced importance of the V635G mutation in cells bearing high levels of TMPRSS2. Similarly while the pattern of partial recovery of the double V635G/C671G mutant remained visible, its entry remained significantly reduced (Unpaired Welch's *t*-test, $p < 0.05$) compared to wild-type Fig. 3d. While reproducible across multiple independent viral stocks and experiments, the phenotypes are modest (two-fold change) in both cell types, except for the marked infectivity defect of the C671G mutant. Representative FACS plots and the gating strategy applied can be found in Supplementary Fig. 8.

Western blotting of purified pseudovirus particles confirmed expression and incorporation of spike Fig. 3e. Notably constructs containing the C671G mutation showed limited incorporation and defective spike processing. While this could be in part due to cleavage, this cysteine residue is identified in a disulphide bond in several crystal structures suggesting that a more likely explanation for the lower entry phenotype in pseudovirus bearing this mutation is down to defective protein folding or stability[37], Supplementary Fig. 9a. The wild-type, S13A and V635G mutants all show S0 (uncleaved) and S1 (cleaved) spike Fig. 3e. Notably the V635G mutant showed significantly increased levels of the cleaved S1, with a near 3:1 ratio of S1 to S0 compared to the wild-type where this is 1:1 Fig. 3f (Unpaired Welch's *t*-test, $p < 0.05$). This could reflect either enhanced cleavage at this location, or increased incorporation of the cleaved V635G S1 into pseudovirus particles. Of note, the increased incorporation of cleaved spike in V635G mutant particles was consistent with the increased infectivity of this mutant Fig. 3c, Supplementary Fig. 9b.

SARS-CoV-2 infection induces proteolytic cleavage of multiple host proteins. Examination of cellular neo-N-termini from infected cells, has the potential to indicate activation of host cell proteases, or the targeting of specific host pathways. Motif analysis from both cell lines revealed strong enrichment of serine residues on the P' region of cleavage sites corresponding to neo-N-termini Fig. 4a, b. Diminished abundance of R/K was also observed, although it should be noted that we excluded neo-N-termini from our analysis which were preceded by R/K and could potentially be artifacts generated by tryptic digestion during sample processing.

Given this strong enrichment for specific residues in our datasets, we sought to identify if enriched or diminished neo-N-termini could be associated with a causal protease using TopFIND 4.0, which links neo-N-termini to causal proteases[38]. As this software does not currently support African Green Monkey datasets, this analysis could be performed only for the Ace2-A549 dataset, the results of which are included in Supplementary Fig. 10. Of those significantly regulated neo-N-termini, causal proteases were identified for a subset, although the analysis did not find the overall activities of these causal proteases to be significantly regulated in infected cells at 24 h post-infection compared to mock-infected cells ($p < 0.05$, Adjusted Fisher's exact test) Supplementary data 8, although this may reflect low numbers of annotated substrates for individual proteases.



Finally, we performed gene ontology analysis of proteins from which the significantly regulated neo-N-termini were derived, identifying significantly regulated biological functions, cellular compartments and molecular functions within each cell line dataset Supplementary Fig. 11, Supplementary Fig. 12. While many of the enriched GO terms were unique to each cell line, common themes included mitochondria, and the cytoskeleton/cell adhesion.

We then sought to identify neo-N-termini that could be targeted directly by the SARS-CoV-2 proteases, rather than host cell proteases. The consensus sequences for coronavirus proteases are conserved between coronaviruses, with PLP recognising a P4

Fig. 4 Increased abundance of novel cellular neo-N-termini consistent with SARS-CoV-2 protease consensus sequences suggests viral protease activity on cellular substrates. Panels **a** and **b** show motif analysis highlighting enriched amino acids proximal to the N-terminus of neo-N-termini enriched in SARS-CoV-2-infected A549-Ace2 and Vero E6 cells, respectively. DAU differential amino acid usage. The P5 to P5′ positions for the cleavage sites are shown following the nomenclature of Schechter and Berger. **c** Consensus motifs for Mpro and PLP. Panels **d** and **e** show the relative abundance of cellular neo-N-termini identified as significantly upregulated ($n = 3$ biological replicates, unpaired two sample *t*-test, multiple-hypothesis corrected $q < 0.05$) and matching or resembling the Mpro or PLP consensus motifs from A549-Ace2 or Vero E6 cells, respectively. Sequence match to the consensus is indicated by the pink or green coloring of the P4 to P1 positions of the relevant cleavage sites indicating match to the Mpro or PLP P4, P2, or P1 positions, respectively. **f** In vitro validation of GFP-tagged PNN, PAICS, and SRC cleavage by SARS-CoV-2 Mpro and PLP, following incubation with 10 M of the respective protease ($n = 3$). **g** Cell-based validation of Mpro cleavage of GOLGA3 and PAICS following transfection of SARS-CoV-2 Nsp4-5 plasmid. Tubulin is included as a loading control ($n = 3$).

to P1 LxGG motif, and Mpro recognising a (A—P—S—T—V) xLQ motif[39] Fig. 4c). No strong preference has been identified for either protease at the P3 residue (Fig. 4a). Analysis of both datasets showed strong enrichment for neo-N-termini consistent with cleavage at Mpro motifs (two-tailed Kolmogorov–Smirnov test, $p < 0.001$, Supplementary Fig. 13a, b). However, no comparable enrichment could be seen for neo-N-termini consistent with cleavage at PLP motifs (Supplementary Fig. 13c, d). This may reflect fewer cellular protein substrates of PLP compared to Mpro, or higher background levels of neo-N-termini generated by cellular proteases with similar P4 to P1 cleavage specificities as PLP.

Neo-N-termini matching, or close to the consensus sequences, for either Mpro or PLP and showing significant upregulation (*t*-test, $q \leq 0.05$ after correction for multiple-hypothesis testing) at 24 h post-infection compared to the 24 h mock sample were selected for further analysis. Perfect matches to the consensus sequences from A549-Ace2 cells included NUP107, PAICS, PNN, SRC, and XRCC1. GOLGA3 and MYLK (MCLK) were identified from Vero E6 cells. Hits from both cell lines that resembled, but did not completely match the consensus sequence were ATAD2, ATP5F1B, BST1, KAT7, KLHDC10, NUCKS1, and WNK1 (Fig. 4d, e). Adding confidence to these observations, approximately half of these hits were also identified in a recent SARS-CoV-2 proximity labelling study (ATP5F1B, GOLGA3, NUP107, PNN, SRC, and WNK)[7], and GOLGA3 was additionally identified in an interactome study as an nsp13 interaction partner[6].

SRC, MYLK, and WNK are all protein kinases, one of the protein families best studied as drug targets[40]. MYLK is especially interesting as dysregulation of MYLK has been linked to acute respiratory distress syndrome—one of the symptoms of severe COVID-19 disease[41]. NUP107 is a member of the nuclear pore complex, with nucleo-cytoplasmic transport a frequent target for viral disregulation[42]. GOLGA3 is thought to play a role in localisation of the Golgi and Golgi-nuclear interactions, and was identified in two recent studies of SARS-CoV-2 interactions[6,7]. PNN is a transcriptional activator, forming part of the exon junction complex, with roles in splicing and nonsense-mediated decay. The coronavirus mouse hepatitis virus has previously been shown to target nonsense-mediated decay, with pro-viral effects of inhibition[43]. PAICS and BST1 both encode enzymes with roles in ADP ribose and purine metabolism-, respectively, with PAICS previously identified as binding the influenza virus nucleoprotein[44].

The majority of these neo-N-termini showed enrichment at 24 h, with levels remaining largely unchanged at earlier timepoints, especially for Mpro substrates (Fig. 4d, e). This matches the timing for peak viral RNA, protein expression, and titres over the timepoints examined (Fig. 1b–d). Exceptions to this trend include the potential PLP substrates, 2/3 of which begin to show increased abundance at 12 h post-infection, with BST1 appearing to peak at 12 h rather than 24 h, indicating a potential temporal regulation of the two viral proteases. Data for all quantified and filtered N- and neo-N-termini from A549-Ace2 and Vero E6 cells is available in Supplementary data 4 and 5, respectively. Paired analysis of enriched neo-N-termini together with our unenriched dataset, allowed inference of cleavage stoichiometry for a subset of these significantly enriched cleaved cellular neo-N-termini by the HIquant approach[45] Supplementary Fig. 14, suggesting that by 24 h in SARS-CoV-2-infected cells, the majority of these proteins are present in the cleaved form.

