Avoiding Regions Symptomatic of Conformational and Functional Flexibility to Identify Antiviral Targets in Current and Future Coronaviruses

Jordon Rahaman[1] and Jessica Siltberg-Liberles[1,2,*]

[1]Department of Biological Sciences, Florida International University, Miami, FL

[2]Department of Biological Sciences, Biomolecular Sciences Institute, Florida International University, Miami, FL

*Corresponding author: E-mail: jliberle@fiu.edu.

Accepted: October 3, 2016

Abstract

Within the last 15 years, two related coronaviruses (Severe Acute Respiratory Syndrome [SARS]-CoV and Middle East Respiratory Syndrome [MERS]-CoV) expanded their host range to include humans, with increased virulence in their new host. Coronaviruses were recently found to have little intrinsic disorder compared with many other virus families. Because intrinsically disordered regions have been proposed to be important for rewiring interactions between virus and host, we investigated the conservation of intrinsic disorder and secondary structure in coronaviruses in an evolutionary context. We found that regions of intrinsic disorder are rarely conserved among different coronavirus protein families, with the primary exception of the nucleocapsid. Also, secondary structure predictions are only conserved across 50–80% of sites for most protein families, with the implication that 20–50% of sites do not have conserved secondary structure prediction. Furthermore, nonconserved structure sites are significantly less constrained in sequence divergence than either sites conserved in the secondary structure or sites conserved in loop. Avoiding regions symptomatic of conformational flexibility such as disordered sites and sites with nonconserved secondary structure to identify potential broad-specificity antiviral targets, only one sequence motif (five residues or longer) remains from the >10,000 starting sites across all coronaviruses in this study. The identified sequence motif is found within the nonstructural protein (NSP) 12 and constitutes an antiviral target potentially effective against the present day and future coronaviruses. On shorter evolutionary timescales, the SARS and MERS clades have more sequence motifs fulfilling the criteria applied. Interestingly, many motifs map to NSP12 making this a prime target for coronavirus antivirals.

Key words: structural disorder, evolutionary dynamics, Coronavirus, evolution, divergence, MERS-CoV.

Introduction

Severe Acute Respiratory Syndrome (SARS)-CoV and Middle East Respiratory Syndrome (MERS)-CoV are two closely related zoonotic coronaviruses. Both have successfully crossed the species barrier to allow animal-to-human transmission, and further to allow human-to-human transmission (Song et al. 2005; Reusken et al. 2016). The SARS outbreak in 2003 had a mortality rate of 10% (Anderson et al. 2010), and SARS-CoV was considered the most aggressive coronavirus compared to other human coronaviruses that commonly cause mild to moderate infection in their hosts (van der Hoek 2007). MERS-CoV is the cause of an ongoing outbreak of the respiratory illness MERS (de Groot et al. 2013). At the time of writing, 1791 MERS cases have been confirmed with a mortality rate of approximately 35% (World Health Organization 2016). Both MERS and SARS have higher mortality rates in elderly and immunosuppressed populations (Gralinski and Baric 2015).

The host changes by MERS-CoV and SARS-CoV suggest that other coronaviruses can potentially cross the species barrier, become zoonotic, and enable human-to-human transmission, ultimately causing high morbidity and mortality. SARS-CoV and MERS-CoV exploited mechanistically different approaches to overcome the human species barrier, but these two viruses have a lot in common (Lu et al. 2015). Here, we aim to identify the vulnerable regions in the proteomes of coronaviruses that neither SARS-CoV nor MERS-CoV nor their contemporary and forthcoming relatives can proliferate without, and address how to

mobilize a defense against the present and future coronaviruses by targeting these regions.

SARS-CoV and MERS-CoV are positive (+)-strand RNA viruses encoding approximately 25 protein products. The MERS-CoV proteome is primarily composed of two polyproteins, ORF1a and ORF1ab; the latter is generated by a -1 ribosomal slippage frameshift. These proteins are cleaved into 16 nonstructural proteins (NSPs). NSPs 1–10 are products of both polyproteins, whereas NSPs 12–16 are only yielded by ORF1ab. NSP11 is unique to ORF1a (van Boheemen et al. 2012). Structural proteins envelope (E), spike (S), membrane (M), and nucleocapsid (N) are elements of the physical structure that encloses the viral genome and come from distinct reading frames, unlike ORF1a and ORF1ab, which come from overlapping reading frames. Additionally, the structural proteins are the product of subgenomic mRNAs that are joined during discontinuous negative RNA strand synthesis (van Boheemen et al. 2012). Finally, NS3 protein (NS3), NS4A protein (NS4A), NS4B protein (NS4B), NS5 protein (NS5), and Orf8b protein encompass the remainder of the proteome and also arise from distinct reading frames (van Boheemen et al. 2012).

