Extensive positive selection drives the evolution of non-structural proteins in lineage C Betacoronaviruses

Diego Forni*, Rachele Cagliani, Alessandra Mozzi, Uberto Pozzoli, Nasser Al-Daghri, Mario Clerici, Manuela Sironi

a Scientific Institute IRCCS E. MEDEA, Bioinformatics, Bosisio Parini, Italy; b Biomarker research program, Biochemistry Department, College of Science, King Saud University, Riyadh, Kingdom of Saudi Arabia (KSA); c Prince Mutaib Chair for Biomarkers of Osteoporosis, Biochemistry Department, College of science, King Saud University, Riyadh, KSA; d Department of Physiopathology and Transplantation, University of Milan, Milan, Italy; e Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy.

Running Head: Positive Selection at MERS-CoV non-structural proteins

*Address for correspondence to Diego Forni, diego.forni@bp.lnf.it

Abstract word count: 249

Main text word count: 4998
Abstract

Middle East Respiratory Syndrome-related Coronavirus (MERS-CoV) spread to humans via the zoonotic transmission from camels. MERS-CoV belongs to the lineage C of Betacoronaviruses (betaCoVs), which also includes viruses isolated from bats and hedgehogs. A large portion of the betaCoV genome consists of two open reading frames (ORF1a and ORF1b) that are translated into polyproteins. These are cleaved by viral proteases to generate 16 non-structural proteins (nsp1-16) which compose the viral replication-transcription complex. We investigated the evolution of ORF1a and ORF1b in lineage C betaCoVs. Results indicated widespread positive selection, acting mostly on ORF1a. The proportion of positively selected sites in ORF1a was much higher than that previously reported for the surface-exposed spike protein. Selected sites were unevenly distributed, with nsp3 representing the preferential target. Several pairs of co-evolving sites were also detected, possibly indicating epistatic interactions; most of these were located in nsp3. Adaptive evolution at nsp3 is ongoing in MERS-CoV strains and two selected sites (G720 and R911) were detected in the protease domain. Whereas position 720 is variable in camel-derived viruses, suggesting that the selective event does not represent a specific adaptation to humans, the R911C substitution was only observed in human-derived MERS-CoV isolates, including the viral strain responsible for the recent South Korean outbreak. It will be extremely important to assess whether these changes affect host range or other viral phenotypes. More generally, data herein indicate that CoV nsp3 represents a major selection target and nsp3 sequencing should be envisaged in monitoring programs and field surveys.
Both SARS-CoV and MERS-CoV originated in bats and spread to humans via an intermediate host. This clearly highlights the potential for coronavirus host shifting and the relevance of understanding the molecular events underlying the adaptation to new host species. We investigated the evolution of ORF1a and ORF1b in lineage C betaCoVs and in 87 sequenced MERS-CoV isolates. Results indicated widespread positive selection, stronger in ORF1a than in ORF1b. Several selected sites were found to be located in functionally relevant protein regions and some of them corresponded to functional mutations in other coronaviruses. The proportion of selected sites we identified in ORF1a is much higher than that for the surface-exposed spike protein. This observation suggests that adaptive evolution in ORF1a might contribute to host shifts or immune evasion. Data herein also indicate that genetic diversity at non-structural proteins should be taken into account when antiviral compounds are developed.
Introduction

Middle East respiratory syndrome-related coronavirus (MERS-CoV, http://talk.ictvonline.org/files/proposals/animal_ssrna_viruses/default.aspx) was identified as the causative agent of a new viral respiratory disease in Saudi Arabia in June 2012 (1). Since then, more than 1,500 cases and 571 deaths have been reported worldwide (as of October 12th, 2015 http://www.who.int/csr/don/12-october-2015-mers-saudi-arabia/en/), although major outbreaks have been confined to the Middle East and, more recently, to South Korea. The rate of human-to-human transmission of MERS-CoV is relatively low, suggesting that a zoonotic reservoir serves as a major source for transmission (2). Recent studies have indicated that MERS-CoV or a closely related virus originated in bats and possibly spread to humans via transmission from dromedary camels (3).

Like SARS-CoV, which causes Severe Acute Respiratory Syndrome and evolved in bats as well, MERS-CoV (order Nidovirales, family Coronaviridae, subfamily Coronavirinae) is a positive-sense RNA (+ssRNA) virus belonging to the C lineage of the Betacoronavirus (betaCoVs) genus (4). CoVs are exceptional among RNA viruses for having long (~30-kb) genomes, a feature associated with a specific genome architecture and with the acquisition of an RNA 3′-to-5′ exoribonuclease activity (exoN) (5). About two thirds of the CoV genome consist of two large overlapping open reading frames (ORF1a and ORF1b), that are translated into the polyproteins pp1a and pp1ab (this latter synthesized via a −1 ribosome frameshift at the 3′ end of ORF1a). These polyproteins are subsequently cleaved by viral proteases to generate 16 non structural proteins (nsp1 to 16), most of which compose the viral replication-transcription complex (RTC). With the exception of nsp1 and 2, whose functions are poorly understood, most nsps have been characterized in some detail for either MERS-CoV, SARS-CoV or mouse hepatitis virus (MHV). Thus, nsp3, a large multidomain and multifunctional protein, was shown to play essential roles in the virus
replication cycle. In fact, the papain-like protease (PLpro) activity of nsp3 is responsible for the initial processing of pp1a. Also, nsp3 together with nsps4 and 6, recruits intracellular membranes to anchor the RTC and to form a reticulovesicular network of double-membrane vesicles (DVMs) and convoluted membranes where viral RNA replication occurs (6). nsp5 encodes a second viral protease (3C-like protease, 3CLpro) that cleaves both pp1a and pp1ab to the final nsp products. nsps7-11 provide the primer-making activity and regulate the function of the main RNA-dependent RNA polymerase (RdRp), this latter encoded by nsp12. Finally, nsp13-16 comprise RNA-modifying enzymes, including the exoN activity in nsp14 (7, 8).

