Evolutionary history of type II transmembrane serine proteases involved in viral priming

Diego Forni1 · Manuela Sironi1 · Rachele Cagliani1

Received: 11 October 2021 / Accepted: 15 January 2022 / Published online: 5 February 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Type II transmembrane serine proteases (TTSPs) are a family of trypsin-like membrane-anchored serine proteases that play key roles in the regulation of some crucial processes in physiological conditions, including cardiac function, digestion, cellular iron homeostasis, epidermal differentiation, and immune responses. However, some of them, in particular TTSPs expressed in the human airways, were identified as host factors that promote the proteolytic activation and spread of respiratory viruses such as influenza virus, human metapneumovirus, and coronaviruses, including SARS-CoV-2. Given their involvement in viral priming, we hypothesized that members of the TTSP family may represent targets of positive selection, possibly as the result of virus-driven pressure. Thus, we investigated the evolutionary history of sixteen TTSP genes in mammals. Evolutionary analyses indicate that most of the TTSP genes that have a verified role in viral proteolytic activation present signals of pervasive positive selection, suggesting that viral infections represent a selective pressure driving the evolution of these proteases. We also evaluated genetic diversity in human populations and we identified targets of balancing selection in TMPRSS2 and TMPRSS4. This scenario may be the result of an ancestral and still ongoing host–pathogen arms race. Overall, our results provide evolutionary information about candidate functional sites and polymorphic positions in TTSP genes.

Introduction
In order to initiate a successful infection, viruses must deliver their genome to the host cell cytosol. To achieve this, enveloped viruses trigger a sequence of events that lead to the fusion of their membrane with that of the target host cell, either by direct fusion with the plasma membrane or by an internalization process into endosomes (Barrett and Dutch 2020; Harrison 2008). Usually, this process is mainly mediated by the interaction of specific viral attachment proteins expressed on the surface of the virion (envelope glycoproteins) with viral receptors. These latter are cell surface molecules that perform specific host cellular functions but are hijacked by viruses to assist infection. Multiple types of molecules, including proteins, carbohydrates, and lipids, can be used as viral receptors. However, proteins are considered to be more suitable receptors because of their stronger affinity and higher specificity for viral attachment, resulting in an increase in the efficiency of viral entry and in an expansion of the viral host range and tropism (Zhang et al. 2019a, b; Maginnis 2018).

Viral receptors favor the recognition and attachment of the virus to the host cell, but the activation of the membrane fusion machinery often requires the involvement of additional host factors. In fact, most viral fusion proteins are synthesized as a precursor and require processing by host cell proteases to trigger the fusion of the viral lipid envelope with the cellular membrane. Several cellular proteases, including proprotein convertases (i.e., furin), trypsin, cathepsins, and type II transmembrane serine proteases (TTSPs), have been described to participate in the proteolytic activation of viral proteins. TTSPs are a family of trypsin-like membrane-anchored serine proteases that play a key role in the regulation of some crucial processes, including cardiac function, digestion, cellular iron homeostasis, epidermal differentiation, and immune responses (Bugge et al. 2009). Consequently, deregulation of these enzymes is an important feature of several pathological conditions (List et al. 2002; Du et al. 2008; Scott et al. 2001), including tumorigenesis (Choi et al. 2009).
TTSPs expressed in the human airways were identified as host factors that promote the proteolytic activation and spread of respiratory viruses such as influenza virus, human metapneumovirus, paramyxoviruses, human parainfluenza viruses, Sendai virus, orthoreoviruses, Group A rotaviruses and coronavirus (Shirogane et al. 2008; Bertram et al. 2010, 2013; Shirato et al. 2013; Sakai et al. 2014; Sasaki et al. 2021; Nygaard et al. 2012), including SARS-CoV-2, the causative agent of the current COVID-19 pandemic (Zang et al. 2020; Hoffmann et al. 2020b).

For a successful infection, influenza viruses depend on the proteolytic activation of their entry-mediating surface protein, hemagglutinin (HA). HA may be activated by diverse host trypsin-like proteases in vitro or in vivo. In particular, influenza viruses (e.g. H1N1 and H7N9 strains) hijack transmembrane serine proteases (including TMPRSS2, TMPRSS4, TMPRSS13, TMPRSS11D, TMPRSS11E, and ST14) (Laporte and Naesens 2017), that cleave HA into subunits HA1 and HA2, a prerequisite for conformational changes that trigger fusion.

In coronaviruses, it is the spike (S) protein that promotes cell entry via membrane fusion (Shang et al. 2020; Li 2016). After receptor engagement, the coronavirus S protein is enzymatically processed by cellular proteases, mediating efficient membrane fusion with host cells (Bertram et al. 2013; Shirato et al. 2013; Hoffmann et al. 2020a, b; Matsuyama et al. 2020; Simmons et al. 2005).

Human coronaviruses can use a variety of cellular proteases to prime their S protein, including TMPRSS2 and other members of the TTSP family (e.g., TMPRSS4 and TMPRSS11) (Zang et al. 2020; Millet and Whittaker 2015). Likewise, coronaviruses can utilize different receptors to promote their entry into the host cells: both SARS-CoV and SARS-CoV-2 use angiotensin-converting enzyme 2 (ACE2), whereas MERS-CoV and HCoV-229E exploit DPP4 (dipeptidyl peptidase 4) and ANPEP (aminopeptidase N or CD13), respectively (Hoffmann et al. 2020b; Raj et al. 2013; Yeager et al. 1992; Forni et al. 2017).

Furthermore, previous studies have shown that SARS-CoV and other coronaviruses interact with multiple cell-surface molecules, such as CD209 (DC-SIGN), CLEC4M (L-SIGN/CD299), CLEC4G (LSECtin), and basigin (BSG or CD147) (Yang et al. 2004; Marzi et al. 2004; Gramberg et al. 2005; Vankadari and Wilce 2020; Wang et al. 2020a). Interestingly, many of these receptors are targets of positive selection (Cui et al. 2013; Bibiana et al. 2020; Guo et al. 2020; Damas et al. 2020; Ortiz et al. 2008; Enard et al. 2016; Sironi et al. 2015; Wang et al. 2020b).

In line with its role in the activation of viral glycoproteins, TMPRSS2 variants were associated with a higher risk of severe disease and increased susceptibility to human influenza (Cheng et al. 2015). Variants in the same gene they may also play a significant role in the interindividual differences, particularly in the gender-related bias, of COVID-19 susceptibility and severity (Alshahawey et al. 2020). Moreover, TMPRSS2 polymorphisms may explain the higher mortality rate among Italians and the association to SARS-CoV-2 infection susceptibility in different populations (Asselta et al. 2020; Irham et al. 2020; Piva et al. 2021; Vargas-Alarcón et al. 2020; Hou et al. 2020).

Given their involvement in viral priming, we reasoned that members of the TTSP family may also represent targets of positive selection, possibly as the result of virus-driven pressure. To test this hypothesis, we systematically analyzed the evolutionary patterns of TTSP genes in a large phylogeny of mammals. We also evaluated their genetic diversity in human populations.

**Material and methods**

**Sequences, alignments, and phylogenetic trees**

Coding sequences for TTSPs were retrieved from the UCSC server (http://genome.ucsc.edu/). Mammalian homologs of human genes were included only if they represented 1-to-1 orthologs, as reported in the EnsemblCompara GeneTrees (Vilella et al. 2009). Sequences that carry stop codons or have a sequence coverage of less than 50% were excluded. The list of species analyzed for each gene is reported in Supplementary Table S1.

The RevTrans 2.0 utility was used to generate multiple sequence alignments (http://www.cbs.dtu.dk/services/RevTrans/, MAFFT v6.240 as an aligner) (Wernersson and Pedersen 2003). Phylogenetic trees were reconstructed using the phyML program with gamma-distributed rates, 4 substitution rate categories, and estimation of transition/transversion ratio and proportion of invariable sites (Guindon et al. 2009).