We sought to validate a subset of prospective SARS-CoV-2 protease substrates in vitro, using the L. tarentolae system previously used to identify cleavage of proteins involved in the immune response by the SARS-CoV-2 proteases[14]. For these assays, the target protein is fused (N- or C-terminally) to GFP, which is then imaged directly in the SDS-PAGE gel. This system sucessfully validated cleavage of PNN, PAICS and SRC by Mpro (PNN, PAICS) and PLP (SRC), respectively (Fig. 4f). It also indicated additional cleavage products of PNN, and SRC not identified in the original mass spectrometry study. SRC cleavage was identified by N-terminomics following a LfGG motif yielding a neo-N-terminus at F67 (Fig. 4f). Two SRC cleavage products migrate at slightly over 30 kDa (including the GFP tag). The second cleavage product found in vitro migrates at slightly higher molecular weight, consistent with cleavage at an LaGG motif 19 amino acids downstream of the first cleavage site, generating a neo-Nterminus at V86 (Fig. 4f). Titration of the amount of PLP included in the reaction resulted in dose-dependent cleavage of SRC (Supplementary Fig. 15).

In addition to a cleavage product consistent with the cleavage at S114 observed in the N-terminomics migrating between the 80–115 kDa markers, a second cleavage product of PNN was also identified, migrating at slightly over 50 kDa including the GFP tag (Fig. 4f). However there are multiple candidate cleavage sites located in this portion of PNN that could explain this cleavage event.

We also validated cleavage of PAICS and GOLGA3 by SARS-CoV-2 in a cell-based assay. To generate functional Mpro, a plasmid containing the nsp4-5 sequence was generated, permitting autocleavage at the nsp4-5 junction and generation of an authentic N-terminus for nsp5 (Mpro). Transfection of HEK 293 cells with this construct resulted in cleavage of PAICS and GOLGA3. Both antibodies recognise the C-terminus of the cleaved proteins. As with the in vitro assay, cleavage of PAICS resulted in the appearence of a single-cleavage product. This assay recognises endogenous rather than tagged PAICS, which is why both uncleaved and cleaved PAICS have differing apparent molecular weights in Fig. 4f and g. Anti-GOLGA3 cleavage results in the appearence of a single-cleavage product at slightly over 70 kDa. This may reflect further proteolytic cleavage of this protein given that the observed cleavage at 365/366 would be expacted to result in a 40k reduction in the apparent molecular weight of GOLGA3, which migrates at slightly over its predicted molecular weight of 167 kDa.

Prospective MPro and PLP substrates are necessary for efficient viral replication, and represent targets for pharmacological intervention. To investigate if the putative cellular substrates of MPro and PLP identified in the N-terminomic analyses are necessary for efficient viral replication, an siRNA screen was conducted Fig. 5. Where proteolytic cleavage inactivates cellular proteins or pathways inhibitory for SARS-CoV-2 replication, siRNA depletion would be anticipated to result in inreased viral titres and/or RNA levels. If proteolysis results in altered function that is beneficial for the virus, we would expect siRNA depletion to result in a reduction in viral titres/RNA levels.

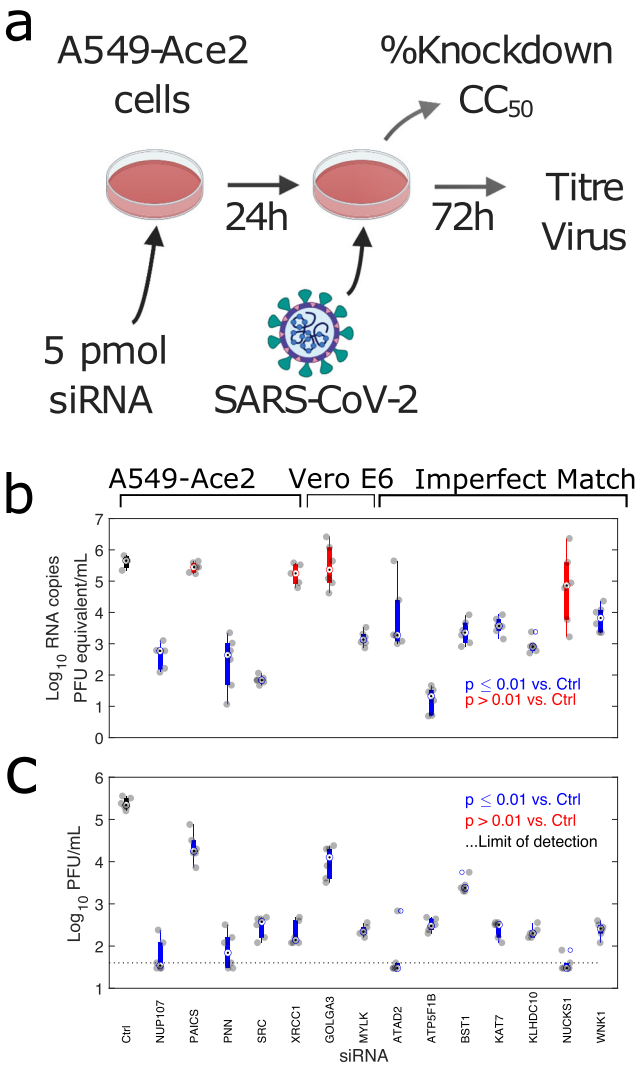


Fig. 5 siRNA depletion of potential MPro and PLP substrates results in significant reductions to viral RNA copies and titres in A549-Ace2 cells. **a** Experimental design, **b** Viral RNA copies ($n = 3$ (control), 5 (NUP107, XRCC1), and 6 (all other samples) biologically independent samples), and **c** titres in supernatant at 72 h post-infection ($n = 6$ biologically independent samples), following infection 24 h post-transfection with the indicated siRNA). Boxplot minima/maxima represent the furthest non-outlier datapoints, centre the median, and bounds of box the interquartile range. Outliers are defined as datapoints >1.5 times the interquartile range from the bottom or top of the box. Individual datapoints are shown in grey. Significance was calculated by one-way ANOVA. Blue bars indicate samples with significantly reduced viral RNA copies or titres ($p \leq 0.01$). Those not meeting this threshold are shown in red. The control siRNA-treated sample is indicated with a black bar. The limit of detection in the plaque assay was calculated to be 40 PFU/mL (dotted line).

Proteins with neo-N-termini showing statistically significant increased abundance during SARS-CoV-2 infection and either matching, or similar to the viral protease consensus sequences were selected for siRNA depletion.

Infection of A549-Ace2 cells was performed 24 h post-transfection with the indicated siRNA and allowed to proceed for 72 h (Fig. 5a). Cell viability for all targets was comparable to untreated controls (Supplementary Fig. 16). siRNA knockdown efficiency at the time of infection was confirmed by qRT-PCR (Supplementary Fig. 17), with a low of 77% efficiency for NUCKS1, and averaging over 95% efficiency for most targets. 10/14 coronavirus protease substrates showed significant reductions (one-way ANOVA, $p \leq 0.01$) in viral RNA levels, averaging a 100–1000-fold median decrease in viral RNA equating to pfu equivalents per ml at 72 h post-infection compared to treatment with a control siRNA (Fig. 5b). PAICS, GOLGA3, NUCKS1, and XRCC1 did not show a significant drop in RNA copy number following siRNA treatment. Plaque assays were then conducted on these samples to determine whether this observed reduction in viral RNA levels reflected a reduction in infectious virus titres (Fig. 5c). All 14 potential substrates showed a statistically significant (one-way ANOVA, $p \leq 0.01$) reduction in viral titres following siRNA depletion. For PAICS and GOLGA3, which did not show reduced RNA levels, these reductions were ~10 fold. Most other siRNA targets showed reduced titres in the 100–1000-fold range. These differences in outcome between viral RNA levels and plaque assays may result from a subset of proteins required for efficient viral replication. While efficient mRNA knockdown was shown for all targets (Supplementary Fig. 17), it is also possible that this discrepancy between viral RNA levels and titres may result from differences in protein half-life of the knockdown targets. This could result in proteins with longer half lives only giving a phenotype at later stages of infection when infectious virus is produced. Finally, we extended our analysis to two previously identified SARS-CoV-2 substrates[14], IRF3 and TAB1 (Supplementary Fig. 18) not identified in our N-terminomic analysis. We found that in line with previous reports[46], the antiviral IRF3 resulted in an aproximately three-fold increase in viral titres and increased RNA copies. In contrast, TAB1 depletion resulted in reduced SARS-CoV-2 RNA copies and viral titres, possibly reflecting the role of TAB1 in negative regulation of antiviral responses[47]. In line with the apparently pro-viral nature of the majority of the prospective SARS-CoV-2 substrates, we observed in a microscopy-based viability assay that siRNA-treated infected cells appeared to show enhanced viability compared to infected cells treated with a scrambled siRNA control, though only reached statistically significance (One-way ANOVA, Tukey's correction for multiple-hypothesis testing) for some substrates (PNN, XRCC1, GOLGA3, ATP5F1B).