Viral RNA represents the major pathogen-associated molecular pattern (PAMP) recognized by the host immune system during CoV infection (9). Both MERS-CoV and SARS-CoV elicit limited interferon (IFN) response in most cell types, indicating that these viruses have evolved efficient strategies to evade innate immune sensing and/or to block IFN induction (9). Indeed, these viruses express antagonists of the IFN response including SARS-CoV ORF6, ORF3b, and nucleoprotein, as well as MERS-CoV structural and accessory proteins M, ORF4a, ORF4b, and ORF5 (10-13). Additional immune evasion strategies, though, rely on nsps. In fact, the enzymatic activities of nsp14 and nsp16 endow the viral RNA of a 2′-O-methylated cap structure that mimics cellular mRNAs and avoids activation of the innate immunity receptors RIG-I and MDA5 (9). In analogy to the exoN and endoribonucleases expressed by other viruses such as Lassa Fever Virus and Pestivirus (14, 15), the ribonuclease activities of nsp14 and nsp15 are also thought to play a role in immune escape by digesting RNA PAMPs (9). Moreover, suppression of IFN responses is mediated by the PLpro in nsp3 through its deubiquitinating and deISGylating activities (16, 17), as well as by nsp1. The latter inhibits IFN-dependent signaling by decreasing the phosphorylation levels of STAT1 and suppresses host protein synthesis (18, 19). Finally, PLpro was shown to physically interact with TRAF3, TBK1, IKKe, STING, and IRF3, which represent key cellular components for
Therefore, the information encoded by CoV ORF1a and ORF1b is essential for viral replication and for immune evasion. For these reasons, inhibitors that interfere with viral enzymatic activities (e.g. proteases, RdRp) are regarded as promising candidates for therapeutic intervention (21).

From an evolutionary standpoint, different observations suggest that nsps may represent targets of natural selection. First, genes encoding molecules that directly interact with the host immune system are thought to be preferential targets of natural selection as a consequence of host-pathogen arms races (22). Second, adaptation to new hosts in other RNA viruses has been associated to selective changes in polymerase genes (23, 24). Finally, the acquisition of a complex replication machinery is evolutionary linked to genome expansion in Nidovirales (5). Nonetheless, evolutionary studies have mainly focused on the analysis of betaCoV spike proteins, as these are surface exposed and represent major determinants of host range and tissue tropism (25). Herein we investigated the evolution of ORF1a and ORF1b in MERS-CoV and in lineage C betaCoVs isolated from bats and hedgehogs. Results indicate widespread positive selection, stronger in ORF1a; within this region, nsp3 represents a preferential selection target and adaptive evolution is ongoing in MERS-CoV strains circulating in the current outbreak.

Material and Methods

Sequences and alignments
ORF1a/ORF1b sequences for 7 lineage C betaCoVs and 87 MERS-CoV strains (available as of July 2015) were retrieved from the NCBI database; a list of accession numbers for the complete genomes is provided in Table S1. Alignment errors are common when divergent sequences are analyzed and can affect evolutionary inference. Thus, we used PRANK (26) to generate multiple sequence alignments and GUIDANCE
(27) for filtering unreliably aligned codons (i.e. we masked codons with a score <0.90), as suggested (28).

The alignments were screened for the presence of recombination events using two methods based on distinct data features: 1) GARD (Genetic Algorithm Recombination Detection) (29) uses phylogenetic incongruence among segments in the alignment to detect the best-fit number and location of recombination breakpoints; the statistical significance of putative breakpoints is then evaluated through Kishino-Hasegawa (HK) tests; 2) GENECONV (30) tests for significant clustering of substitutions along sequences; statistical significance is assessed through permutation with multiple-comparison correction. For both methods recombination breakpoints were considered significant if the $p$ value was < 0.05. No breakpoint was detected in any analysis.

Detection of positive selection

Gene trees were generated by maximum-likelihood using the program phyML with a GTR plus gamma-distributed rates model and 4 substitution rate categories (31).

Positive selection can be defined when the nonsynonymous/synonymous rate ratio ($\omega$) is higher than 1; to analyze the presence of episodic positive selection in lineage C betaCoVs viruses we applied the branch-site test (32) from the PAML suite (33). The test is based on the comparison between two nested models: a model (MA) that allows positive selection on one or more lineages (called foreground lineages) and a model (MA1) that does not allow such positive selection. Twice the difference of likelihood for the two models ($\Delta \ln$) is then compared to a $\chi^2$ distribution with one degree of freedom (32). A false discovery rate correction was applied to take into account a multiple hypothesis issue generated by analyzing different branches on the same phylogeny (34). When the likelihood ratio test suggested the action of positive selection, the Bayes Empirical Bayes (BEB)
analysis was used to evaluate the posterior probability that each codon belongs to the site class of positive selection on the foreground branch.

BUSTED (branch-site unrestricted statistical test for episodic diversification) (35) is a recently developed software designed to describe episodic positive selection that is acting on specific branches in the phylogeny at a proportion of sites within the alignment. An alternative model that allows the action on positive selection on foreground branches is compared with a null model that doesn't allow \( \omega > 1 \). Twice the \( \Delta \ln \) of the two models is then compared to a \( \chi^2 \) distribution (degrees of freedom=2); if the null model is rejected, at least one site is under positive selection on the foreground branch(es). To detect selection at individual sites, twice the difference of the likelihood for the alternative and the null model at each site is compared to a \( \chi^2 \) distribution (degree of freedom=1). BUSTED is implemented in the HYPHY package (36).