**Evolutionary analysis in mammals**

All alignments were screened for the presence of recombination using genetic algorithm recombination detection (GARD) (Kosakovsky Pond et al. 2006). GARD is a genetic algorithm implemented in the HYPHY suite, which uses phylogenetic incongruence among segments in the alignment to detect the best-fit number and location of recombination breakpoints.

To obtain an overview of the selective patterns of mammalian TTSPs coding genes, the average nonsynonymous substitution (dN)/synonymous substitution (dS) rate ratio was calculated using the single-likelihood ancestor counting (SLAC) method (Kosakovsky Pond and Frost 2005).

To detect positive selection, the codon-based codeml program implemented in the phylogenetic analysis by maximum likelihood (PAML) suite was applied (Yang 2007).
Specifically, a model (M8, positive selection model) that allows a class of sites to evolve with \( dN/dS > 1 \) was compared to two models (M7 and M8a, neutral models) that do not allow \( dN/dS > 1 \). To assess statistical significance, twice the difference of the likelihood (\( \Delta \ln L \)) for the models (M8a vs M8 and M7 vs M8) is compared to a \( \chi^2 \) distribution (1 or 2 degrees of freedom for M8a vs M8 and M7 vs M8 comparisons, respectively). To assure reliability, two different codon substitution models were used (F3 x 4 and F61).

Positively selected sites were identified using the following methods: (1) the Bayes empirical Bayes (BEB) analysis (with a cutoff of 0.90), which calculates the posterior probability that each codon is from the site class of positive selection (under model M8) (Anisimova et al. 2002); (2) fast unbiased Bayesian AppRoximation (FUBAR) (Murrell et al. 2013), an approximate hierarchical Bayesian method that generates an unconstrained distribution of selection parameters to estimate the posterior probability of positive diversifying selection at each site in a given alignment (with a cutoff \( \geq 0.90 \)); (3) the fixed effects likelihood (FEL) (Kosakovsky Pond and Frost 2005), a maximum-likelihood (ML) approach to infer \( dN/dS \) on a per-site basis, assuming that the selection pressure for each site is constant along the entire phylogeny (with a cutoff of \( \geq 50 \)). To be conservative, a site was considered under positive selection if it was detected by at least two methods. Protein structures were visualized with the PyMOL software (Schrödinger 2017), which was also used to create 3D protein figures.

GARD, FEL, FUBAR, and SLAC analyses were run locally through the HyPhy suite (Pond et al. 2005).

In order to identify specific branches with a proportion of sites evolving with \( dN/dS > 1 \), the adaptive Branch-Site Random Effects Likelihood method (aBS-REL) was used. This method applies sequential likelihood ratio tests to estimate the posterior probability of positive diversifying selection at each site in a given alignment (with a cutoff \( \geq 0.90 \)); (3) the fixed effects likelihood (FEL) (Kosakovsky Pond and Frost 2005), a maximum-likelihood (ML) approach to infer \( dN/dS \) on a per-site basis, assuming that the selection pressure for each site is constant along the entire phylogeny (with a cutoff of \( \geq 50 \)). To be conservative, a site was considered under positive selection if it was detected by at least two methods. Protein structures were visualized with the PyMOL software (Schrödinger 2017), which was also used to create 3D protein figures.

**Population genetic analyses**

Data from the 1000 Genomes Phase 1 Project (1000 Genomes Project Consortium et al. 2010) for individuals with East Asian (AS; Japanese plus Chinese) or European ancestry (CEU), and Yoruba subjects from Nigeria (YRI), were used to calculate nucleotide diversity and site frequency spectrum statistics for seven TTSP genes. Nucleotide diversity over the whole gene was measured as the average number of pairwise sequence nucleotide differences, \( \pi \) (Nei and Li 1979) and the expected per site heterozygosity, \( \theta_N \) (Watterson 1975). The same statistics were also calculated in 5 kb sliding windows moving with a step of 250 bp. Three site frequency spectrum (SFS) statistics (Tajima’s \( D \), \( DT \)) (Tajima 1989), Fu and Li’s \( D^* \) and \( F^* \) (Fu and Li 1993) and nucleotide diversity indicators were calculated for gene regions showing high level of diversity in the sliding-windows approach.

In order to obtain a statistical significance for these values, the same tests were performed for a control set of ~2000 genes, as previously described (Forni et al. 2014, 2013). Briefly, ~2000 genes were randomly selected, with the only limitations to be longer than 5000 bp and to have more than 80% human-outgroup (chimpanzee or orangutan genomes) aligning bases (Forni et al. 2014, 2013). This approach allowed us to generate a distribution of the above statistics and the calculation of empirical percentiles. We considered as statistically significant those values with a percentile rank higher than the 95th. For the gene region analysis, we selected 5 kb independent gene regions from the 2000 random genes and we used them to generate SFS distributions.

Orthologous genomic regions in chimpanzee, orangutan or macaque were retrieved using the liftOver tool (Kuhn et al. 2013).

**Results and discussion**

**Evolutionary analysis of TTSPs in mammals**

In humans, the TTSP family comprises 17 known members that share a number of common structural features (Fig. 1A). These include a short N-terminal cytoplasmic domain, a transmembrane domain, a variable stem region, and a C-terminal serine protease domain with a catalytic triad (Ser-His-Asp) (Böttcher-Friebertshäuser 2018). The stem region contains an assortment of modular structural domains that contribute to TTSP activation and interaction with substrates (Hooper et al. 2001; Antalis et al. 2010). Based on the composition of the extracellular protein portion and similarity in domain structure, chromosomal localization, and phylogenetic analysis of the serine protease domains, TTSPs are grouped in four subfamilies known as HAT/DESC, hepsin/TMPRSS (transmembrane protease/serine S1), matriptase, and corin (Böttcher-Friebertshäuser 2018; Szabo and Bugge 2008) (Fig. 1A).

TTSPs expressed in the human airways have been identified as host cell factors that may support the proteolytic activation and spread of respiratory viruses, including influenza viruses and coronaviruses. In particular, in vitro studies have demonstrated the involvement of 8 TTSP proteases (TMPRSS11D, TMPRSS11E, TMPRSS11A, TMPRSS2, TMPRSS4, TMPRSS13, TMPRSS15, and ST14) in the activation of HA and of spike proteins (Zang et al. 2020; Hoffmann et al. 2020b; Böttcher-Friebertshäuser 2018; Hayashi et al. 2018). Other viruses also require TTSPs (at
least in vitro) for entry into the host cell. Indeed mammalian orthoreovirus, paramyxoviruses, human metapneumovirus, and human parainfluenza viruses require proteolytic activation through TMPRSS2 and/or TMPRSS11D at the viral entry step (Nygaard et al. 2012; Böttcher-Friebertshäuser 2018). More recently, Sasaki and colleagues demonstrated that TMPRSS2 and TMPRSS11D also mediate proteolytic activation and trypsin-independent infection in group
A rotavirus (Sasaki et al. 2021) (Supplementary Table S2, Fig. 1A).

To comprehensively analyze the selective pressure acting on TTSPs and to obtain an estimate of the extent of constraint acting on these genes, we retrieved coding sequence information for all available mammalian species in the UCSC database (Supplementary Table 1). In particular, we retrieved sequences for at least 46 species for 16 TTSP genes. We excluded TMPRSS11B from the analysis, as the existence of a human pseudogene (TMPRSS11BNL) could affect evolutionary inference. Concerning TMPRSS13, we excluded from the analysis the region encoding the first 89 amino acids, as it showed limited similarity across species.

Because recombination can generate false positive inferences of positive selection (Sironi et al. 2015; Anisimova et al. 2003), all alignments were screened for the presence of recombination using genetic algorithm recombination detection (GARD) (Kosakovsky Pond et al. 2006). Evidence of recombination was detected only for the CORIN gene. The CORIN coding alignment was thus split on the basis of the recombination breakpoints in three sub-regions (CORIN-reg1-3) and then used for the analyses.