A subset of the prospective viral protease substrates have commercially available inhibitors, notably SRC and MYLK. In the case of SRC these include tyrosine kinase inhibitors in current clinical use. In light of the siRNA screening results we concluded that pharmacological inhibition of SARS-CoV-2 protease substrates could represent a viable means to inhibit SARS-CoV-2 infection. Dose-response experiments were conducted with seven inhibitors to determine whether pharmacological inhibition of SARS-CoV-2 protease substrates could be employed as a potential therapeutic strategy (Fig. 6; Supplementary Fig. 19). Of these, two tyrosine kinase inhibitors: Bafetinib and Sorafenib showed inhibition at concentrations which did not result in cytotoxicity in the human cell line A549-Ace2 (Fig. 6). In the case of Bafetinib, a Lyn/Bcr-Abl inhibitor, which has off-target activity against SRC the IC50 was in the nanomolar range (IC50: 0.79 µM, 95% confidence interval 0.23–1.35 µM). Bafetinib has recently been independently identified as an inhibitor of the coronaviruses

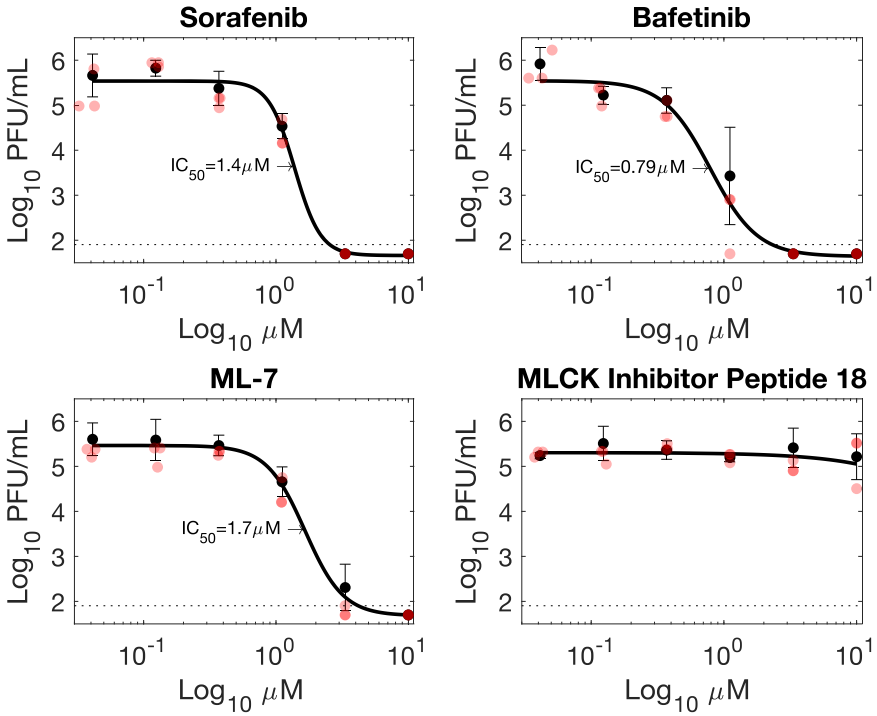


Fig. 6 Inhibitors targeting viral protease substrates reduce SARS-CoV-2 titres in A549-Ace2 cells. Sorafenib is a tyrosine kinase inhibitor previously shown to inhibit SARS-CoV-2 replication[4]. Bafetinib is a dual ABL/LYN inhibitor with off-target activity against SRC kinase. ML-7 and MLCK target myosin light chain kinase (MYLK/MLCK). Black circles and error bars represent mean and standard deviation. $n = 3$ biologically independent samples. Red circles indicate individual datapoints. The limit of detection in the plaque assay was calculated to be 40 PFU/mL (dotted line). PFU plaque-forming units.

OC43 and SARS-COV-2 in a large-scale drug repurposing screen[48]. Inhibition with Sorafenib, which was included as a positive control and does not directly target any of the protease substrates, was in the low micromolar range (Fig. 6), in line with a previously published report[4]. Two inhibitors were trialled against MYLK. These were MLCK inhibitor peptide 18, and ML-7. Only ML-7 showed inhibition of SARS-CoV-2, with inhibition in the low micromolar range (IC50: 1.7 μM, 95% confidence interval 1.51–1.80 μM), at concentrations which did not induce cytotoxicity (Fig. 6; Supplementary Fig. 20). ML-7 and MLCK inhibitor peptide 18 have different mechanisms of action, with MLCK inhibitor peptide 18 outcompeting kinase substrate peptides, and ML-7 inhibiting ATPase activity. All four had CC50 values over the 10 μM maximum concentration tested, except ML-7, which had a CC50 of 5 μM. Bafetinib did show reduced viability at the two highest concentrations tested (10 μM, 3.3 μM), although not reaching 50% reduction (Supplementary Fig. 20).

The other 3 tyrosine kinase inhibitors tested (Bosutinib, Saracatinib, Dasatinib) all showed inhibition Supplementary Fig. 19; however, cytotoxicity results obtained with the assay used were also high preventing the unambiguous determination of whether inhibition was specific or due to cytotoxicity Supplementary Fig. 20. However, it should be noted that these agents have been reported to be cytostatic in A549 cells, and the CellTiter-Glo assay used to assess viability measures cellular metabolism so will not distinguish between cytostatic and cytotoxic effects.

Discussion

We employed a mass spectrometry approach to study proteolytic cleavage events during SARS-CoV-2 infection. Substrates of viral proteases are frequently inferred through studies of related proteases[49]. However, such approaches are unable to identify novel substrates, and even closely related proteases can differ in their substrate specificity[50]. Mass spectrometry-based approaches to identify protease substrates by identifying the neo-N-terminal peptides generated by protease activity have existed for a number of years[22,51–53]; however, they have seen only limited application to the study of viral substrates[54,55], and have not been previously applied to the study of proteolysis during coronavirus infection.

While our approach identified multiple novel viral and cellular cleavage sites, it also failed to identify multiple known cleavage sites, including the FURIN cleavage site in spike, and multiple cleavage sites within the viral polyprotein. This can be understood from the dependence of the approach on the specific protease used for mass spectrometry analysis. Isobaric labelling prior to trypsin digestion blocks tryptic cleavage at lysine residues and causes trypsin to cleave solely after arginine residues. This results in the generation of long peptides and if the specific cleavage site does not produce a peptide of suitable length for analysis (typically 8–30 amino acids) then it will be missed. This can be alleviated through the application of multiple mass spectrometry-compatible proteases in parallel, yielding multiple peptides of different length for each cleavage site[56,57]. This would both increase the number of sites identified and cross-validate previously identified cleavage sites. These methods will likely prove a fruitful avenue for future investigations of proteolysis during infection with SARS-CoV-2 and other viruses that employ protease-driven mechanisms of viral replication.

Our approach identified multiple cleavage sites within viral proteins. In some cases, such as the nucleocapsid protein, cleavage by cellular proteases has been observed for SARS-CoV[18,19], although the number of cleavage products observed was much higher in our study (Fig. 2). Compared to the gel-based approaches used in the past, our approach is much more sensitive for detecting when protease activity results in N-termini with ragged ends, due to further exoproteolytic activity. Examples of this in our data are particularly evident in Fig. 2 for the nucleocapsid and ORF3a where neo-N-termini appear in clusters.

Cleavage sites within the nucleocapsid and spike protein are of particular interest as these are the two viral antigens to which research is closely focuses for both testing and vaccination purposes. In this context, neo-N-termini are of interest as N-termini can be recognised by the immune response, as they are typically surface-exposed. Antibodies recognising neo-N-termini such sites will not be detected in tests using complete or recombinant fragments that do not account for such cleavage sites. Indeed, a recent study revealed altered antigenicity of proteolytic proteoforms of the SARS-CoV-2 nucleocapsid following autolysis[25], and loss of antibody responses to spike following mutations in circulating SARS-CoV-2 variants altering signal peptide cleavage[33]. Understanding cleavage events can also inform interpretation of protein structural analysis, for example in the ORF3a viroporin[26]. Knowledge of cleavage sites can permit further analysis of spike entry mechanisms, and vaccine design, and inform genotype-to-phenotype predictions.

Formation of the most prominent neo-N-terminus we identified in SARS-CoV-2 spike at 637 appeared dependent on cathepsins and/or calpains, as its appearance was limited by calpeptin treatment. Mutation of the P2 residue in the putative cleavage site, resulting in mutant V365G, led to an increased incorporation of cleaved spike in pseudotyped viral particles. Consistent with an increased content of fusion-competent cleaved spike, the V635G mutant showed an increased infectivity in HEK-Ace2 target cells.

Why blocking the formation of the 637 neo-N-terminus promotes cleaved spike incorporation remains to be elucidated. A possible explanation may lie in a competition between different cleavage sites in the producer cell, with cleavage at 637 inhibiting cleavage at the FURIN site, or inhibiting the incorporation of spike trimers already cleaved at the FURIN site. The capacity of producer cells to cleave viral glycoproteins at alternative sites may thus be viewed as an intrinsic defense mechanism. In contrast, the capacity of viral proteases to cleave multiple host proteins, as demonstrated in this study, contributes to well-established mechanisms aiming at inhibiting innate host responses, including in particular the interferon pathway[14]. Therefore, the diversity of proteolytic cleavage events revealed by N-terminomics may reflect another layer in the dynamic evolutionary conflict between viruses and their hosts.