Conservatively, we considered a site as selected if it showed a \( p \) value \( \leq 0.05 \) in BUSTED and a posterior probability \( \geq 0.90 \) in the BEB analysis.

The site models implemented in PAML were applied for the analysis of nsp3 sequences from MERS-CoVs isolates. To detect selection, two different pairs of nested site models (M1a/M2a and M7/M8) were fitted to the data (33); the M2a and M8 allow a class of sites to evolve with \( \omega > 1 \), whereas M1a and M7 do not. Positively selected sites were identified using the BEB analysis (from model M8) (37). Sites were validated using MEME (38) (with a cutoff \( \leq 0.1 \)), which allows the distribution of \( \omega \) to vary from site to site and from branch to branch at a site. MEME (38) analyses were performed through the DataMonkey server (39).

Detection of co-evolving sites

To detect co-evolving sites in the ORF1a and ORF1b alignments we applied two different methods:
BGM (Bayesian Graphical Model)-Spidermonkey (40) and the Mutual Information Server To Infer Coevolution (MISTIC) (41). Spidermonkey is a tool implemented in the HYPHY package that identifies co-evolving sites from an alignment of coding sequences; a BGM is used to evaluate the connection among codons (represented by the nodes of the network). Significant statistical associations between nodes are indicated by the edges of the network, suggesting functional or structural interactions between codons.

MISTIC estimates the relationship between two or more position in an alignment. The co-evolutionary association is estimated by Mutual Information (MI), that evaluates how much the information from the aminoacid at the first position can help to predict the aminoacid identity at the second position.

For BGM-Spidermonkey sites were filtered based on a minimum count of 4 substitutions across the phylogeny. To be conservative, we considered a pair of residues as co-evolving if they showed a posterior probability >0.75. This threshold corresponds to 0.02% and 1.42% of all analyzed site pairs in ORF1a and ORF1b, respectively. Likewise, for MISTIC site pairs were required to display a MI rank higher that the 99th percentile calculated using all MI scores from the alignment. Pairs of sites exceeding the thresholds for both methods were declared to be co-evolving.

Membrane topology, glycosylation site predictions, and 3D structure mapping

The membrane protein topology for MERS-CoV nsp3 and nsp4 was predicted by using TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) (42). N-Glycosylation sites were predicted with NetNGlyc (http://www.cbs.dtu.dk/services/NetNGlyc/), a program that uses artificial neural networks to examine the sequence context of Asn-X-Ser/Thr motifs.

Sites were mapped onto structures using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.2 Schrödinger, LLC).
Results

nsp3 in ORF1a is a major selection target in betaCoVs

The lineage C of betaCoVs includes two bat species closely related to MERS-CoV, namely Ty-BatCoV HKU4 and Pi-BatCoV HKU5, isolated from the lesser bamboo bats (*Tylonycteris pachypus*) and Japanese pipistrelles (*Pipistrellus abramus*), respectively (4). Additional viruses belonging to the lineage C of betaCoVs have been described in bats (BtCoV/133, BtVs-BetaCoV/SC2013) and hedgehogs (Hedgehog coronavirus, EriCoV) (21, 43-45). Recently, a virus belonging to the same species as MERS-CoV was isolated in *Neoromicia* bats (NeoCoV) (46). To investigate the evolutionary history of ORF1a, we obtained sequence information for these viruses and for 6 MERS-CoV strains isolated from either humans or camels and belonging to the major groups described to date (47) (Fig. 1). The sequence alignment was pruned of unreliably aligned codons (see Material and Methods), a procedure that resulted in the masking of the almost entire acidic domain in nsp3. Indeed, this region was previously shown to be highly divergent among CoVs (48). We next analyzed the alignment for the presence of recombination breakpoints using GARD (Genetic Algorithm Recombination Detection) (29) and GENECONV (30). No evidence of recombination was detected.

The pylogenetic tree of ORF1a obtained with phyML was consistent with previously reported ones (46, 47). An estimate of the extent of functional constraint along ORF1a was obtained by identification of negatively selected sites (total number =903) and calculation of their distribution among nsps. This analysis indicated that the average fraction of negatively selected sites is ~0.24, with weakest selection in nsp1 and strongest constraint in nsp6-9 (Fig. 1). Evidence of episodic positive selection along the internal branches of the ORF1a phylogeny was
searched for using branch-site tests. Specifically, we applied two different methods: the branch-site
unrestricted statistical test for episodic diversification (BUSTED) (35) and the maximum-likelihood
models (MA/MA1) implemented in the PAML suite (33). These two approaches rely on different
assumptions of ω (nonsynonymous/synonymous rate ratio) variation among branches. To be
conservative, episodic positive selection at each tested branch was declared only if statistically
significant support was obtained with both methods. Using this criterion, we found 4 branches with
evidence of episodic selection (Fig. 1, Table 1). Selected sites along these branches were identified
using the Bayes empirical Bayes (BEB) procedure from model MA and with BUSTED; again, only
sites detected by both methods were considered. A total of 55 selected sites were detected; these
were scattered along the entire ORF1a and located in different nsps, with the exclusion of nsp7 and
nsp9, which were not targeted by selection (Fig. 1, Table S2). To determine whether any nsp
represented a preferential target of episodic positive selection, we performed random uniform
sampling (i.e. assuming a random distribution of selected sites, we calculated the likelihood of
identifying a number of sites equal to or higher than the one we observed for each nsp). This
analysis indicated that nsp3 was preferentially targeted by episodic selection (Bonferroni corrected
p value= 0.010) during the evolution of MERS-related CoVs.

Finally, we searched for evidence of co-evolution between sites in the ORF1a alignment. To this
aim, we applied BGM-Spidermonkey (40) and MISTIC (41) (see Material and Methods for details).
Six pairs of co-evolving sites were detected by both methods. Most site pairs were located in nsp3
(Fig. 1, Table S3).