For coding genes, the strength of selection can be quantified by comparing the rate of non-synonymous nucleotide substitutions per non-synonymous site (dN) with that of synonymous substitutions per synonymous site (dS) (Hughes and Nei 1989). We thus calculated the average dN/dS values calculated for TTSP genes. p values from the Wilcoxon rank sum test for each gene are reported. TTSPs involved (red) or not involved (blue) in the activation of viral glycoproteins are indicated. Branch-site analysis of positive selection was applied using the site models implemented in the codeML tool from the PAML suite (see methods). Using likelihood ratio tests (LRTs), codeML compares models of gene evolution that allow (model M8) or disallow (models M8a and M7) a class of codons to evolve with dN/dS > 1. Specifically, M8 represents the positive selection model that is tested against the neutral M8a and M7 models.

To assess statistical significance, twice the difference of the likelihood (ΔlnL) for the models (M7 vs M8 and M8a vs M7) was compared to a χ² distribution (2 degrees of freedom for the M7 vs M8 comparison, 1 degree of freedom for M8a vs M8). To be conservative and to obtain robust results we applied two different codon frequency models (F3 × 4 and F61) and we called a gene as a target of positive selection only if it was detected by both comparisons (M7 versus M8, M8a versus M8) with both models.

Applying these criteria, we found seven TTSP genes (TMPRSS11D, TMPRSS11E, TMPRSS11A, TMPRSS15, TMPRSS2, TMPRSS4, CORIN-reg1 and CORIN-reg2) to be targets of positive selection in mammals (Table 1). To identify specific sites targeted by positive selection, we again applied a conservative strategy and we called a site as positively selected only if it was detected by at least two of the following methods: BEB analysis from the M8 model, FUBAR and FEL methods (Kosakovsky Pond and Frost 2005; Murrell et al. 2013). Positively selected sites were identified in each of the seven genes and, as evident in Fig. 1A, they were not clustered, but rather distributed along the protein sequences. The only exception was TMPRSS11D, where a preponderance of selected sites was observed in the protease S1 domain.

We next extended our analysis to explore possible variations in selective pressure across lineages. In particular, we aimed to assess whether specific internal branches in the mammalian phylogenies evolved under episodic positive selection. Because we did not want to make any a priori assumption about which lineages were more likely to have experienced adaptive evolution, we applied the branch-site random effects likelihood (aBS-REL) method (Smith et al. 2004).
Table 1  Likelihood ratio test statistics for models of variable selective pressure among sites (F3 × 4 and F61 codon frequency model) for membrane-associated type II transmembrane serine protease genes

| Gene/LRT model | n. of species | n. of species | F3 × 4 – 2ΔlnL<sup>a</sup> | p value<sup>b</sup> | F61 – 2ΔlnL<sup>a</sup> | p value<sup>b</sup> | Positively selected sites<sup>c</sup> |
|----------------|---------------|---------------|-----------------------------|-----------------|-----------------------------|-----------------|----------------------------------|
| TMPRSS11D      | 57            | 57            | 19.9                        | 2.81 × 10<sup>-10</sup> | 29.29                       | 1.95 × 10<sup>-14</sup> | 75T, 156E, 215S, 248F, 254R, 259L, 260L, 265 K, 314Q |
| M7 vs M8       | 24.57         | 35.03         |                             |                 |                             |                 |                                  |
| TMPRSS11E      | 57            | 57            | 15.62                       | 2.28 × 10<sup>-8</sup> | 17.68                       | 2.74 × 10<sup>-9</sup> | 98E, 252K, 324Y, 347A          |
| M7 vs M8       | 27.33         | 27.12         |                             |                 |                             |                 |                                  |
| TMPRSS11A      | 54            | 54            | 2.95                        | 0.0151          | 3.73                        | 0.0063          | 14N, 39V, 62I, 413N             |
| M7 vs M8       | 9.37          | 8.42          |                             |                 |                             |                 |                                  |
| TMPRSS11F      | 56            | 56            | 0                           | 1               | 0                           | 1               |                                  |
| M7 vs M8       | 0             | 0             |                             |                 |                             |                 |                                  |
| HPN            | 57            | 57            | 2.12                        | 0.0395          | 4.36                        | 0.0031          | 103C, 140G, 301P, 433V, 489I    |
| M7 vs M8       | 9.6           | 11.63         |                             |                 |                             |                 |                                  |
| TMPRSS3        | 54            | 54            | 0                           | 1               | 0.16                        | 0.57            |                                  |
| M7 vs M8       | 10.62         | 5.53          |                             |                 |                             |                 |                                  |
| TMPRSS4        | 56            | 56            | 9.26                        | 1.68 × 10<sup>-5</sup> | 8.56                       | 3.51 × 10<sup>-5</sup> | 103A, 161D, 166E, 179S, 298F, 319T |
| M7 vs M8       | 12.71         | 13.74         |                             |                 |                             |                 |                                  |
| TMPRSS13       |               |               | 1.33                        | 0.1             | -0.18                       | 1               |                                  |
| M7 vs M8       | 6.92          | 3             |                             |                 |                             |                 |                                  |
| TMPRSS5        | 47            | 47            | 1.17                        | 0.13            | 0.03                        | 0.81            |                                  |
| M7 vs M8       | 2.76          | 0.063         |                             |                 |                             |                 |                                  |
| TMPRSS15       | 57            | 57            | 33.74                       | 2.13 × 10<sup>-16</sup> | 39.03                       | 9.99 × 10<sup>-19</sup> | 12H, 26A, 389A, 486I, 499Y, 545S, 621 V, 669E, 701 T, 918 V, 997L |
| M7 vs M8       | 59.31         | 61.51         |                             |                 |                             |                 |                                  |
| ST14           | 58            | 58            | 0.12                        | 0.62            | 9.68                        | 1.08 × 10<sup>-5</sup> |                                  |
| M7 vs M8       | 13.17         | 17.05         |                             |                 |                             |                 |                                  |
| TMPRSS6        | 57            | 57            | 0                           | 1               | 1.15                        | 0.13            |                                  |
| M7 vs M8       | 11.86         | 17.12         |                             |                 |                             |                 |                                  |
| TMPRSS7        | 56            | 56            | 0.12                        | 0.62            | 0.82                        | 0.2             |                                  |
| M7 vs M8       | 7.97          | 9.01          |                             |                 |                             |                 |                                  |
| TMPRSS9        | 56            | 56            | 0.546                       | 0.3             | 6.58                        | 0.0003          |                                  |
| M7 vs M8       | 6.666         | 11.66         |                             |                 |                             |                 |                                  |
| CORIN-reg1     | 57            | 57            | 0.12                        | 0.62            | 0.82                        | 0.2             |                                  |
| M7 vs M8       | 7.97          | 9.01          |                             |                 |                             |                 |                                  |
2015) to the seven positively selected TTSPs. For three genes, aBS-REL identified the presence of a few positively selected branches with borderline p values (Fig. 1C and Supplementary Fig. S1). In particular, evidence of selection was detected for the branch leading to Eptesicus fuscus and the Myotis clade for TMPRSS11D, to the myomorpha suborder of Rodentia for TMPRSS11A, and to some lineages of the carnivora order for TMPRSS2. However, these results were not confirmed using the codeml branch-site LRT models (Zhang et al. 2005), suggesting that positive selection in TTSP genes is mainly pervasive rather than episodic.

As mentioned above, an involvement of TTSP genes in the activation and spread of influenza viruses and/or coronaviruses was verified for 8 TTSP genes (Böttcher-Friebertshäuser 2018; Hayashi et al. 2018). Notably, six of the seven positively selected genes (TMPRSS11D, TMPRSS11E, TMPRSS11A, TMPRSS15, TMPRSS2, and TMPRSS4) are involved in the processing of viral glycoproteins, strongly suggesting that viral infections could represent the selective pressure driving the evolution of these TTSPs. Conversely, proteases with no known role in viral cell entry displayed no positive selection signals. The only exception was the CORIN gene, in which positive selection signals were identified in two of the 3 analyzed sub-regions.