Proteolytic cleavage can alter protein function in several ways, including inactivation, re-localisation, or altered function including the removal of inhibitory domains. Our siRNA screen showed knockdown of the majority of potential protease targets we identified was inhibitory to SARS-CoV-2 replication (Fig. 5). Indeed, with the exception of the previously reported antiviral protein IRF3, no siRNA treatment resulted in higher viral titres or RNA levels, suggesting that inactivation is not the prime purpose of these cleavage events. This suggests that in many cases, proteolytic cleavage by viral proteases may be extremely targeted, serving to fine-tune protein activity, rather than merely serving as a blunt instrument to shut down unfavorable host responses. It is also worth noting the low overlap between our infection-based study, and a subsequently released N-terminomics dataset, which used incubation of cell culture lysates with recombinant Mpro[58]. Only GOLGA3 was common to both studies, in spite of the larger number of cleavage events identified in the Koudelka et al. study. However in such a lysate-based experiment subcellular compartmentalisation and regulation of relative enzyme and substrate localisation & abundance is lost, so can risk identifying cleavage events not possible in vivo during genuine infection.

In this study, we used two cell line models to characterise the effects of SARS-CoV-2 infection on protease activity and the generation of viral and cellular cleavage products. Notably, we tested the efficiency of several inhibitors against SARS-CoV-2 infection only in the context of the A549-Ace2 cell line model. Other model systems such as Calu-3 or lung organoids may yield different results for a subset of cleavage events due to differences in SARS-CoV-2 entry pathways[59]. Given previous experiences translating cell culture findings for SARS-CoV-2 to the clinic, it is important to note our results present preliminary data that must be further validated in other models, in vivo, and through clinical trials before use in patients for the treatment of COVID-19 disease.

In conclusion, we have presented a first study of proteolysis and the resulting proteolytic proteoforms generated in the context of SARS-CoV-2 infection for two mammalian cell lines: A549-Ace2 and Vero E6. We identified multiple previously unknown cleavage sites in viral proteins including close to characteristic mutations in circulating variants of concern. We also identified 14 novel substrates of the coronavirus Mpro and PLP proteases, finding that these substrates appear pro-viral in the context of siRNA and inhibitor assays in A549-Ace2 cells. An improved understanding of the exact ways in which proteolytic cleavage is regulated, modulates protein activity, and serves to benefit viral replication will be crucial for targeting cellular substrates of viral proteases as a therapeutic strategy.

Methods

Cell culture and virus. Vero E6 (Vero 76, clone E6, Vero E6, ATCC® CRL-1586TM) authenticated by ATCC and tested negative for mycoplasma contamination prior to commencement were maintained in Dulbecco's modified Eagle's medium (DMEM; Thermo Fisher Scientific) containing 10% (v/v) fetal bovine serum (FBS, ThermoFisher Scientific) and penicillin/streptavidin (ThermoFisher Scientific). A549-Ace2 cells, a human lung epithelial cell line that over-expresses ACE2, were kindly provided by Oliver Schwartz (Institut Pasteur)[60]. A549-Ace2 cells were cultured in DMEM supplemented with 10% FBS, penicillin/streptavidin and 10 µg/ml blasticidin (Sigma) and maintained at 37 °C with 5% CO_2. The SARS-CoV-2 isolate BetaCoV/France/IDF0372/2020 was supplied through the European Virus Archive goes Global (EVAg) platform. Viral stocks were prepared by propagation in Vero E6 cells in DMEM supplemented with 2% FBS. For protease inhibitor experiments employing calpeptin and camostat mesylate, these drugs or an equal volume of vehicle (DMSO) were supplemented to the medium 12 h post-infection at 50 mM final concentration. All experiments involving live SARS-CoV-2 were performed in compliance with Institut Pasteur Paris's guidelines for Biosafety Level 3 (BSL-3) containment procedures in approved laboratories. All experiments were performed in at least three biologically independent samples.

For spike-pseudotyped lentivector production and infections, HEK293Tn were maintained in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum and 100 g/mL penicillin/streptomycin (complete medium), and cultured at 37 °C under 5% CO2. HEK293T-hACE2-TMPRSS2 (called HEK-ACE2-TMPRSS2) with inducible TPMRSS2 expression were a gift from Julian Buchrieser and Olivier Schwartz[60]. These cells were maintained in complete medium with blasticidin (10 µg/mL, InvivoGen) and puromycin (1 µg/mL, Alfa Aesar), and TMPRSS2 expression was induced by addition of doxycycline (0.5 µg/mL, Sigma).

SARS-CoV-2 titration by plaque assay. Vero E6 cells were seeded in 24-well plates at a concentration of 7.5 × 104 cells/well. The following day, serial dilutions were performed in serum-free MEM media. After 1 h absorption at 37 °C, 2× overlay media was added to the inoculum to give a final concentration of 2% (v/v) FBS/MEM media and 0.4% (w/v) SeaPrep Agarose (Lonza) to achieve a semi-solid overlay. Plaque assays were incubated at 37 °C for 3 days. Samples were fixed using 4% Formalin (Sigma–Aldrich) and plaques were visualised using crystal Violet solution (Sigma–Aldrich).

Infections for N-terminomic/proteomic analysis. N-terminomic sample preparation is based around Weng et al. 2019 Mol. Cell. Proteomics, adapted for TMTpro-based quantitation[22,61]. A protocol for this TMTpro-adapted method can be found at https://www.protocols.io/view/tmtprohunter-n-terminomics-bi44kgyw. Vero E6 or A549-Ace2 cells were seeded using 2 × 106 cells in T25 flasks. The following day cells were either mock infected or infected with SARS-CoV-2 at a MOI of 1 in serum-free DMEM at 37 °C for 1 h. After absorption, the 0 h samples were lysed immediately, while the media for other samples was replaced with 2% FBS/DMEM (ThermoFisher Scientific) and incubated at 37 °C for times indicated before lysis. Cells were washed 3× with PBS (ThermoFisher Scientific) before lysing them in 100 mM HEPES pH 7.4 (ThermoFisher Scientific), 1% Igepal (Sigma–Aldrich), 1% sodium dodecyl sulfate (SDS; ThermoFisher Scientific), and protease inhibitor (mini-cOmplete, Roche). Samples were then heated to 95 °C for 5 min, before immediately freezing at −80 °C. Samples were then thawed and

incubated with benzonase for 30 min at 37 °C. Sample concentrations were normalised by BCA assay, and 25 µg of material from each sample was used for downstream processing.

DTT was added to 10 mM and incubated at 37 °C for 30 min, before alkylation with 50 mM 2-chloroacetamide at room temperature in the dark for 30 min. DTT at 50 mM final concentration was added to quench the 2-chloroacetamide for 20 min at room temperature. Samples were washed by SP3-based precipitation[62]. Each sample was resuspended in 22.5 µL 6 M GuCl, 30 µL of 0.5 M HEPES pH8, and 4.5 µL TCEP (10 mM final) and incubated for 30 min at room temperature.

0.5 mg of individual TMTpro aliquots (Lot VB294905) were resuspended in 62uL of anhydrous DMSO. 57 µL of the TMTpro was then added to each sample, mixed and incubated for 1.5 h. Label allocation was randomised using the Matlab Randperm function. Excess TMTpro was quenched with the addition of 13 µL of 1 M ethanolamide and incubated for 45 min. All samples were combined for downstream processing. SP3 cleanup was performed on the combined samples. These were resuspended in 400 µL of 200 mM HEPES pH8, containing Trypsin gold at a concentration of 25 ng/µL and incubated overnight at 37 °C.

Samples were placed on a magnetic rack for 5 min. 10% of the samples was retained for the unenriched analysis. The remaining material was supplemented with 100% ethanol to a final concentration of 40%, undecanal added at an undecanal:peptide ratio of 20:1 and sodium cyanoborohydride to 30 mM. pH was confirmed to be between pH 7-8 and the samples were incubated at 37 °C for 1 h. Samples were then sonicated for 15 s, and bound to a magnetic stand for 1 min. The supernatant was retained and then acidified with 5% TFA in 40% ethanol. Macrospin columns (Nest group) were equilibrated in 0.1% TFA in 40% ethanol. The acidified sample was applied to the column, and the flow through retained as the N-terminal-enriched sample.