**Weaker selective pressure for ORF1b**

The same approach described above was applied to ORF1b sequences. The degree of constraint in
nsp12-16 was comparable to that observed for ORF1a (Fig. 1A). Statistical support of episodic positive selection was detected for 2 branches only (Fig. 1B, Table 1) and fewer selected sites were detected compared to ORF1a. Most sites were targeted by selection on the branch leading to bat CoVs (HKU5, BtCoV/133, HKU4) (Fig. 1, Table S2). No nsp resulted a preferential selection target. One pair of co-evolving sites was detected by MISTIC and BGM-Spidermonkey (Fig. 1A, Table S3).

Ongoing selection at nps3 in MERS-CoV

Given the results above, we investigated whether positive selection also occurred at nsp3 during the recent evolution of MERS-CoV. To this purpose, we retrieved 87 sequences from MERS-CoVs isolated from camels or humans (available strains as of July 2015) (Table S1). No recombination breakpoint was detected by either GARD or GENECONV and the codeml site models were applied (33). Results showed that models of gene evolution that allow a class of codons to evolve with ω >1 (NSsite models M2a and M8) better fit the data than the neutral models (NSsite models M1a and M7), strongly supporting the action of positive selection (Table 2). Positively selected sites in nsp3 were identified using two methods, the BEB procedure implemented in M8 and MEME (Mixed Effects Model of Evolution). Two sites (G720 and R911) were detected by both methods (Table 2, Fig. 2). Four different residues are observed at position 720 in MERS-CoV sequences, and three of them are present in viruses isolated from camels (Fig. 2), suggesting that adaptive evolution at this site started before the spread to humans. Conversely, the R911C substitution only occurs in viruses derived from humans: all of them belong to group 3 and include the virus responsible for the recent South Korean outbreak (MERS-CoV/KOR/KNIH) (Fig. 2).
Positively selected sites in the context of betaCoV biology

To shed light into the functional role of selected sites we exploited available structural, genetic, and biochemical data for nsps (mainly obtained for SARS-CoV and MHV), as well as in silico prediction of transmembrane helices and glycosylation sites.

Most positively selected sites were located in nsp3 (Fig. 3A). In the PLpro domain, four of the positively selected sites, including two that are selected in the viral stains from the current MERS-CoV epidemic, were clustered in a spatially confined region opposed to the catalytic crevice. This region corresponds to the “palm” of the right-handed architecture of the protease, whereas three additional sites are located on the “fingers” (Fig. 3A). One of these three surface exposed sites, Q830, is part of a non-canonical nsp5 cleavage site, unique for MERS-CoV, that could contribute to the formation of new cleavage products (49) (Fig. 3A).

Additional sites were identified in the other domains of the protein (Fig. 3A). One positively selected site in the UbI1 domain (C91) precedes a conserved acidic loop that when mutated in SARS-CoV determines a lethal phenotype (50). Among sites in the ADPR domain, none was located within the ADP-ribose binding pocket (Fig. 3A) (51). As for the nucleic-acid binding domain (NAB) (Fig. 3A), one of the positively selected site (T1080) is part of the α 2 helix, that together with the loop preceding the β 6 strand, forms a structure possibly interacting with ssRNA via a charged patch (52).

Finally, positively selected sites were found in the so-called transmembrane region of nsp3 (Fig. 3). We performed an in silico prediction of transmembrane helices and glycosylation sites. As previously observed for MHV and SARS-CoV (53), the topology model did not fit the general Nendo/Cendo shared structure and was inconsistent with the location of the glycosylation site. Thus, in analogy to other CoVs, some hydrophobic regions predicted to be transmembrane are
unlikely to span the lipid bilayer. Intriguingly, we identified two positively selected sites in the lumenal loop; one of them (A1386) is located between the first two conserved cysteine residues (Fig. 3B) of the zinc finger motif (ZF). Other positively selected sites were found in the third hydrophobic region (that may or may not span the membrane).

In nsp4, four transmembrane regions were predicted (in SARS-CoV and MHV all of these are membrane-spanning) (54) (Fig. 3C). One of the positively selected sites is located on the large lumenal loop 1 (Fig. 3C) and does not affect (or introduces) a predicted glycosylation site. Variants in the same loop of MHV nps4 were shown to alter DVM morphology or number, irrespective of their effect on glycosylation (55). Two additional positively selected sites (G453 and A479) are located in the C-terminal domain of the protein (Fig. 3C). This domain is cytosolic and interacts with viral and host proteins (56).

We also detected 4 positively selected sites in the other viral protease, 3CLpro (nsp5) (Fig. 4). Interestingly, two of them (H8 and V132) were proposed to affect the correct domain orientation required for dimerization of MERS-CoV nsp5 (57). Moreover, site-directed mutagenesis of 3CLpro from MHV indicated that the two corresponding sites affect protease activity in a temperature-dependent fashion (58, 59) (Fig. 4).

Concerning selected sites in nsp10 and ORF1b, interesting findings relate to the nps10-nsp14 and nsp10-nsp16 interactions. Based on the crystal structure of the corresponding SARS-CoV protein complexes, one of the positively selected sites in nsp16 (K249) is located at the direct interaction surface with nsp10 (Fig. 5A). Also, the only positively selected site we detected in nsp10 (S61) is involved in the interaction with both nsp16 (60) and nsp14 (61) (Fig. 5). In fact, mutation of residue 61 in SARS-CoV nsp10 (A61V mutant) strongly reduces interaction with nsp14. In MHV and SARS-CoV, S61 is located on an exposed loop that also includes Q65 (Fig. 5); mutation of Q65 results in a temperature sensitive (ts) phenotype in MHV and, through a poorly understood
mechanism, disrupts nsp5 function (62). This observation suggests that this loop may be involved in interaction with other nsps, or that aminoacid changes in this loop can modify nsp10 conformation (62).