Corin is a cardiac membrane serine protease and plays a key role in maintaining cardiac function and regulating blood pressure (Yan et al. 2000). Corin deficiency causes spontaneous hypertension and cardiac hypertrophy in mice (Chan et al. 2005), and variants in the human CORIN gene can alter corin protein conformation, inhibit corin zymogen activation, and contribute to hypertension and cardiovascular diseases (Zhou and Wu 2014; Peng et al. 2015).

The N-terminal cytoplasmic tail and the transmembrane domain are unnecessary for corin expression and catalytic activity (Knappe et al. 2003). However truncation of the cytoplasmic tail can create a new N-terminus that inhibits corin trafficking in the Golgi (He et al. 2020). Interestingly, three out of the 16 positively selected sites localized to the N-cytoplasmic tail. In particular, positively selected site Cys13 is polymorphic in human populations (rs2289433, p.Cys13Tyr) and it has been associated with serum corin levels (Zhang et al. 2019a, b).

Taking into consideration the physiological functions of this protein and that, up to date, the involvement of corin in viral activation and spread has not been reported, the signal of positive selection is most likely driven by non-viral factors.

For the other six positively selected genes, their role in the processing of viral glycoproteins has been widely described (Böttcher-Friebertshäuser 2018; Hayashi et al. 2018; Zmora et al. 2018; Kishimoto et al. 2021; Zang et al. 2020), supporting the hypothesis that viral infections represented the main selective force in the evolutionary history of these genes. For instance, in swine and mouse, TMPRSS2 can activate HA with monobasic cleavage site, indicating that homologous proteases are involved in HA cleavage in different host species (Tarnow et al. 2014; Peitsch et al. 2014).

In addition to the canonical activator TMPRSS2, other TTSPs have been reported to function as entry cofactors of different viral fusion proteins, at least in vitro. For example, it was recently reported that TMPRSS4 works synergistically with TMPRSS2 to promote SARS-CoV-2 infection of small intestinal enterocytes (Zhang et al. 2020), and three members of the HAT/DESC subfamily (TMPRSS11D, TMPRSS11A, and TMPRSS11E) are able to activate the SARS-CoV, SARS-CoV-2, and MERS-CoV spike proteins.
as well as influenza HA (Böttcher-Friebertshäuser 2018; Hayashi et al. 2018; Zmora et al. 2018; Kishimoto et al. 2021; Zhang et al. 2020; Bertram et al. 2011).

The redundancy of function of the different proteases may appear to be in contrast with the presence of positive selection signals in the TTSP genes. However, we wish to mention that several lines of evidence suggest that this might not be the case. First, not all TTSPs have the same tissue/cell expression and not all of them may have the same cleavage efficiency. Second, protease activation may occur through different mechanisms in cells, as recently suggested by Liu and colleagues based on proteomic interaction data (Liu et al. 2021). Third, most data suggesting redundancy of TTSP function in terms of viral protein cleavage have been obtained in vitro and do not necessarily translate to the situation in vivo or in natural infections. Finally, since more proteases are able to cleave proteins of different viruses and the same viral protein can be cleaved by more proteases, different infectious agents may have exerted distinct selection pressures on the same gene, and this could explain, at least in part, the presence of selected sites distributed in different domains.

As stated above, positively selected sites in TTSPs genes are not clustered, but rather distributed along the protein sequences. However, all TTSPs involved in the processing of viral glycoproteins have at least one selected site in the peptidase domain (Fig. 1A). Although such sites are not located near the active site (catalytic triad) of the protease domain (Fig. 2), we cannot exclude that they may have a role in modulating the proteolytic activity of these proteins. Likewise, we cannot rule out the possibility that the sites identified in the protease domains, as well as in other protein regions, are involved in direct interactions with the viral target proteins (e.g., spike protein in coronavirus, HA in influenza viruses, VP4 protein in group A rotaviruses). It is currently unknown whether proteases are able to establish direct interactions with their substrates.

In summary, evolutionary analyses indicate that most of the TTSP genes that have a verified role in proteolytic activation and spread of influenza viruses and/or coronaviruses present pervasive positive selection signals. This evidence strongly suggests that viral infections represent a major selective pressure driving the evolution of these type II transmembrane serine proteases in mammals. Nonetheless, we cannot exclude that the selection signals are due, at least partially, to other selective pressures associated with the physiological functions of these proteins. Future studies are required to clarify this aspect and to evaluate the functional relevance and selective advantages associated with these genetic changes in TTSP proteins during evolution.

### Population genetic analysis in humans

Infectious diseases have exerted strong selective pressures throughout human history (Haldane 1932; Fumagalli et al. 2011). Indeed, a number of human loci have been shown to evolve in response to such pressures; the best documented example is the major histocompatibility complex (MHC) (Bernatchez and Landry. 2003), which is highly polymorphic in several species as a consequence of balancing selection. This is a process whereby genetic variability is maintained in populations due to some selective pressure. Molecular signals of balancing selection have also been described for several genes involved in immune defense, including toll-like receptors (Fumagalli et al. 2009; Ferrer-Admetlla et al. 2008), various cytokine genes (Fumagalli et al. 2009; Ferrer-Admetlla et al. 2008), antimicrobial peptides (AMPs) (Cagliani et al. 2008; Hollox and Armour. 2008), and TRIM5 (Cagliani et al. 2010).

Thus, we addressed the role of natural selection in the shaping of genetic diversity at the seven positively selected TTSP genes in human populations. To this aim, we calculated $\theta_W$ (an estimate of the expected per site heterozygosity) (Watterson 1975) and $\pi$ (the average number of pairwise sequence nucleotide differences between haplotypes) (Nei and Li 1979) in individuals with European ancestry (CEU), in Yoruba subjects from Nigeria (YRI), and in East Asian subjects (AS; Japanese plus Chinese). We then compared the values with the distribution obtained from 2000 randomly selected genes (see Methods for details).

As reported in Supplementary Table S3, only TMPRSS2 and TMPRSS4 showed evidence of high nucleotide diversity in all populations, although the rank of $\theta_W$ and $\pi$ did not reach the 95th percentile in all of them.

High nucleotide diversity might be suggestive of balancing selection. Yet, as a result of mutations and recombination, balancing selection signatures tend to extend over relatively short genomic intervals (Wiuf et al. 2004; Charlesworth. 2006). Thus, in order to map the region(s) responsible for high nucleotide diversity in TMPRSS2 and TMPRSS4, we calculated $\theta_W$ and $\pi$ in 5 kb sliding-windows moving along these genes. In order to analyze potential regulatory regions, we extended the analysis to 20 kb upstream and downstream each gene. Overall, the same trend of $\theta_W$ and $\pi$ was evident in all populations for both genes (Figs. 3, 4). In TMPRSS2, sliding-windows analysis highlighted 3 regions with high nucleotide diversity in at least two populations (Fig. 3 and Table 2): one located in the 3′ region of TMPRSS2 (TMPRSS2-reg1) and falling in the MX1 gene, one in TMPRSS2 intron 8 (TMPRSS2-reg2), and one covering the promoter region (TMPRSS2-reg3). In the case of TMPRSS4, only two intragenic regions were identified (Fig. 4 and Table 2): one covering the 2 kb surrounding
exon 2 (TMPRSS4-reg1) and one covering the entire 3′UTR (TMPRSS4-reg2).

An effect of balancing selection is a distortion of the site frequency spectrum (SFS) toward intermediate frequency alleles. Deviation in the SFS is identified by neutrality tests, including Tajima’s $D$ (DT) (Tajima 1989) and Fu and Li’s $D^*$ and $F^*$ (Fu and Li 1993). DT tests the departure from neutrality by comparing $\theta_W$ and $\pi$, and positive values indicate an excess of intermediate frequency variants. Fu and Li’s $F^*$ and $D^*$ are also based on SNP frequency spectra and differ from DT in that they also take into account whether mutations occur in external or internal branches of a genealogy. For TMPRSS2-reg1 and TMPRSS2-reg2, tests indicated departure from neutrality with significantly
positive values for most statistics in all populations (Table 2). Conversely, for TMPRSS2-reg3, neutrality tests showed more significant values in the CEU population. With respect to TMPRSS4, significantly positive values for most statistics were observed only in the region covering the 3′-UTR (Table 2). Overall, these analyses indicate that
both \textit{TMPRSS2} and \textit{TMPRSS4} may be targets of ongoing balancing selection in human populations.