Both unenriched and enriched samples were desalted on macrospin columns (Nest group), before drying down again. Off-line basic reverse phase fractionation for both unenriched and enriched samples was performed on a Waters nanoAcquity with an Acquity UPLC M-Class CSH C18 130 A 1.7 µm, 300 × 150 µm column. The sample was run on a 70 min gradient at 6 µL/min flow rate. Gradient parameters were 10 min 3% B, 10–40 min 3–34% B, 40–45 min 34–45% B, 45–50 min 45–99% B, 50–60 min 99% B, 60.1–70 min 3% B. Buffers A and B were 10 mM ammonium formate pH 10, and 10 mM ammonium formate pH 10 in 90% acetonitrile, respectively. Both samples were resuspended in buffer A, and 1 min fractions were collected for 1–65 min of the run. These were concatenated into 12 (1:13:24…) or 5 fractions (1:6:11…) for unenriched and enriched samples, respectively, using a SunChrom Micro Fraction Collector. Samples were dried down and resuspended in 1% formic acid for LC-MS/MS analysis.

Mass spectrometry. LC-MS/MS analysis was conducted on a Dionex 3000 coupled in line to a Q-Exactive-HF mass spectrometer. Digests were loaded onto a trap column (Acclaim PepMap 100, 2 cm × 75 µm inner diameter, C18, 3 microM, 100 ˚A) at 5 µL per min in 0.1%(v/v) TFA and 2%(v/v) acetonitrile. After 3 min, the trap column was set in line with an analytical column (Easy-Spray PepMap® RSLC 15 × 50 cm inner diameter, C18, 2 microlM, 100 ˚A) (Dionex). Peptides were loaded in 0.1%(v/v) formic acid and eluted with a linear gradient of 3.8–50% buffer B (HPLC grade acetonitrile 80%(v/v) with 0.1%(v/v) formic acid) over 95 min at 300 nl per min, followed by a washing step (5 min at 99% solvent B) and an equilibration step (25 min at 3.8% solvent). All peptide separations were carried out using an Ultimate 3000 nano system (Dionex/Thermo Fisher Scientific). The Q-Exactive-HF was operated in data-dependent mode with survey scans aquired at a resolution of 60,000 at 200 *m/z* over a scan range of 350–2000 *m/z*. The top 16 most abundant ions with charge states +2 to +5 from the survey scan were selected for MS2 analysis at 60,000 *m/z* resolution with an isolation window of 0.7 *m/z*, with a (N)CE of 30%. The maximum injection times were 100 and 90 ms for MS1 and MS2, respectively, and AGC targets were 3e6 and 1e5, respectively. Dynamic exclusion (20 s) was enabled.

Data analysis. All data were analysed using Maxquant version 1.6.7.0[63]. Custom modifications were generated to permit analysis of TMTpro 16plex-labelled samples. FASTA files corresponding to the reviewed Human proteome (20,350 entries, downloaded 8 May 2020), and African Green monkey proteome (Chlorocebus sabeus, 19,223 entries, downloaded 16 May 2020). A custom fasta file for SARS-CoV-2 was generated from the Uniprot-reviewed SARS-CoV-2 protein sequences (2697049). This file was modified to additionally include the processed products of pp1a and pp1b, novel coding products identified by ribo-seq[24], as well as incorporate two coding changes identified during sequencing (spike: V367F, ORF3a: G251V). All FASTA files, TMT randomisation strategy, and the modifications.xml file containing TMTpro modifications have been included with the mass spectrometry data depositions. Annotated spectra covering peptide N-termini of interest were prepared using xiSPEC v2[64].

Several different sets of search parameters were used for analysis of different experiments.

For analysis of unenriched material from fractionated lysates:

Default MaxQuant settings were used with the following alterations. As the experimental design meant unenriched samples contained a majority of peptides lacking N-terminal TMT labelling, quantification was performed at MS2-level with the correction factors from Lot VB294905 on lysine labelling only with the N-Terminal label left unused. Digestion was trypsin/p with a maximum of three missed cleavages. Carbamidomethylation of cysteines was selected as a fixed modification. Oxidation (M), Acetylation (Protein N-terminus), and N-terminal TMTpro labelling were selected as variable modifications. PSM and Protein FDR were set at 0.01.

For analysis of fractionated, N-terminally enriched material:

Default MaxQuant settings were used with the following alterations. Quantification was performed at MS2-level with the correction factors from Lot VB294905. Digestion was semi-specific ArgC, as TMTpro labelling of lysines blocks trypsincleavage. Carbamidomethylation of cysteines was selected as a fixed modification. Oxidation (M), Acetylation (Protein N-terminus), Gln/Glu to pyroglutamate were selected as variable modifications. PSM and Protein FDR were set at 0.01.

For analysis of viral protein neo-N-termini from fractionated, N-terminally-enriched material:

Default MaxQuant settings were used with the following alterations. MS1based quantitation was selected. Digestion was ArgC, sei-specific N-terminus. Carbamidomethylation of cysteines was selected as a fixed modification. Oxidation (M), Acetylation (Protein N-terminus), Gln/Glu to pyroglutamate, and TMTpro modification of N-termini and lysine residues were selected as variable modifications. PSM and Protein FDR were set at 0.01.

All downstream analysis was conducted in Matlab, external packages used include BreakYAxis[65]. Reverse hits and contaminants were removed, peptides were filtered to meet PEP ≤ 0.02. For quantitative analysis, peptides were further filtered at PIF ≥ 0.7. TMTpro data were normalised for differences in protein loading by dividing by the label median, rows were filtered to remove rows with more than 2/3 missing data. Missing data were KNN imputed, and individual peptides were normalised by dividing by their mean abundance accross all TMTpro channels. As the objective was to identify protein cleavage events, peptides were further filtered to remove those beginning at the first or second amino acid in a protein sequence that represent the native N-terminus. ± methionine. neo-N-termini were annotated if they matched known signal peptides. For non-quantitative analysis (e.g. mapping of viral neo-N-termini), peptides were filtered to retain only blocked (acetylated, TMTpro labelled, and pyroglutamate) N-termini. pyroglutamate-blocked N-termini were discarded if they were preceeded by arginine or lysine as these could represent artifactual cyclisation of tryptic N-termini. Fractional protein or peptide intensity was calculated as the total intensity for the protein or peptide, multiplied by the fraction of the summed normalised TMTpro intensity represented by a particular TMTpro label of interest.

Inference of cleaved and uncleaved proteoforms of protease substrates was performed using the HIquant approach[45], using the unenriched and N-terminally enriched data, processed as above. To ensure that unenriched tryptic peptides represented a simple, rather than complex mixture of proteoforms, inference was performed for cellular substrates where the unenriched tryptic peptides formed a single cluster based on the Euclidean distance using the Matlab evalcluster ('linkage') implementation.

Visualisation of the Y636/S637 cleavage site within the SARS-CoV-2 spike glycoprotein structure was performed using PDB: 6X6P[27], in UCSF ChimeraX v1.0[66].

Data analysis for motif analysis, association of causal proteases, and GO enrichment. Data were normalised for loading by dividing each TMT intensity column by the column median. The rows with 75% or more missing values were removed. The remaining missing values were imputed using K-nearest neighbour algorithm. Each row was then normalised by the row mean. Peptides that required imputation of more than one value per group were excluded. Neo-N-peptides were selected by the following criteria: start position at amino acid ¿ 2, not among known signal peptides, not among predicted signal peptides, not preceded by arginine or lysine. Prediction of potential signal peptides was performed using SignalP v5.0[67] tool—only peptides with prediction confidence greater than 0.9 and predicted cleavage site located 5 or less amino acids from a cleavage site of a detected peptide were considered.

All analysis described in this section was performed using the R statistical programming environment[68]. Differential abundance analysis was performed using limma[69] package. The statistical significance of the results was assessed using Storey's q-values[70] (q-value < 0.05). Geneset enrichment analysis was performed using clusterProfiler[71] package in R using BioSigDB[72,73] genesets. Significant pathways selected by *p* values, adjusted for multiple testing using Benjamini-Hochberg method (p-value < 0.05). Peptide motif analysis was done using the dagLogo[74] package. Peptide regions of five amino acids before and after the cleavage site were selected and motif analysis was performed using Fisher's exact test (p-value < 0.05). All figures were produced using package ggplot2[75]. Data analysis was performed by members of the University of Liverpool Computational Biology Facility.

Production of spike-pseudotyped lentivectors. Lentiviral particles encoding the SARS-CoV-2 spike were prepared by transient transfection of HEK293Tn cells using the CaCl2 method. The lentiviral vector pCDH-EF1a-GFP (System Bioscience), the packaging plasmid psPAXII (Addgene), the spike expression

vector phCMV-SARS-CoV-2-Spike (a gift from O. Schwartz), and the pRev plasmid (a gift from P. Charneau) were mixed at a 2:2:1:1 ratio and transfected at 252 µg DNA per 175 cm^2 cell flask. The pQCXIP-Empty plasmid was used as a negative control for spike expression. At 48 h after transfection, supernatants were collected and concentrated by ultracentrifugation at 23,000 × *g* for 1 h 30 m at 4 °C on a 20% sucrose cushion. Viral particles were resuspended in PBS and frozen in aliquots at −80 °C until use. Gag p24 antigen concentration was measured with the Alliance HIV-1 p24 Antigen ELISA kit (Perkin Elmer).