Additional selected sites in nps14 and 16 do not involve the contact interface with nsp10. In nsp14, T228 is located within one of the zinc fingers of the protein (Fig. 5); these domains are essential for the ExoN proofreading activity of nsp14 (63). Another selected residue (S284) flanks a position (F286 in SARS, highly conserved in coronaviruses) that is involved in the hydrophobic interaction between the ExoN and N7-methyltransferase domain (63).

As for nsp16, the T151 selected site is in close proximity to a residue (L153) (Fig. 5) that originates a ts phenotype in MHV; specifically, the mutation has an effect on RNA synthesis (64). Finally, the positively selected site 138 (Asn in MERS-CoV) is located on a solvent-exposed flexible loop (not present in the structure) on the RNA binding groove; this loop may be involved in interaction with RNA (65).

Discussion

In recent years, the emergence of SARS-CoV and MERS-CoV as dangerous zoonoses stirred great interest in the ecology and evolution of coronaviruses. Both viruses originated in bats and spread to humans via an intermediate host (3, 66). This clearly highlights the potential for CoV host shifting and the relevance of understanding the molecular events underling the adaptation to new species. Since the identification of MERS-CoV, a number of related viruses were isolated from bats (and other mammals) all over the world, suggesting a wide distribution of betaCoVs. Unfortunately, most of these studies reported partial viral sequences, hampering a fully resolved phylogenetic analysis (67-69). Herein we only included viral species/strains with complete information for ORF1a and
ORF1b, and the sequences we analyzed are relatively divergent. This may introduce a high false positive rate in the inference of positive selection due to two major issues: the unreliability of sequence alignments and the saturation of substitution rates. To circumvent these possible problems we adopted a stringent alignment filtering criterion and we computed substitution rates over phylogenies. This latter procedure allows breaking of long branches, resulting in improved rate estimation. In fact, branch-site tests are relatively insensitive to the saturation issue (70). Moreover, we supported all claims of positive selection by the combined use of two methods. Although this may have resulted in a loss of power (that is intrinsically low for branch site tests (32)), we were able to detect several positively selected sites along different branches of the phylogenies. In both ORF1a and ORF1b episodic selection was particularly strong on the branch leading to bat CoVs (HKU4, HKU5, and BtCoV133), possibly reflecting specific characteristics of these host species (large population sizes, high seroprevalence, and wide geographic distributions).

Recently, a study of the spike (S) protein in MERS-CoV-related viruses analyzed a very similar set of viral sequences as the ones herein and detected positive selection at 9 sites (25). The proportion of selected sites we identified in the ORF1a region is much higher than that for the S gene. Albeit counter-intuitive, as the S protein is exposed on the viral surface and functions as a central determinant of host range and of antibody response, this observation indicates that ORF1a represented a major target of positive selection during the evolution of lineage C betaCoVs. Therefore, adaptive evolution in this region might contribute to host shifts or immune evasion. In line with these results, studies on avian influenza A viruses showed that viral surface proteins are not the sole determinants of host range. Conversely, adaptation of avian influenza to mammalian hosts was found to be critically dependent on changes in the polymerase genes (23, 24). Recently, an extended analysis indicated that the adaptation of avian influenza virus to swine was accompanied by substitutions in almost all viral genes. Mutation accrual continued for a long time.
after the host shift, suggesting that multiple mutations progressively optimize viral fitness in the
new host (71); many of these mutations were suggested to represent compensatory or epistatic
changes (71). Interestingly, we also found evidence of co-evolution between site pairs, possibly
suggesting epistatic interactions. The vast majority of co-evolving sites were located in nsp3, in line
with the view that epistatically interacting substitutions are enriched in protein regions undergoing
adaptive evolution (72). Epistasis is very common in viruses and is thought to play an important
role in the evolution of immune evasion, host shifts, and drug resistance (73, 74); this latter is likely
not at play in the case of MERS-CoV and related viruses. Previous studies on RNA viruses have
mainly focused on the effect of epistatic interaction on the emergence of drug resistance (e.g.
influenza virus hemagglutinin and neuraminidase, HCV NS3 protease, HIV-1 protease and reverse
transcriptase) or antibody escape (influenza virus hemagglutinin and neuraminidase) (75-78). Little
information is available on the possible role of co-evolving sites in the context of innate immunity
or adaptation to new hosts. Nonetheless, epistatic interaction between two sites in the Chikungunya
virus E1 envelope glycoprotein underlies the ability of the pathogen to infect a new mosquito vector
(79). These observations suggest that intra- and inter-gene epistasis contributes to determine the
evolutionary trajectories of viral species whenever the environment (broadly defined, including the
host species) changes (74). The co-evolving sites we identified are located in distinct domains of
nsp3 and the pair in nsp12 and nsp14 does not involve residues at the interaction surface between
the two proteins. Thus, we are presently unable to infer the molecular mechanism underlying their
interaction. The generation of mutant viruses or recombinant viral proteins carrying different amino
acids at co-evolving site pairs will be necessary to address the nature of their interaction and the
effect on the viral phenotype. We note, however, that long-range interaction are likely to be
common in ORF1a and ORF1b, both at the intra- and inter-protein level, as demonstrated by
experiments with SARS-CoV and MHV mutants (58, 62). For instance, a temperature sensitive (ts)
mutation in MHV nsp10 affects viral replication by blocking the activity of nsp5 (62). Also, ts mutations in MHV nsp5 can be rescued by second-site suppressor changes located in physically distant protein regions (58).