The action of natural selection obviously implies the presence of a functional variant with an effect on fitness and a selective pressure to act on it. Recently, variants in \textit{TMPRSS2} with a potential functional role and/or associated to viral infection susceptibility have been identified (Asselta et al. 2020; Irham et al. 2020; Piva et al. 2021; Vargas-Alarcón et al. 2020). Specifically, two variants in \textit{TMPRSS2}-reg2 (rs55964536 and rs7364083) define two distinct haplotypes with different frequencies among populations and were predicted to modulate \textit{TMPRSS2} expression (Asselta et al. 2020) (Fig. 3). The rs55964536 polymorphism was reported as an eQTL in lung tissue (genotype-tissue expression, GTEx, V6 data release).
expression, GTEx) (GTEx Consortium et al. 2017) and the haplotype to which rs55964536 belongs is functionally linked to another eQTL variant (rs8134378) located at a known androgen-responsive enhancer for TMPRSS2 (Asselta et al. 2020). Based on these observations, Asselta and colleagues suggested that upregulation of TMPRSS2 expression level through androgens could explain the sex bias in COVID-19 severity. On the other hand, rs7364083 belongs to the second haplotype. This latter is also defined by a variant (rs2070788) associated with increased gene expression, a higher risk of severe disease with the 2009 H1N1 pandemic influenza, and increased susceptibility to human H7N9 flu (Cheng et al. 2015).

In TMPRSS2-reg2, another variant (rs2298661) in association with both COVID-19 disease susceptibility and with increased risk of worse disease outcomes was identified through a trans-ancestry meta-analysis (Horowitz et al. 2021). The rs2298661 polymorphism is in tight linkage disequilibrium (LD \( R^2 > 0.8 \)) with a non-synonymous variant (rs12329760, p.V160M). V160M falls in an exonic splicing enhancer site (Srp40) and is also predicted to decrease protein stability in structural models, with a plausible role in viral entry (Vishnubhotla et al. 2020).

Taken together, these evidences suggest that the balancing selection signal in TMPRSS2-reg2 could be secondary to the presence of variants with a potential functional effect, most likely on gene expression regulation. Overall, these variants with different allele frequencies in different populations could play an important role in viral entry, helping to explain the heterogeneity observed in different ethnic groups with regard to incidence, severity, progression, and therapeutic response to COVID-19 (Alshahawey et al. 2020).

As reported in Fig. 3, other eQTL variants fall in TMPRSS2-reg1 and -reg3. In particular, a cluster of eQTLs modulating expression levels of TMPRSS2 and MX1 in lung and thyroid falls in TMPRSS2-reg1, whereas whole-blood MX2 eQTLs are present in TMPRSS2-reg3. MX1 and MX2 are involved in antiviral response. In particular, MX1 can restrict a wide variety of viruses, including orthomyxoviruses (e.g., influenza and Thogoto), bunyaviruses (e.g., La Crosse and rift valley viruses), and hepatitis B virus (Verhelst et al. 2013). Conversely, MX2 is known to counteract retroviruses (Kane et al. 2013), and to exert a weak effect against vesicular stomatitis virus infection (Liu et al. 2012). Moreover, MX1 expression is higher in COVID-19 patients compared to non-infected subjects and tends to decrease with age (Bizzotto et al. 2020). Variants in TMPRSS2-reg1 might thus contribute to modulate both TMPRSS2 and MX1 expression, acting on the ability to infect and to trigger an effective antiviral response.

As stated above, TMPRSS2-reg1 falls within the MX1 gene, hence deviation from neutrality with significantly positive values for most SFS statistics in human populations may be attributable to a specific selection signal for the MX1 gene rather than TMPRSS2. However, within this region, there are eQTL variants for TMPRSS2; these variants may act to determine a different expression of the protease and,

---

**Table 2** Nucleotide diversity and neutrality tests for the TMPRSS2 and TMPRSS4 gene regions

| Gene (chromosomal position) | Population | \( S^a \) | \( \pi \times 10^{-5} \) Value Rank \( b \) | \( \theta_W \times 10^{-4} \) Value Rank \( b \) | Tajima’s \( D \) Value Rank \( b \) | Fu and Li’s \( D^* \) Value Rank \( b \) | Fu and Li’s \( F^* \) Value Rank \( b \) |
|-----------------------------|------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| TMPRSS2-reg1 (chr21:42,818,732–42,822,732) | CEU | 38 | 29.96 99 | 16.57 99 | 2.38 97 | 1.5 97 | 2.25 99 |
| | YRI | 58 | 38.95 99 | 25.24 99 | 1.65 99 | 0.98 80 | 1.53 97 |
| | AS | 33 | 29.12 99 | 14.12 99 | 3.05 99 | 2.06 99 | 2.98 99 |
| TMPRSS2-reg2 (chr21:42,843,982–42,849,232) | CEU | 86 | 50.54 99 | 28.57 99 | 2.39 98 | 1.55 97 | 2.32 99 |
| | YRI | 99 | 54.68 99 | 32.82 99 | 2.08 99 | 1.88 99 | 2.37 99 |
| | AS | 78 | 47.47 99 | 25.43 99 | 2.66 98 | 1.23 93 | 2.25 99 |
| TMPRSS2-reg3 (chr21:42,886,982–42,888,232) | CEU | 9 | 27.71 99 | 12.56 96 | 2.79 99 | 1.28 93 | 2.17 99 |
| | YRI | 10 | 25.3 98 | 13.93 83 | 1.94 99 | -0.19 36 | 0.69 81 |
| | AS | 10 | 28.52 99 | 13.7 98 | 2.54 97 | 0.56 77 | 1.52 93 |
| TMPRSS4-reg1 (chr11:117,964,293–117,966,293) | CEU | 21 | 20.03 98 | 18.31 99 | 0.26 56 | -0.15 48 | 0.01 51 |
| | YRI | 21 | 17.41 92 | 18.28 95 | -0.13 67 | 0.81 74 | 0.53 75 |
| | AS | 21 | 24.05 99 | 17.97 99 | 0.91 73 | 0.3 69 | 0.66 74 |
| TMPRSS4-reg2 (chr11:117,988,543–117,992,043) | CEU | 31 | 23.79 99 | 15.45 99 | 1.55 89 | 1.65 98 | 1.93 97 |
| | YRI | 31 | 25.76 98 | 15.42 89 | 1.93 99 | 1.65 97 | 2.12 99 |
| | AS | 29 | 27.32 99 | 14.18 99 | 2.62 98 | 0.39 73 | 1.58 94 |

Significant values are in bold

\( ^a \)Number of segregating sites

\( ^b \)Percentile rank relative to a distribution of 5-kb regions from 2000 random genes
consequently, a different ability to activate the viral proteins that are TMPRSS2 protease targets.

With respect to TMPRSS4, information on its regulation is more limited. However, eQTLs are present in TMPRSS4-reg2. Specifically, different eQTLs for the esophageous mucosa map to this genomic region.

In summary, high nucleotide diversity and a shift in the SFS toward intermediate-frequency alleles suggest that the natural selection signals in these TMPRSS2 and TMPRSS4 regions are driven by the presence of polymorphic variants with a functional role that could contribute to infection with and spread of viruses, and/or to the modulation of disease clinical course.