Point mutations to generate the mutants were introduced by site-directed mutagenesis of phCMV-SARS-CoV-2-Spike using Q5 polymerase (Thermo Scientific) and validated by sanger sequencing. The primers used are listed in the supplementary information.

Infection with spike-pseudotyped lentivectors. The day before infection, 100,000 HEK-ACE2 or HEK-ACE2-TMPRSS2 cells were plated in 96-well plates and TMPRSS2 was induced by the addition of doxycycline. HEK-ACE2 ± TMPRSS2 were infected with the equivalent of 2 µg of p24 Gag for each spike lentivector, in final volume of 100 µL. Infection was quantified by measuring the percentage of GFP + cells 2 days post-infection by flow cytometry. Cells were harvested, washed in PBS, and stained with the viability dye eF780 (eBioscience) for 30 min at 4 °C. After two washes in PBS, cells were fixed with paraformaldehyde 2% (ThermoFisher) and acquired on an Attune NxT flow cytometer. Results were analyzed with FlowJo software (v10.7.1), with statistical analyses carried out with the GraphPad Prism software (v9).

Western blotting of spike-pseudotyped lentivectors. To prepare protein extracts, cells were lysed in buffer with NaCl 150 mM, Tris HCl 50 mM (pH8), 1% Triton, EDTA 5 mM, supplemented with protease inhibitors (Roche) for 30 min on ice. For lentiviral particle extracts, an equivalent of 500 ng of p24 Gag was lysed in buffer with 1% Triton (ELISA kit, Alliance Perkin Elmer) for 30 min on ice. To preserve antibody reactivity, samples were not heated nor reduced before being run in a 4–12% acrylamide denaturing gel (NP0323, NuPAGE, ThermoFisher), and then transferred onto a nitrocellulose membrane (IB23001, ThermoFisher). The membrane was blocked with 5% dried milk in PBS Tween 0.1%, before incubation with the primary antibody for 1 h at RT, followed by three washes, and incubation with the secondary antibody for 30 min at RT. After three more washes, the fluorescent signal was revealed on a LiCor Odyssey 9120 imaging system. Images were quantified with the ImageStudioLite (v5.2.5) software, using a mode with automated background subtraction. Primary antibodies consisted in the human anti-spike mAb 48 (1:1,000; a gift from H. Mouquet) or the mouse anti-p24 Gag MAB7360 (R&D Systems; 1:1000). Anti-human or mouse IgG secondary antibodies, conjugated to DyLight-800 (A80-304D8, Bethyl Laboratories) or DyLight-680 (SA5-35521, ThermoFisher), respectively, were used at a 1:10,000 concentration.

In vitro cleavage assays. In vitro cleavage assays were performed using the Leishmania tarentolae (LTE) system as described[14]. SRC, PAICS, PNN, and RPA2 (control) were cloned as GFP fusion proteins into dedicated Gateway vectors for cell-free expression. Open Reading Frames (ORFs) in pDonor were sourced from the Human ORFeome collection, version 8.1 and transferred into Gateway destination vectors that include N-terminal (SRC) or C-terminal (PAICS, PNN and RPA2) Fluorescent proteins. The specific Gateway vectors were created by the laboratory of Pr. Alexandrov and sourced from Addgene (Addgene plasmid # 67137; http://n2t.net/addgene:67137; RRID:Addgene 67137). LTE extracts for in vitro expression were prepared in-house as described[76]. Purified recombinant Mpro and PLP were generated by the UNSW protein production facility as described previously[14].

The SRC, PAICS, PNN, and RPA2proteins were expressed individually in 10 µL reactions (1 µL DNA plasmid at concentrations ranging from 400 to 2000ng/L added to 9 µL of LTE reagent). The mixture was incubated for 30 min at 27 °C to allow the efficient conversion of DNA into RNA. The samples were then split into controls and protease-containing reactions. The proteases PLpro (nsp3) and 3CLpro (nsp5) were added at various concentrations, and the reactions were allowed to proceed for another 2.5 h at 27 °C before analysis. The controls and protease-treated LTE reactions were then mixed with LDS (Bolt LDS Sample Buffer, ThermoFisher) and loaded onto SDS-page gels (412% Bis-Tris Plus gels, ThermoFisher); the proteins were detected by scanning the gel for green (GFP) fluorescence using a ChemiDoc MP system (BioRad) and proteolytic cleavage was assessed from the changes in banding patterns. Note that in this protocol, the proteins are not treated at high temperature with the LDS and not fully denatured, to avoid destruction of the GFP fluorescence. As proteins would retain some folding, the apparent migration on the SDS-page gels may differ slightly from the expected migration calculated from their molecular weight. We have calibrated our SDS-page gels and ladders using a range of proteins, as shown previously[14].

Transfection and cell-based validation of proteolytic cleavage by western blotting. A mammalian expression plasmid expressing the coding sequence of SARS-Cov-2 Nsp4-Nsp5 in a pCDNA3.1 backbone was synthesised (GeneArt™ Gene synthesis, ThermoFisher, USA). HEK 293 T cells in a six-well plate were transfected with polyethylenimine (PEI) and 2 ug of Nsp4-Nsp5 fusion construct, or with pCDNA3.1 control, in three biological replicates. After 48 h, cells were lysed with RIPA buffer (ThermoFisher, USA) in presence of Phosphatase/Protease inhibitors cocktail (ThermoFisher, USA). The lysate was centrifuged (15 min at 4 °C and 17,005 × *g*) and the supernatant was collected. Protein concentration was quantified using Pierce™ BCA Protein Assay kit (ThermoFisher, USA) and western Blot performed with standard protocol. Briefly, proteins were seperated on a precast 4–20% gradient gel (Biorad, USA) and transferred on a nitrocellulose membrane using a semi-dry Trans-Blot Turbo Transfer System and Trans-Blot Turbo Transfer Buffer (Biorad, USA). Membranes were blocked for 1 h with 5% milk in TBST (Tris-Buffered Saline and Tween 20) buffer, rinsed, and incubated overnight at 4 °C with primary antibodies in 2% BSA in TBST. Membranes were washed with TBST and incubated for 2 h at room temperature with Horseradish peroxidase (HRP)-linked secondary secondary antibody (Cell signaling #7074). Chemiluminescent signal was revealed using SuperSignal™ West Pico PLUS Substrate (ThermoFisher, USA) and imaged with an Azure 600 Imaging system (Azure Biosystem, USA). The primary antibodies used were PAICS (Bethyl A304–547A-T), GOLGA3 (Bethyl A303–404A-T) and -Tubulin (Cell Signaling #2128). Primary and secondary antibodies were used at 1/1000 and 1/5000 dilutions, respectively. Importantly, primary antibodies against PAICS and GOLGA3 recognised C-terminal immunogens, ensuring that cleaved proteins could be detected.

Virus infections in siRNA-based cellular protein knockdowns. Host proteins were knocked-down in A549-Ace2 cells using specific dsiRNAs from IDT. Briefly, A549-Ace2 cells seeded at 1x104 cells/well in 96-well plates. After 24 h, each well was transfected with 5 pmol of individual dsiRNAs using Lipofectamine RNAiMAX (Thermo Fisher Scientific) according to the manufacturer's instructions. 24 h post-transfection, the cell culture supernatant was removed and replaced with virus inoculum (MOI of 0.1 PFU/cell). Following a 1 h adsorption at 37 °C, the virus inoculum was removed and replaced with fresh 2% FBS/DMEM media. Cells were incubated at 37 °C for 3 days before supernatants were harvested. Samples were either heat-inactivated at 80 °C for 20 min and viral RNA was quantified by RT-qPCR, using previously published SARS-CoV-2 specific primers targeting the N gene[77]. RT-qPCR was performed using the Luna Universal One-Step RT-qPCR Kit (NEB) in an Applied Biosystems QuantStudio 7 thermocycler, using the following cycling conditions: 55 °C for 10 min, 95 °C for 1 min, and 40 cycles of 95 °C for 10 sec, followed by 60 °C for 1 min. The quantity of viral genomes is expressed as PFU equivalents, and was calculated by performing a standard curve with RNA derived from a viral stock with a known viral titer. Alternatively, infectious virus titers were quantified using plaque assays as described above.

To quantify siRNA-based cellular protein knockdowns, A549-Ace2 cells were seeded and transfected with individual dsiRNAs as described above. After 24 h incubation at 37 °C cells were lysed and RNA was extracted using Trizol (ThermoFisher Scientific) followed by purification using the Direct-zol-96 RNA extraction kit (Zymo) following the manufacturer's instructions. RNA levels of target proteins were subsequently quantified by using RT-with the Luna Universal One-Step RT-qPCR Kit (NEB) in an Applied Biosystems QuantStudio 7 thermocycler using gene-specific primers. Expression levels were compared to scrambled dsiRNA-transfected cells und normalised to expression of human beta-actin. Knockdown efficiencies were calculated using Ct in Matlab.