The CoV RTC is extremely complex and comprises a number of enzymatic activities that constitute a unique repertoire among RNA viruses. Several molecular details on the assembly and functioning of the RTC are still missing and the mechanisms underlying the phenotypic effects of specific changes are often poorly understood. For instance, Stokes and coworkers described a conservative mutation (M575I) located in the N-terminal portion of the ADPR domain of MHV nps3 that determines a ts phenotype with impaired RNA synthesis (80). How the mutation exerts its effects, though, remains unknown. We thus interpreted the effect of positively selected sites in light of what is known of CoV biology, and we compared the location of selected sites to the few substitutions (in either SARS-CoV or MHV) with a known effect, as detailed above. Clearly, the selected sites we identified herein lend themselves to experimental testing to assess their impact on BetaCoV phenotypes. Indeed, evolutionary analyses can provide information on the location and nature of adaptive changes, thus highlighting the presence of functional genetic variants. For instance, the major selection target, nsp3, has multiple enzymatic functions. It would be extremely interesting to use site-directed mutagenesis and biochemical analyses to assess whether the selected sites in the PLpro domain affect the enzyme's specificity not only towards the viral polyprotein, but also in terms of deubiquitination and deISGylation activities, which, in turn, may modulate the viral ability to evade host immune responses (17). Likewise, expression of mutant nsp3 and nsp4 proteins carrying different aminoacids at the positively selected sites in the large luminal loops will be instrumental to determine whether, as shown for other CoVs (54, 55), these changes affect the formation of membrane rearrangements onto which the RTC assembles. Ultimately, the generation of mutant CoVs followed by in vitro infection will clarify the effect of specific changes on viral
fitness, at least in cell cultures.

It is noteworthy that we found nsp3 to represent a preferential target of positive selection and that the adaptive process is ongoing in circulating MERS-CoV strains. A previous genome-wide study of positive selection in MERS-CoV detected only one positively selected site in the S gene (2); however, fewer strains than those we analyzed herein were available at the time of the study and the authors did not include sequences obtained from camels (2). The two selected sites we identified are located in the PLpro domain of nsp3 and, together with two sites selected in bats, map to the “palm” portion of the right-handed structure of the protease. The palm also accommodates the catalytic triad, but the selected sites are located at the opposite side of the crevice, suggesting that they may exert an effect by altering the conformational structure of the protease or via interaction with other viral components or host proteins. In addition to its role in viral replication PLpro functions as a multitasking inhibitor of IFN responses and physically associates with several host innate immunity molecules (20). Because antagonism of the host immune system is considered a major driver of evolutionary change in viruses (81), the G720K variant is an excellent candidate as a modulator of host responses. However, variability at position 720 is also observed among viruses isolated from camels, suggesting that the selective event ensued in these animals or in a previous host and does not represent a specific adaptation to humans. A similar observation was previously reported for the positively selected site in the S protein (25). Conversely, variation at the second selected site (R911) in PLpro was only observed for viruses isolated from human patients. Although this finding may simply reflect the sparse sampling of camel-derived viruses, adaptation of a zoonotic virus to a new host is expected to result in selective events that optimize viral fitness in terms of replication efficiency, transmissibility, and immune evasion (73).

In general, deeper understanding of the adaptive events that underlie host shifts in CoVs and other viruses will be pivotal to predict and prevent future zoonoses. Bats alone host a variety of CoVs that
represent potential threats to human health (67-69). Data herein suggest that monitoring programs and field surveys of CoV diversity and prevalence should envisage molecular characterization of nsp3.

**Funding information**

NAD is supported by the Deanship of Scientific Research, Prolific Research Group Program (PRG-1436-15), Vice Rectorate for Graduate Studies and Scientific Research in King Saud University (KSU), Riyadh, Saudi Arabia. The funder had no role in study design, data collection and interpretation, or the decision to submit the work of publication.
Figure Legends

Figure 1. (A) Schematic representation of ORF1a/ORF1b and their nsps products. nsps are colored in hues of blue depending on the percentage of negatively selected sites. nsp11 is shown in gray because it was not analyzed (because it is too short). Positively selected sites are represented by triangles, with color corresponding to the selected branch in the phylogeny (see panel B). Co-evolving sites are shown below the nsp structure, with different symbols indicating each pair of co-evolving sites. (B) Maximum likelihood phylogenies for ORF1a (left) and ORF1b (right) in lineage C betaCoVs. Branches set as foreground lineages in independent branch-site tests are highlighted with different colors and numbered. Thick branches yielded statistically significant evidence of positive selection. Branch length is proportional to synonymous substitution rate (dS). Co-evolving sites are also reported, with different symbols as indicated in panel A. Positions are relative to each nsp, see also Table S3.

Figure 2. Maximum likelihood phylogeny for nsp3 sequences in a subset of isolates representing MERS-CoV major groups. The amino acid alignment of the region surrounding the two positively selected sites (magenta) in MERS-CoV isolates is also shown. Asterisks denote viruses isolated from dromedary camels.

Figure 3. (A) Representation of nsp3 domain architecture. Positively selected sites are indicated by triangles, co-evolving sites with symbols (see Figure 1 legend). In the enlargements, positively selected sites were mapped onto known domain 3D-structures of MERS-CoV or SARS-CoV (PDB IDs: 2GRI, 3EWR, 4RNA, and 2K87). The acidic domain is shown in gray because it was not analyzed (see text). Topology maps and probability diagrams of transmembrane helices for MERS-CoV nsp3 transmembrane domain (B) and MERS-CoV nsp4 (C). The conserved cysteine residues
and the predicted N-glycosylation sites are mapped onto the lumenal loops. Color codes are as in
Figure 1; yellow denotes protein regions or sites known to be functional and mentioned in the text.