Conclusions

Type II transmembrane serine proteases (TTSPs) are involved in the proteolitic activation of different viruses. In particular, the cleavage of surface glycoproteins by some proteases has been reported for influenza viruses and coronaviruses, but also for other respiratory viruses. These observations led us to hypothesize that the members of this family may be targets of virus-driven selective pressure. It is known that natural selection events acted on the receptors used by viruses to promote their entry into the cell during the early stages of infection. From this point of view, coronavirus receptors, such as ACE2, DPP4, and ANPEP have been extensively studied. However, until now, systematic evolutionary studies had never been carried out on TTSP gene family.

In this study, the application of specific molecular evolution tests showed that 7 TTSP genes evolved adaptively in mammals. Six of these are involved in the proteolytic activation of different viral proteins, suggesting that host–pathogen arms races drive the evolution of these genes.

Since more proteases are able to prime proteins of different viruses, we may speculate that they play an important role in determining viral tropism and disease progression, explaining, at least partially, the apparent redundancy of the function of these proteases. Regarding the positively selected sites, we observed that all the TTSP genes have at least one site in the protease domain, that these sites are not in close proximity to the catalytic domain and that in general the selected sites are distributed along the gene. Taken together, these results suggest that different infectious agents may have exerted distinct selection pressures on the same gene and that sites targeted by positive selection in the protease and in other domains could participate in the formation of bonds or in stabilizing the reciprocal position of the host and viral proteins.

Of particular interest, especially at the present time, is TMPRSS2. This protease represents the canonical activator of SARS-CoV-2, although its role in priming has been confirmed for many other respiratory viruses.

Notably, TMPRSS2 presents pervasive positive selection signals in the mammalian phylogeny and could be target of ongoing balancing selection in human populations. In the latter case, the signals of natural selection could be driven by the presence of polymorphic variants with a functional role that may contribute to the infection with and the spread of viruses, and/or to the modulation of the clinical course of the disease.

As the world continues to wrestle with the COVID-19 pandemic, research on the virus aims to develop therapies to control the health and socio-economic emergency.

Different vaccine platforms are being used for vaccine development (Kyriakidis et al. 2021), most of them aiming to induce neutralizing antibodies against the spike protein (S). Some of these are already used in vaccination campaigns worldwide, others are close to receiving approval or authorization for large-scale vaccinations. However, the emergence of new lineages of SARS-CoV-2 carrying variants in the spike protein raises concerns about a reduction in the efficacy of the vaccines currently in use. Furthermore, vaccination during the pandemic could exert selective pressures that might favor the expansion of vaccine-resistant SARS-CoV-2 lineages. In this context, it is important to combine vaccinations with alternative therapies aimed at impairing the virus capacity to spread through the host cells. Thus, several studies have identified serine proteases as potential drug targets for the treatment of SARS-CoV-2 infection (Baughn et al. 2020; Qiao et al. 2020; Tang et al. 2020; Zarubin et al. 2020; Deng et al. 2021).

TMPRSS2 is a well-characterized TTSP and represents one of the essential host factors in viral activation and cell entry of influenza viruses, such as pandemic influenza H1N1, and different coronaviruses, including SARS-CoV-2. TMPRSS2 is dispensable for development and homeostasis in mice (Kim et al. 2006). Thus, this protease is potentially one of the most promising drug targets for COVID-19 therapy. Therapeutics targeting TMPRSS2 expression and/or activity may be beneficial not only for coronavirus-infected patients but also for those infected with influenza viruses. Indeed, promising data have been obtained from the use of TMPRSS2 inhibitors such as camostat and nafamostat mesylate, which seem to have good efficacy against coronaviruses (Hoffmann et al. 2020b, c). The design of selective inhibitors will be facilitated by a better understanding of functional genetic diversity, both in humans and in other mammals that are used as model organisms for drug development.

Data reported in this study indicate that, in analogy to other genes with a role in viral infections, TTSPs have evolved adaptively in mammals (Sironi et al. 2015). In particular a continuum of selective pressure acting on TMPRSS2
and \textit{TMPRSS4} was observed as these genes also represent selection targets in human populations. This scenario may be the result of an ancestral and still ongoing host–pathogen arms race.

Overall, our results provide evolutionary information about candidate functional sites that may shape infection outcomes across mammals and within humans.

\textbf{Supplementary Information} The online version contains supplementary material available at \url{https://doi.org/10.1007/s00439-022-02435-y}.

\textbf{Author contributions} Conceptualization, RC and MS; formal analysis, RC, DF; and MS; investigation, RC; visualization, RC and DF; writing—original draft, RC, MS and DF; supervision, RC and MS. All authors reviewed the manuscript.

\textbf{Funding} This work was supported by the Italian Ministry of Health (“Ricerca Corrente 2021” to MS).

\textbf{Declarations}

\textbf{Conflict of interest} The authors declare no competing interests.

\textbf{References}

1000 Genomes Project Consortium, Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation across population-scale sequencing. Nature 467:1061–1073. \url{https://doi.org/10.1038/nature09534}.

Alshahawey M, Raslan M, Sabri N (2020) Sex-mediated effects of ACE2 and TMPRSS2 on the incidence and severity of COVID-19: the need for genetic implementation. Curr Res Transl Med 68:149–150 (\textit{S2452-3186(20)30054-4}).

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958.

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236.

Antalis TM, Buzaa MS, Hodge KM, Hooper JD, Netzel-Arnett S (2010) The cutting edge: membrane-anchored serine protease activities in the pericellular microenvironment. Biochem J 428:325–346. \url{https://doi.org/10.1042/BJ20100046}.

Asselta R, Paraboschi EM, Mantovani A, Duga S (2020) ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. Aging (abing NY) 12:10087–10098. \url{https://doi.org/10.18632/aging.103415}.

Barrett CT, Dutch RE (2020) Viral membrane fusion and the transmembrane domain. Viruses 12:693. \url{https://doi.org/10.3390/v12070693}.

Baughn LB, Sharma N, Elhai E, Sekulic A, Bryce AH, Fonseca R (2020) Targeting TMPRSS2 in SARS-CoV-2 infection. Mayo Clin Proc 95:1989–1999 (\textit{S0025-6196(20)30626-1}).

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol 16:363–377. \url{https://doi.org/10.1046/j.1420-9110.2003.00531.x}.

Bertram S, Glowacka I, Blazewiejka P, Soilleux E, Allen P, Danisch S, Steffen I, Choi SY, Park Y, Schneider H, Schughart K, Pöhllmann S (2010) TMPRSS2 and TMPRSS4 facilitate trypsin-independent spread of influenza virus in Caco-2 cells. J Virol 84:10016–10025. \url{https://doi.org/10.1128/JVI.00239-10}.

Bertram S, Glowacka I, Müller MA, Lavender H, Gnirs K, Nehlmeier I, Niemeyer D, He Y, Simmons G, Drosten C, Soilleux EJ, Jahn O, Steffen I, Pöhllmann S (2011) Cleavage and activation of the severe acute respiratory syndrome coronavirus spike protein by human airway trypsin-like protease. J Virol 85:13363–13372. \url{https://doi.org/10.1128/JVI.05300-11}.

Bertram S, Dijkman R, Habjan M, Heurich A, Gierer S, Glowacka I, Welsch K, Winkler M, Schneider H, Hofmann-Winkler H, Thiel V, Pöhllmann S (2013) TMPRSS2 activates the human coronavirus 229E for cathepsin-independent host cell entry and is expressed in viral target cells in the respiratory epithelium. J Virol 87:6150–6160. \url{https://doi.org/10.1128/JVI.03372-12}.

Bibiana SOF, Vargas-Pinilla P, Amorim CEG, Sortica VA, Bortolini MC (2020) ACE2 diversity in placental mammals reveals the evolutionary strategy of SARS-CoV-2. Genet Mol Biol 43:e20200104. \url{https://doi.org/10.1590/1678-4685-GMB-2020-0104}.

Bizzotto J, Sanchis P, Abbate M, Lager-Vickers S, Lavignolle R, Toro A, Olsevicki S, Sabater A, Cascardo F, Vazquez E, Cotigolina J, Gueron G (2020) SARS-CoV-2 infection boosts MX1 antiviral effector in COVID-19 patients. iScience 23:101585. \url{https://doi.org/10.1016/j.isci.2020.101585}.