To assess cell viability after siRNA knockdowns, cells were seeded and transfected as described above. 24 h after transfection cell viability was measured using alamarBlue reagent (ThermoFisher Scientific), media was removed and replaced with alamarBlue and incubated for 1 h at 37 °C and fluorescence measured in a Tecan Infinite M200 Pro plate reader. Percentage viability was calculated relative to untreated cells (100% viability) and cells lysed with 20% ethanol (0% viability), included in each plate.

For cell counting to determine cell numbers, cells were fixed in formalin to deactivate virus. The fixed cells were stained with 5 µg/ml of Hoechst 33258 (Sigma). The assay plates were imaged on an IX-83 automated inverted microscope (Olympus) using a ×10 objective. The DAPI settings (Ex UV 377/50, Em 415–480) were used to image Hoechst 33258. The acquisition setup was configured to image four sites per well. The nuclei were identified using the object detection module in the ScanR analysis software.

Drug screens and cytotoxicity analysis. Black with clear bottom 384-well plates were seeded with 2 × 103 A549-Ace2 cells per well. The following day, individual compounds were added using the Echo 550 acoustic dispenser at concentrations indicated 2 h prior to infection. DMSO-only (0.5%) and remdesivir (10 µM; SelleckChem) controls were added in each plate. After the pre-incubation period, the drug-containing media was removed, and replaced with virus inoculum (MOI of 0.1 PFU/cell). Following a one-hour adsorption at 37 °C, the virus inoculum was removed and replaced with 2% FBS/DMEM media containing the individual drugs at the indicated concentrations. Cells were incubated at 37 °C for 3 days. Supernatants were harvested and heat-inactivated at 80 °C for 20 min. Detection of viral genomes from heat-inactivated was performed by RT-qPCR as described above. Cytotoxicity was determined using the CellTiter-Glo luminescent cell viability assay (Promega). White with clear bottom 384-well plates were seeded with 2 × 103 A549-Ace2 cells per well. The following day, individual compounds were added using the Echo 550 acoustic dispenser at concentrations indicated. DMSO-only

(0.5%) and camptothecin (10 μM; Sigma–Aldrich) controls were added in each plate. After 72 h incubation, 20 μl/well of Celltiter-Glo reagent was added, incubated for 20 min and the luminescence was recorded using a luminometer (Berthold Technologies) with 0.5 sec integration time. Curve fits and IC50/CC50 values were obtained in Matlab.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The proteomics data generated in this study have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange. org) via the PRIDE repository[78]. Specifically the A549-Ace2 datasets can be found under PRIDE accession numbers PXD021145, PXD021152, and PXD021402, Vero E6 datasets under PXD021154, PXD021153, and PXD021403, and Protease Inhibitor experiments under PXD023539, PXD023538, and PXD023540. Raw images corresponding to cell viability have been deposited in the Zenodo repository under https://doi.org/10.5281/zenodo.4984272. Other reagent and oligo sequence details are described in Supplementary data 9. Source data are provided with this paper.

Code availability

All Matlab (R2019b) and R (v4.0.3) scripts used to process data and produce the figures in this manuscript can be accessed through the Emmott Lab Github page at: https://github.com/emmottlab/sars2nterm/, and on Zenodo, https://doi.org/10.5281/zenodo.5153020. The Matlab scripts used to process the mass spectrometry data and produce the figures in this manuscript have been tested in Matlab versions R2019b with the Statistics Machine Learning Toolbox, on Mac OS Catalina.

Received: 4 February 2021; Accepted: 24 August 2021;
Published online: 21 September 2021

References

1. Wang, C. et al. A novel coronavirus outbreak of global health concern. *Lancet* **395**, 470–473 (2020).
2. Zhu, N. et al. A novel coronavirus from patients with pneumonia in china, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
3. Davidson, A. D. et al. Characterisation of the transcriptome and proteome of sars-cov-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome Med.* **12**, 68 (2020).
4. Klann, K. et al. Growth factor receptor signaling inhibition prevents sars-cov-2 replication. *Mol. Cell*. https://doi.org/10.1016/j.molcel.2020.08.006 (2020).
5. Bojkova, D. et al. Proteomics of sars-cov-2-infected host cells reveals therapy targets. *Nature* **583**, 469–472 (2020).
6. Gordon, D. E. et al. A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
7. Laurent, E. M. N. et al. Global bioid-based sars-cov-2 proteins proximal interactome unveils novel ties between viral polypeptides and host factors involved in multiple covid19-associated mechanisms. *bioRxiv*. https://doi.org/10.1101/2020.08.28.272955 (2020).
8. Bouhaddou, M. et al. The global phosphorylation landscape of sars-cov-2 infection. *Cell* **182**, 685–712.e19 (2020).
9. Stukalov, A. et al. Multi-level proteomics reveals host perturbation strategies of sars-cov-2 and sars-cov. *bioRxiv*. https://doi.org/10.1101/2020.06.17.156455 (2020).
10. Ou, X. et al. Characterization of spike glycoprotein of sars-cov-2 on virus entry and its immune cross-reactivity with sars-cov. *Nat. Commun.* **11**, 1620 (2020).
11. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.e8 (2020).
12. Jin, Z. et al. Structure of mpro from sars-cov-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).
13. Rut, W. et al. Activity profiling and structures of inhibitor-bound sars-cov-2-plpro protease provides a framework for anti-covid-19 drug design. *bioRxiv*. https://doi.org/10.1101/2020.04.29.068890 (2020).
14. Moustaqil, M. et al. Sars-cov-2 proteases cleave irf3 and critical modulators of inflammatory pathways (nlrp12 and tab1): implications for disease presentation across species and the search for reservoir hosts. *bioRxiv* https://doi.org/10.1101/2020.06.05.135699 (2020).
15. Papa, G. et al. Furin cleavage of sars-cov-2 spike promotes but is not essential for infection and cell-cell fusion. *bioRxiv* https://doi.org/10.1101/2020.08.13.243303 (2020).
16. Shang, J. et al. Cell entry mechanisms of sars-cov-2. *Proc. Natl Acad. Sci. USA* **117**, 11727–11734 (2020).
17. Nelson, C. A. et al. Structure and intracellular targeting of the sars-coronavirus orf7a accessory protein. *Structure* **13**, 75–85 (2005).
18. Diemer, C. et al. Cell type-specific cleavage of nucleocapsid protein by effector caspases during sars coronavirus infection. *J. Mol. Biol.* **376**, 23–34 (2008).
19. Mark, J. et al. Sars coronavirus: unusual lability of the nucleocapsid protein. *Biochem. Biophys. Res. Commun.* **377**, 429–433 (2008).
20. Forni, G. et al. Covid-19 vaccines: where we stand and challenges ahead. *Cell Death Differ.*, **28**, 626–639 (2021).
21. Struwe, W. et al. The covid-19 ms coalition | accelerating diagnostics, prognostics, and treatment. *Lancet* **395**, 1761–1762 (2020).
22. Weng, S. S. H. et al. Sensitive determination of proteolytic proteoforms in limited microscale proteome samples. *Mol. Cell. Proteomics* **18**, 2335–2347 (2019).
23. Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
24. Finkel, Y. et al. The coding capacity of sars-cov-2. *Nature* https://doi.org/10.1038/s41586-020-2739-1 (2020).
25. Lutomski, C. A. et al. Proteoforms of the sars-cov-2 nucleocapsid protein are primed to proliferate the virus and attenuate the antibody response. *bioRxiv* https://doi.org/10.1101/2020.10.06.328112 (2020).
26. Kern, D. M. et al. Cryo-EM structure of the sars-cov-2 3a ion channel in lipid nanodiscs. *bioRxiv* https://doi.org/10.1101/2020.06.17156554 (2020).
27. Herrera, N. G. et al. Characterization of the sars-cov-2 s protein: biophysical, biochemical, structural, and antigenic analysis. *bioRxiv* https://doi.org/10.1101/2020.06.14.150607 (2020).
28. Chelius, D. et al. Formation of pyroglutamic acid from n-terminal glutamic acid in immunoglobulin gamma antibodies. *Anal. Chem.* **78**, 2370–2376 (2006).
29. Zhao, P. et al. Virus-receptor interactions of glycosylated sars-cov-2 spike and human ace2 receptor. *Cell Host Microbe* **28**, 586–601.e6 (2020).
30. Liang, J. G. et al. S-trimer, a covid-19 subunit vaccine candidate, induces protective immunity in nonhuman primates. *Nat. Commun.* **12**, 1346 (2021).
31. Fielding, B. C. et al. Characterization of a unique group-specific protein (u122) of the severe acute respiratory syndrome coronavirus. *J. Virol.* **78**, 7311–7318 (2004).
32. Hodcroft, E. B. *Covariants: Sars-cov-2 mutations and variants of interest*. https://covariants.org (2021).
33. McCallum, M. et al. SARS-CoV-2 immune evasion by variant B.1.427/B.1.429. *bioRxiv* https://doi.org/10.1101/2021.03.31.437925 (2021).
34. Schechter, I. & Berger, A. On the size of the active site in proteases. i. papain. *Biochem. Biophys. Res. Commun.* **27**, 157–162 (1967).
35. Biniossek, M. L. et al. Proteomic identification of protease cleavage sites characterizes prime and non-prime speci city of cysteine cathepsins b, l, and s. *J. Proteome Res.* **10**, 5363–5373 (2011).
36. Wrobel, A. G. et al. Sars-cov-2 and bat ratg13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.* **27**, 763–767 (2020).
37. Walls, A. C. et al. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell* **181**, 281–292.e6 (2020).
38. Fortelny, N. et al. Proteome TopFIND 3.0 with TopFINDer and PathFINDer: database and analysis tools for the association of protein termini to pre- and post-translational events. *Nucleic Acids Res.* **43**, D290–D297 (2014).
39. Rut, W. et al. Substrate specificity profiling of sars-cov-2 main protease enables design of activity-based probes for patient-sample imaging. *bioRxiv*. https://doi.org/10.1101/2020.03.07.981928 (2020).
40. Knapp, S. New opportunities for kinase drug repurposing and target discovery. *Br. J. Cancer* **118**, 936–937 (2018).
41. Szilagyi, K. L. et al. Epigenetic contribution of the myosin light chain kinase gene to the risk for acute respiratory distress syndrome. *Transl. Res. J. Lab. Clin. Med.* **180**, 12–21 (2017).
42. Nofrini, V., Di Giacomo, D. & Mecucci, C. Nucleoporin genes in human diseases. *Eur. J. Hum. Genet.* **24**, 1388–1395 (2016).
43. Wada, M. et al. Interplay between coronavirus, a cytoplasmic rna virus, and nonsense-mediated mrna decay pathway. *Proc. Natl Acad. Sci. USA* **115**, E10157–E10166 (2018).
44. Generous, A. et al. Identification of putative interactions between swine and human influenza a virus nucleoprotein and human host proteins. *Virol. J.* **11**, 228–228 (2014).
45. Malioutov, D. et al. Quantifying homologous proteins and proteoforms. *Mol. Cell. Proteomics* **18**, 162–168 (2019).
46. Yin, X. et al. Mda5 governs the innate immune response to sars-cov-2 in lung epithelial cells. *Cell Rep.* **34**, 108628 (2021).
47. Xia, Z. et al. Inducible TAP1 negatively regulates the antiviral innate immune response by targeting the TAK1 complex. *J. Immunol.* **198**, 3690–3704 (2017).
48. Drayman, N. et al. Drug repurposing screen identifies masitinib as a 3clpro inhibitor that blocks replication of sars-cov-2 in vitro. *bioRxiv*. https://doi.org/10.1101/2020.08.31.274639 (2020).