Figure 4. (A) Structure of the dimeric form of MERS-CoV nsp5 (PDB ID: 4YLU). Positively
selected sites affecting the dimerization process are labeled onto the structure. Color codes are as in
Figure 1. Catalytic residues are in yellow. (B) Amino acid alignment of the nsp5 regions
surrounding selected sites. Residues that confer a temperature sensitive phenotype when mutated in
MHV are underlined (see text).

Figure 5. Ribbon representation of SARS-CoV nsp10-nsp16 complex (PDB ID: 2XYQ) (A), and
SARS-CoV nsp10-nsp14 complex (PDB ID: 5C8U ) (B). Positively selected sites in lineage C
betaCoVs are shown in green (see Figure 1), residue involved in inter- or intra-protein interactions
are shown in yellow. An amino acid alignment of the nsp10 exposed loop containing S61 and Q65
is also shown. Functional residues in MHV/SARS-CoV are underlined (see text).
References


ORF1a

nsp1  nsp2  nsp3  nsp4  nsp5  nsp6  nsp8  nsp7  nsp10  nsp9  nsp12  nsp13  nsp14  nsp15  nsp16

ribosomal frame-shift

% negatively selected sites

10-15
15-20
20-25
>25

ORF1b

nsp12  nsp13  nsp14  nsp15  nsp16

MERS-CoV-FRA/UAE
MERS-CoV-Al-Hasa17 2013
MERS-CoV-Wadi-Ad-Dawasir1 2013
MERS-CoV-Riyadh1 2012
MERS-CoV-EMC/2012
NeoCoV
BtVs-BetaCoV/SC2013
Pi-BatCoV HKU5
BtCoV/133
Ty-BatCoV HKU4

EriCoV/2012-216
EriCoV/2012-174

MERS-CoV-Hafr-Al-Batin1 2013

EriCoV/2012-174

NeoCoV

BtVs-BetaCoV/SC2013

Pi-BatCoV HKU5

BtCoV/133

Ty-BatCoV HKU4

EriCoV/2012-216
EriCoV/2012-174
A

monomer A

monomer B

V132

H8

B

<table>
<thead>
<tr>
<th>virus</th>
<th>SGLVKMSHPSGDVEAMQVTVCG...YNRPITGFTTVMRPNYTIKGSFLCSGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERS-CoV-EMC/2012</td>
<td>SGLVKMSHPSGDVEACMQVTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>MERS-CoV-Al-Hasa17</td>
<td>SGLVKMSHPSGDVEACMVQTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>MERS-CoV-Wadi-Ad-Dawasir1</td>
<td>SGLVKMSHPSGDVEACMVTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>MERS-CoV-FRA/ UAE</td>
<td>SGLVKMSHPSGDVEAMQVTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>MERS-CoV-Riyadh1</td>
<td>SGLVKMSHPSGDVEAMQVTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>MERS-CoV-Hafr-Al-Batin1</td>
<td>SGLVKMSHPSGDVEAMQVTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>NeoCoV</td>
<td>SGLVKMSHPSGDVEAMQVTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>BtVs-BetaCoV/SC2013</td>
<td>SGLVKMSHPSGAVACMVQTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>Pi-BatCoV HKU5</td>
<td>SGLVKMSHPSGAVACMVQTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>BtCoV/133</td>
<td>SGLVKMSHPSGAVACMVQTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>Ty-BatCoV HKU4</td>
<td>SGLVKMSHPSGAVACMVQTVCG...YNRPITGFTTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>EriCoV/2012-174</td>
<td>SGLVKMSAPSGAVACMVQTVCG...YNRPITGFTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>EriCoV/2012-216</td>
<td>SGLVKMSAPSGAVACMVQTVCG...YNRPITGFTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>SARS CoV</td>
<td>SGLVKMSAPSGAVACMVQTVCG...YNRPITGFTVMRPNYTIKGSFLSCGC</td>
</tr>
<tr>
<td>MHV-A59</td>
<td>SGLVKMSAPSGAVACMVQTVCG...YNRPITGFTVMRPNYTIKGSFLSCGC</td>
</tr>
</tbody>
</table>
A

<table>
<thead>
<tr>
<th>Virus</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERS-CoV-EMC/2012</td>
<td>SVKPESTADQET</td>
</tr>
<tr>
<td>MERS-CoV-Al-Hasa17</td>
<td>SVKPESTADQET</td>
</tr>
<tr>
<td>MERS-CoV-Wadi-Ad-Dawasir1</td>
<td>SVKPESTADQET</td>
</tr>
<tr>
<td>MERS-CoV-FRA/UAE</td>
<td>SVKPESTADQET</td>
</tr>
<tr>
<td>MERS-CoV-Riyadh1</td>
<td>SVKPESTADQET</td>
</tr>
<tr>
<td>MERS-CoV-Hafr-Al-Batin1</td>
<td>SVKPESTADQET</td>
</tr>
<tr>
<td>NeoCoV</td>
<td>SVKPESTADQET</td>
</tr>
<tr>
<td>BtVs-BetaCoV/SC2013</td>
<td>SVKPSNADQET</td>
</tr>
<tr>
<td>Pi-BatCoV HKU5</td>
<td>SVKPEANADQET</td>
</tr>
<tr>
<td>BtCoV/133</td>
<td>SVKPEATADQET</td>
</tr>
<tr>
<td>Ty-BatCoV HKU4</td>
<td>SVKPEATADQET</td>
</tr>
<tr>
<td>EriCoV/2012-174</td>
<td>SVKPESNLDQET</td>
</tr>
<tr>
<td>EriCoV/2012-216</td>
<td>SVKPESNLDQET</td>
</tr>
<tr>
<td>SARS CoV</td>
<td>TVTPEATNQES</td>
</tr>
<tr>
<td>MHV-A59</td>
<td>TIKPEATNQES</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Residue</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>S61</td>
<td></td>
</tr>
<tr>
<td>T228</td>
<td></td>
</tr>
<tr>
<td>S284</td>
<td></td>
</tr>
<tr>
<td>S371</td>
<td></td>
</tr>
<tr>
<td>K214</td>
<td></td>
</tr>
<tr>
<td>T151</td>
<td></td>
</tr>
<tr>
<td>nsp14</td>
<td></td>
</tr>
<tr>
<td>nsp10</td>
<td></td>
</tr>
<tr>
<td>nsp16</td>
<td></td>
</tr>
</tbody>
</table>