Böttcher-Friebertshäuser E (2018) Membrane-anchored serine proteases: host cell factors in proteolytic activation of viral glycoproteins. In: Böttcher-Friebertshäuser E, Garten W, Klenk HD (eds) Activation of Viruses by Host Proteases. Springer, Switzerland, pp 153–203.

Bugge TH, Antalis TM, Wu Q (2009) Type II transmembrane serine proteases: host cell factors in proteolytic activation of viral glycoproteins. In: Böttcher-Friebertshäuser E, Garten W, Klenk HD (eds) Activation of Viruses by Host Proteases. Springer, Switzerland, pp 153–203.

Bujalowski M, Bielawski JP, Yang Z (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950–958.

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Menozzi G, Bresolin N, Sironi M (2008) The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biol 9:R143. \url{https://doi.org/10.1186/gb-2008-9-9-r143}.

Cagliani R, Fumagalli M, Biasin M, Picentini L, Riva S, Pozzoli U, Bonaglia MC, Bresolin N, Clerici M, Sironi M (2010) Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. Hum Genet. \url{https://doi.org/10.1007/s00439-010-0884-6}.

Chan JC, Knudson O, Wu F, Morser J, Dole WP, Wu Q (2005) Hypertension in mice lacking the proenatriuretic peptide convertase corin. Proc Natl Acad Sci USA 102:785–790. \url{https://doi.org/10.1073/pnas.0407234102}.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64. \url{https://doi.org/10.1371/journal.pgen.0020064}.

Cheng Z, Zhou J, To KK, Chu H, Li C, Wang D, Yang D, Zheng S, Hao K, Bossé Y, Obeidat M, Brandsma CA, Song YQ, Chen Y, Zheng BJ, Li L, Yuen KY (2015) Identification of TMPRSS2 as a susceptibility gene for severe 2009 pandemic A(H1N1) influenza. J Infect Dis 212:1214–1221. \url{https://doi.org/10.1093/infdis/jiv246}.

Choi SY, Bertram S, Glowacka I, Park YW, Pöhllmann S (2009) Type II transmembrane serine proteases in cancer and viral infections. Trends Mol Med 15:303–312. \url{https://doi.org/10.1016/j.molmed.2009.05.003}.

Cui J, Eden JS, Holmes EC, Wang LF (2013) Adaptive evolution of bat dipeptidyl peptidase 4 (dpp4): implications for the origin and emergence of Middle East respiratory syndrome coronavirus. Virol J 10:304:422X-10–304. \url{https://doi.org/10.1186/1743-422X-10-304}.

Damas J, Hughes GM, Keough KC, Painter CA, Persky NS, Corbo M, Hiller M, Koepli KP, Pfening AR, Zhao H, Genereux DP, Swofford R, Pollard KS, Ryder OA, Nweeia MT, Lindblad-Toh
Forni D, Cagliani R, Pozzoli U, Colleoni M, Riva S, Biasin M, Filippi
Enard D, Cai L, Gwennap C, Petrov DA (2016) Viruses are a dominant
Du X, She E, Gelbart T, Truksa J, Lee P, Xia Y, Khovananth K, Mudd S, Mann N, MoreSCO EM, Beutler E, Beutler B (2008) The serine protease TMPRSS6 is required to sense iron defi-
Science 320:1088–1092. https://doi.org/10.1126/science.1157121
Enard D, Cai L, Gwennap C, Petrov DA (2016) Viruses are a dominant
Forni D, Cagliani R, Clerici M, Sironi M (2017) Molecular evolu-
1016/j.immunol.2013.04.008
Forni D, CaglJani R, Pozzoli U, De Gioia L, Filippi G, Riva S, Menozzi G, Colleoni M, Biasin M, Lo Caputo S, Mazzotta F, Comi GP, Bresolin N, Clerici M, Sironi M (2013) A 175 million year history of T cell regulatory molec-
Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R (2011) Signatures of environmental genetic adapta-
Gramberg T, Hofmann H, Möller P, Lalor PF, Marzi A, Geier M, Krumbiegel M, Pöhlmann S (2020c) Nafamostat mesylate blocks activation of the serine protease SARS-CoV-2 is essen-
Guindon S, Delsuc F, Dufayard JF, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol 537:113–137. https://doi.org/10.1007/978-1-59745-251-9_6
Guo H, Hu BJ, Yang XL, Zeng LP, Li B, Ouyang S, Shi ZL (2020) Evolutionary arms race between virus and host drives genetic diversity in bat severe acute respiratory syndrome-related coron-
Haldane JBS (1932) The causes of evolution. Harper & Row, New York
Harrison SC (2008) Viral membrane fusion. Nat Struct Mol Biol 15:690–698. https://doi.org/10.1038/nsmb.1456
Hayashi H, Kubo Y, Izumida M, Takahashi E, Kido H, Sato K, Yamaya M, Nishimura H, Nakayama K, Matsuyama T (2018) Enterokinase enhances influenza A virus infection by activat-
He M, Zhang Y, Li H, Liu M, Dong N, Wu Q (2020) A common CORIN variant in hypertension reduces corin intracellular trafficking by exposing an inhibitory N-terminus. Biochem Biophys Res Commun 530:35–41
Hoffmann M, Kleine-Weber H, Pöhlmann S (2020a) A Multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Mol Cell 78:779–784.e5 (S1097-2765(20)30264-1)
Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu NH, Nitsche A, Müller MA, Drosten C, Pöhlmann S (2020b) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181:271-280.e8 (S0092-8674(20)30229-4)
Hoffmann M, Schroeder S, Kleine-Weber H, Müller MA, Drosten C, Pöhlmann S (2020c) Nafamostat mesylate blocks activation of SARS-CoV-2: new treatment option for COVID-19. Antimicrob Agents Chemother 64:e00754-e020. https://doi.org/10.1128/AAC.00754-20 (Print 2020 May 21)
Hollox EJ, Armour JA (2008) Directional and balancing selection in human beta-defensins. BMC Evol Biol 8:113. https://doi.org/10.1186/1471-2148-8-113
Hooper JD, Clements JA, Quigley JP, Antalis TM (2001) Type II transmembrane serine proteases. Insights into an emerg-
Hou Y, Zhao J, Martin W, Kallianpur A, Chung MK, Jehi L, Sharifi N, Erzurum S, Eng C, Cheng F (2020) New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2
Sasaki M, Itakura Y, Kishimoto M, Tabata K, Uemura K, Ito N, Sugiyama M, Wastika CE, Orba Y, Sawa H (2021) Host serine proteases TMPRSS2 and TMPRSS11D mediate proteolytic activation and trypsin-independent infection in group A rotaviruses. J Virol 95:e00398-e241. https://doi.org/10.1128/JVI.00398-21

Schrödinger L (2017) The PyMOL molecular graphics system, version 2.0 Schrödinger, LLC. Google Scholar there is no corresponding record for this reference

Scott HS, Kudoh J, Wattenhofer M, Shibuya K, Berry A, Chrast R, Guipponi M, Wang J, Kawasaki K, Asakawa S, Minoshima S, Younus F, Mehdi SQ, Radhakrishna U, Papasavvas MP, Gehrig C, Rossier C, Korostischovsky M, Gal A, Shimitzu N, Bonne-Tamir B, Antonarakis SE (2001) Insertion of beta-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness. Nat Genet 27:59–63. https://doi.org/10.1038/83768

Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F (2020) Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci USA 117:11727–11734. https://doi.org/10.1073/pnas.2003138117

Shirato K, Kawasaki M, Matsuyama S (2013) Middle East respiratory syndrome coronavirus infection mediated by the transmembrane serine protease TMPRSS2. J Virol 87:12552–12561. https://doi.org/10.1128/JVI.01890-13

Shirogane Y, Takeda M, Iwasaki M, Ishiguro N, Takeuchi H, Nakatsu Y, Tahara M, Kikuta H, Yanagi Y (2008) Efficient multiplication of human metapneumovirus in Vero cells expressing the transmembrane serine protease TMPRSS2. J Virol 82:8942–8946. https://doi.org/10.1128/JVI.00676-08