49. Emmott, E. et al. Norovirus-mediated modification of the translational landscape via virus and host-induced cleavage of translation initiation factors. *Mol. Cell. Proteomics* **16**, S215–S229 (2017).
50. Emmott, E., Sweeney, T. R. & Goodfellow, I. A cell-based fluorescence resonance energy transfer (fret) sensor reveals inter- and intragenogroup variations in norovirus protease activity and polyprotein cleavage. *J. Biol. Chem.* **290**, 27841–27853 (2015).
51. Gevaert, K. et al. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted n-terminal peptides. *Nat. Biotechnol.* **21**, 566–569 (2003).
52. McDonald, L. & Beynon, R. J. Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterization. *Nat. Protoc.* **1**, 1790–1798 (2006).
53. Kleifeld, O. et al. Isotopic labeling of terminal amines incomplex samples identifies protein n-termini and protease cleavage products. *Nat. Biotechnol.* **28**, 281–288 (2010).
54. Jagdeo, J. M. et al. N-terminomics tails identifies host cell substrates of poliovirus and coxsackievirus b3 3c proteinases that modulate virus infection. *J. Virol.* https://doi.org/10.1128/JVI.02211-17 (2018).
55. Saeed, M. et al. Defining the proteolytic landscape during enterovirus infection. *PLoS Pathog.* **16**, 1–28 (2020).
56. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry based proteomics. *J. Proteome Res.* **9**, 1323–1329 (2010).
57. Giansanti, P. et al. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat. Protoc.* **11**, 993–1006 (2016).
58. Koudelka, T. et al. N-terminomics for the identification of in vitro substrates and cleavage site specificity of the sars-cov-2 main protease. *Proteomics* **21**, 2000246 (2021).
59. Mykytyn, A. Z. et al. SARS-CoV-2 entry into human airway organoids is serine protease-mediated and facilitated by the multibasic cleavage site. *Elife* **10**, e64508 (2021).
60. Buchrieser, J. et al. Syncytia formation by sars-cov-2-infected cells. *EMBO J.* **39**, e106267 (2020).
61. Li, J. et al. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, 399–404 (2020).
62. Hughes, C. S. et al. Singlepot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).
63. Tyanova, S., Temu, T. & Cox, J. The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
64. Kolbowski, L., Combe, C. & Rappsilber, J. xiSPEC: web-based visualization, analysis and sharing of proteomics data. *Nucleic Acids Res.* **46**, W473–W478 (2018).
65. MikeCF. *Break y axis* https://www.mathworks.com/matlabcentral/fileexchange/45760-break-y-axis, matlab central file exchange. (2021).
66. Goddard, T. D. et al. Ucsf chimerax: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
67. Armenteros, J. J. A. et al. Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
68. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021).
69. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
70. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* **64**, 479–498 (2002).
71. Yu, G. et al. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
72. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci USA.* **102**, 15545–15550 (2005).
73. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
74. Ou, J. et al. daglogo: An r/bioconductor package for identifying and visualizing differential amino acid group usage in proteomics data. *PLoS ONE* **15**, 1–20 (2020).
75. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).
76. Hunter, D. J. B. et al. Unexpected instabilities explain batch-to-batch variability in cell-free protein expression systems. *Biotechnol. Bioeng.* **115**, 1904–1914 (2018).
77. Chu, D. K. W. et al. Molecular diagnosis of a novel Coronavirus (2019-nCoV) causing an outbreak of pneumonia. *Clin. Chem.* **66**, 549–555 (2020).
78. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2018).

Acknowledgements

We thank members of the Centre for Proteome Research, especially Rob Beynon and Jos Sarsby, as well as Nikolai Slavov & Aleksandra Petelski (Northeastern University) for constructive comments. We also thank Agnès Zettor and Soizick Lucas-Staat from the Chemogenomic and Biological Screening Platform for their technical assistance. We thank Julian Buchrieser, Olivier Schwartz, Pierre Charneau, Cyril Planchais, and Hugo Mouquet for the gift of reagents. The A549-Ace2, HEK-ACE2, and HEK-ACE2-TMPRSS2 cells were a gift from Olivier Schwartz (Institut Pasteur). We would like to thank Katherina Michie, Jack Bennett, and key personnel at the protein production facility of UNSW for the purification of PLpro and 3CLpro of SARS-CoV-2. We would like to thank Emma Ollivier for the initial in vitro cleavage work on SRC and for useful discussions and comments, and Dominic J.B Hunter for production of cell-free extracts. We would also like to thank Prof. Alexandrov for the cell-free plasmids compatible with LTE protein production. This work was supported by the Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (grant ANR-10-LABX-62-IBEID) to M.V. S.G. is the recipient of a MESR/Ecole Doctorale BioSPC ED562, Université de Paris fellowship. L.A.C. is supported by Institut Pasteur TASK FORCE SARS COV-2 (Tropicoro project), DIM ELICIT Region Ile-de-France, and ANRS. E.E. is supported by startup funding from the University of Liverpool, as well as a Wellcome Trust ISSF Interdisciplinary & Industry Award. E.E. is grateful for the support of GoFundMe donors for sponsoring SARS-CoV-2 research in his laboratory.

Author contributions

E.E. conceived the study with B.M. B.M., J.C., S.G., D.B., L.D., M.W., Y.G., E.S., and E.E. performed experiments and analysed data. F.A., L.C., C.C., C.E., P.E., Y.G., A.J., E.S., E.V., M.V., and E.E. advised on experimental design, provided reagents, and supervision. E.E., A.G., and A.J. performed bioinformatic analysis of the proteomic data. P.B. provided technical expertise. All authors commented on the results and implications, and commented on the draft. E.E. and B.M. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25796-w.

Correspondence and requests for materials should be addressed to Edward Emmott.

Peer review information *Nature Communications* thanks Andreas Pichlmair, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.