MERS-CoV-EMC/2012 SVKPESTADQET
MERS-CoV-Al-Hasa17 SVKPESTADQET
MERS-CoV-Wadi-Ad-Dawasir1 SVKPESTADQET
MERS-CoV-FRA/UAE SVKPESTADQET
MERS-CoV-Riyadh1 SVKPESTADQET
MERS-CoV-Hafr-Al-Batin1 SVKPESTADQET
NeoCoV SVKPESTADQET
BtVs-BetaCoV/SC2013 SVKPSNADQET
Pi-BatCoV HKU5 SVKPEANADQET
BtCoV/133 SVKPEATADQET
Ty-BatCoV HKU4 SVKPEATADQET
EriCoV/2012-174 SVKPESNLDQET
EriCoV/2012-216 SVKPESNLDQET
SARS CoV TVTPEATNQES
MHV-A59 TIKPEATNQES
Table 1. Likelihood ratio test statistics for branch-site tests.

<table>
<thead>
<tr>
<th>Region</th>
<th>Foreground branch(a)</th>
<th>-2ΔlnL(b) (MA vs MA1)(c)</th>
<th>MA vs MA1 p value</th>
<th>-2ΔlnL(b) (BUSTED)</th>
<th>BUSTED p value</th>
<th>N. of sites identified by BEB and BUSTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF1a</td>
<td>Branch 18</td>
<td>4.35</td>
<td>0.0369 (0.0369)</td>
<td>0.50</td>
<td>0.778</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Branch 17</td>
<td>49.18</td>
<td>2.34x10^{-12} (3.90x10^{-12})</td>
<td>23.18</td>
<td>9.25x10^{-6}</td>
<td>1 (nsp3)</td>
</tr>
<tr>
<td></td>
<td>Branch 16</td>
<td>75.22</td>
<td>4.21x10^{-16} (1.05x10^{-17})</td>
<td>85.9</td>
<td>2.58x10^{-10}</td>
<td>13 (1 nsp1, 1 nsp2, 8 nsp3, 2 nsp4, 1 nsp8)</td>
</tr>
<tr>
<td></td>
<td>Branch 23</td>
<td>93.39</td>
<td>4.29x10^{-12} (2.14x10^{-21})</td>
<td>71.76</td>
<td>2.61x10^{-19}</td>
<td>41 (6 nsp2, 25 nsp3, 4 nsp4, 4 nsp5, 1 nsp6, 1 nsp10)</td>
</tr>
<tr>
<td></td>
<td>Branch 25</td>
<td>26.84</td>
<td>2.20x10^{-7} (2.75x10^{-7})</td>
<td>7.52</td>
<td>0.023</td>
<td>-</td>
</tr>
<tr>
<td>ORF1b</td>
<td>Branch 18</td>
<td>0.78</td>
<td>0.38 (0.475)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Branch 17</td>
<td>8.53</td>
<td>0.0035 (0.0058)</td>
<td>2.8</td>
<td>0.246</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Branch 16</td>
<td>12.70</td>
<td>3.6x10^{-4} (0.0009)</td>
<td>17.36</td>
<td>1.70x10^{4}</td>
<td>1 (nsp12)</td>
</tr>
<tr>
<td></td>
<td>Branch 23</td>
<td>37.83</td>
<td>7.73x10^{-10} (3.86x10^{-6})</td>
<td>31.94</td>
<td>1.16x10^{-7}</td>
<td>11 (2 nsp12, 1 nsp13, 3 nsp14, 1 nsp15, 4 nsp16)</td>
</tr>
<tr>
<td></td>
<td>Branch 25</td>
<td>0.23</td>
<td>0.63 (0.63)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

\(a\) Branches are numbered as in figure 1.

\(b\) 2ΔlnL is twice the difference of the natural logs of the maximum likelihood of the models being compared.

\(c\) MA and MA1 are branch-site models that assume four classes of sites: the MA model allows a proportion of codons to have ω ≥ 1 on the foreground branches, whereas the MA1 model does not.
### Table 2. Likelihood ratio test statistics for models of variable selective pressure among sites.

<table>
<thead>
<tr>
<th>region</th>
<th>LRT model</th>
<th>-2ΔlnL</th>
<th>p value</th>
<th>% of sites (average ω)</th>
<th>Positively selected sites (BEB and MEME)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsp3</td>
<td>M1a vs M2a&lt;sup&gt;a&lt;/sup&gt;</td>
<td>15.25</td>
<td>0.00048</td>
<td>0.6 (17.24)</td>
<td>G720, R911</td>
</tr>
<tr>
<td></td>
<td>M7 vs M8&lt;sup&gt;b&lt;/sup&gt;</td>
<td>14.15</td>
<td>0.00085</td>
<td>1.6 (11.59)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> M1a is a nearly neutral model that assumes one ω class between 0 and 1, and one class with ω=1; M2a (positive selection model) is the same as M1a plus an extra class of ω >1.

<sup>b</sup> M7 is a null model that assumes that 0<ω<1 is beta distributed among sites; M8 (positive selection model) is the same as M7 but also includes an extra category of sites with ω>1.

<sup>c</sup> 2ΔlnL: twice the difference of the natural logs of the maximum likelihood of the models being compared.