Simmons G, Gosalia DN, Rennekamp AJ, Reeves JD, Diamond SL, Bates P (2005) Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry. Proc Natl Acad Sci USA 102:11876–11881 (20050551)

Siroiri C, Kawase M, Matsuyama S (2013) Middle East respiratory syndrome coronavirus infection mediated by the transmembrane serine protease TMPRSS2. J Virol 87:12552–12561. https://doi.org/10.1128/JVI.01890-13

Siroti M, Cagliani R, Forni D, Clerici M (2015) Evolutionary insights into host–pathogen interactions from mammalian sequence data. Nat Rev Genet 16:224–236. https://doi.org/10.1038/nrg3905

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosaikovsky Pond SL (2015) Less is more: an adaptive branch-site model of elevated adaptive evolution in viral receptors. J Virol 94:e00684-e720. https://doi.org/10.1128/JVI.00684-20 (Print 2020 Jul 30)

Sobrosa R, Bugge TH (2008) Type II transmembrane serine proteases in development and disease. Int J Biochem Cell Biol 40:1297–1316. https://doi.org/10.1016/j.biocel.2007.11.013

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tang T, Bidon M, Jaimes JA, Whittaker GR, Daniel S (2020) Coronavirus membrane fusion mechanism offers a potential target for antiviral development. Antivir Res 178:104792 https://doi.org/10.1016/j.antiviral.2020.104792

Tarnow C, Engels G, Arendt A, Schwalm F, Sediri H, Preuss A, Nelson PS, Garten W, Klenk HD, Gabriel G, Böttcher-Friebertshäuser E (2014) TMPRSS2 is a host factor that is essential for pneumovirus and pathogenicity of H7N9 influenza A virus in mice. J Virol 88:4744–4751. https://doi.org/10.1128/JVI.03799-13

Villega AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblComparas GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19:327–335. https://doi.org/10.1101/gr.073585.107; 10.1101/gr.073585.107

Vishnubhota R, Vankadari N, Ketavarapu V, Amanchry R, Avanthi S, Bale G, Reddy DN, Sasikala M (2020) Genetic variants in TMPRSS2 and Structure of SARS-CoV-2 spike glycoprotein and TMPRSS2 complex. BioRxiv. https://doi.org/10.1101/2020.06.30.179663

Wang K, Chen W, Zhang Z, Deng Y, Lian QJ, Du P, Wei D, Zhang Y, Sun XX, Gong L, Yang X, He L, Zhang L, Yang Z, Geng JJ, Chen R, Zhang H, Wang B, Zhu YM, Nan G, Jiang JL, Li L, Wu J, Lin P, Huang W, Xie L, Zheng ZH, Zhang K, Miao JL, Cui HY, Huang M, Zhang J, Fu L, Yang XM, Zhao Z, Sun S, Gu H, Wang Z, Wang CF, Lu Y, Liu YY, Wang QY, Bian H, Zhu P, Chen ZN (2020a) CD147-spike protein is a novel route for SARS-CoV-2 infection to host cells. Signal Transduct Target Ther 5:283. https://doi.org/10.1038/s41392-020-00426-x

Wang W, Zhao H, Han GZ (2020b) Host–virus arms races drive elevated adaptive evolution in viral receptors. J Virol 94:e00684-e0720. https://doi.org/10.1128/JVI.00684-20 (Print 2020 Jul 30)

Watterson GA (1975) On the number of segregating sites in genetic models without recombination. Theor Popul Biol 7:256–276

Wernersson R, Pedersen AG (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 31:3537–3539

Wuif C, Zhao K, Innan H, Nordborg M (2004) The probability and chromosomal extent of trans-specific polymorphism. Genetics 168:2363–2372. https://doi.org/10.1034/j.1171-1670.02.10.1073pnas.2003138117

Wang W, Zhao H, Han GZ (2020b) Host–virus arms races drive elevated adaptive evolution in viral receptors. J Virol 94:e00684-e720. https://doi.org/10.1128/JVI.00684-20 (Print 2020 Jul 30)

Watterson GA (1975) On the number of segregating sites in genetic models without recombination. Theor Popul Biol 7:256–276

Wernersson R, Pedersen AG (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 31:3537–3539

Yuan W, Wu F, Morser J, Wu Q (2000) Corin, a transmembrane carboxyl terminal serine protease, acts as a pro-atrial natriuretic peptide-converting enzyme. Proc Natl Acad Sci USA 97:8525–8529 (150149907)

Yang Z (2019) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088

Yang ZY, Huang Y, Ganesh L, Leung K, Kong WP, Schwartz O, Subbarao K, Nabel GJ (2004) pH-dependent entry of severe acute respiratory syndrome coronavirus is mediated by the spike glycoprotein and enhanced by dendritic cell transfer through DC-SIGN. J Virol 78:5642–5650 (78/11/5642)

Yeager CL, Ashmun RA, Williams RK, Cardellichio CB, Shapiro LH, Look AT, Holmes KV (1992) Human aminopeptidase N is a receptor for human coronavirus 229E. Nature 357:420–422 (BF357420a)

Zang R, Gomez Castro MF, McCune BT, Zeng Q, Rothlauf PW, Sonnek NM, Liu Z, Bruoloi KW, Wang X, Greenberg HB, Diamond MS, Ciorba MA, Whelan SPI, Ding S (2020) TMPRSS2 and TMPRSS4 promote SARS-CoV-2 infection of human small intestinal enterocytes. Sci Immunol 5:eabc582. https://doi.org/10.1126/sciimmunol.abc1582

Zarubin A, Stepanov V, Markov A, Kolesnikov N, Marusin A, Khitrinnaya I, Savrovskaya M, Litvinov S, Ekomasova N, Dzhabaermanov M, Maksimova N, Sukhomyasova A, Shytgasycheva O, Khussanitdinova E, Radzhahov M, Kharkov V (2020) Structural variability, expression profile, and pharmacogenetic properties of TMPRSS2 gene as a potential target for COVID-19 therapy. Genes (Basel) 12:19. https://doi.org/10.3390/genes12010019

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479. https://doi.org/10.1093/molbev/msi237

Zhang H, Mo X, Qian Q, Zhou Z, Zhu Z, HuangFu X, Xu T, Wang A, Guo Z, Lei S, Zhang Y (2019a) Associations between potentially...
functional CORIN SNPs and serum corin levels in the Chinese Han population. BMC Genet 20:99-019-0802–4. https://doi.org/10.1186/s12863-019-0802-4

Zhang Z, Zhu Z, Chen W, Cai Z, Xu B, Tan Z, Wu A, Ge X, Guo X, Tan Z, Xia Z, Zhu H, Jiang T, Peng Y (2019b) Cell membrane proteins with high N-glycosylation, high expression and multiple interaction partners are preferred by mammalian viruses as receptors. Bioinformatics 35:723–728. https://doi.org/10.1093/bioinformatics/bty694

Zhang C, Zhang Y, Zhang S, Wang Z, Sun S, Liu M, Chen Y, Dong N, Wu Q (2020) Intracellular autoactivation of TMPRSS11A, an airway epithelial transmembrane serine protease. J Biol Chem 295:12686–12696. https://doi.org/10.1074/jbc.RA120.014525

Zhou Y, Wu Q (2014) Corin in natriuretic peptide processing and hypertension. Curr Hypertens Rep 16:415-013-0412–7. https://doi.org/10.1007/s11906-013-0415-7

Zmora P, Hoffmann M, Kollmus H, Moldenhauer AS, Danov O, Braun A, Winkler M, Schughart K, Pöhlmann S (2018) TMPRSS11A activates the influenza A virus hemagglutinin and the MERS coronavirus spike protein and is insensitive against blockade by HAI-1. J Biol Chem 293:13863–13873. https://doi.org/10.1074/jbc.RA118.001273

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