Our approach utilizes genomic sequence data, which is readily available for viruses known to cause disease. However, because most viruses pose no major threat to their host, they pass by unnoticed leaving the majority of virus genome space uncharted. With the availability of cost-efficient genome sequencing technology, and recent developments in the field of viral metagenomics, large-scale identification of viral genome space is on the rise (Rosario and Breitbart 2011; Mokili et al. 2012). By exploring viral diversity, critical components constituting a viral genus' fitness can be evaluated. Examples such as the common influenza virus illustrate the rapidity of viral gene mutation and in order to maintain immune protection, an annual flu vaccination is recommended. Underway efforts aim to generate broadly neutralizing vaccines whose design accounts for the genomic sequences of multiple types of influenza virus to eliminate frequent re-vaccination against the flu (Giles and Ross 2011, 2012). Development of broadly neutralizing vaccines often relies on the consensus or ancestral sequences of extant viral sequences in order to provide greater coverage for related viruses (Kesturu et al. 2006). Unfortunately, consensus sequences can be misleading, and ancestral sequence reconstruction is error-prone for quickly diverging sequences (McCloskey et al. 2014). In addition, viruses with compact genomes often express proteins with structural disorder that may undergo structural transformations. Although these transformer proteins, like VP40 in Ebola, are masters at changing their structure, and thus expanding their functional repertoire as needed for the life cycle of the virus (Bornholdt et al. 2013), flexible regions are potentially important in rewiring protein–protein interactions between the virus and its host (Le Breton et al. 2011; Ortiz et al. 2013; Gitlin et al. 2014). The flexibility trait of many viral proteins is a complicating factor in vaccine development. For instance, Dengue virus exhibits serotype-specific antibody affinity that causes antibody-dependent enhancement, an obstacle in the development of Dengue vaccines that protects against all four serotypes (Flipse and Smit 2015). To overcome the hurdle posed by structural flexibility, we propose an additional screening step in identifying potential vaccine or antiviral targets that considers the structural flexibility of the viral proteins. The Structural Genomics Initiatives increased their success rate by excluding proteins predicted to be structurally disordered (Slabinski et al. 2007). A similar approach can perhaps benefit vaccine development. Furthermore, to make this approach robust to potential mutations, minimizing loss in efficacy or resistance, the evolutionary context of sequence and structure must be considered. Thus, we suggest expanding the concept of broadly neutralizing vaccines/antivirals by increasing the diversity of viruses considered if possible. Sites conserved for sequence, structure, and with low disorder propensity among diverse virus protein homologs are very likely to be constrained from 1) changing sequence on evolutionary time scales and 2) undergoing real-time structural transitions. These sites have potential as targets for broad-specificity antivirals or vaccines because conservation makes them broad-specificity and low dynamics avoids targeting a conformational ensemble, which is not only difficult (Yu et al. 2016), but that may change as the sequence diverges (Siltberg-Liberles et al. 2011).

A recent large-scale study of structural disorder in >2,000 viral genomes in 41 viral families found the amount of disorder in different virus families varying from 2.9% to 23.1% (Pushker et al. 2013). It was reported that *Coronaviridae* has very low disorder content (mean disorder 3.68%) (Pushker et al. 2013). *Coronaviridae* contains two subfamilies: *Coronavirinae* and *Torovirinae*. SARS-CoV and MERS-CoV are part the *Coronavirinae* subfamily, from here on referred to as coronavirus (CoV). The lack of disorder is intriguing because it may be important for rewiring interactions between viral proteins and host proteins (Ortiz et al. 2013) and providing opportunities to acquire novel functional sequence motifs (Gitlin et al. 2014). Structural disorder has also been proposed to be important for viral viability, enabling multifunctionality and vigor in response to changes in the environment (Xue et al. 2014). Given the low fraction of structural disorder reported across *Coronaviridae*, we set out to investigate the conservation of structural disorder and secondary structure across CoV. Sites identified as conserved for structure and lacking disorder can be considered to be vulnerable and druggable in the proteomes of coronaviruses. The structural divergence capacity of these regions is limited, leaving a wider range of the present and emergent coronaviruses susceptible to the effects of potential broadly neutralizing anti-CoV therapies targeting these sites. We will refer to these sites as target sites.

Materials and Methods

Protein Family Reconstruction

Protein sequences were identified by individual BLAST searches with MERS-CoV (Taxonomy ID: 1335626) proteins ORF1ab (YP_009047202.1; polyprotein), S protein (YP_009047204.1), M protein (YP_009047210.1), E protein (YP_009047209.1), and N protein (YP_009047211.1) against coronaviruses. BLAST searches of the ORF1ab protein were performed, using start and end positions as detailed in the ORF1ab NCBI Reference Sequence file, against the refseq_protein database. The sequences retrieved from the BLAST output maintained the following cutoff: >30% sequence identity and >50% coverage relative to MERS-CoV sequence query. The 30% sequence identity and 50% query coverage cutoff strikes a balance between alignment quality and at least 10 sequences for most protein families. NSP1 (YP_009047202.1; 1-193), NSP2 (YP_009047202.1; 194-853), NS3 (YP_009047205.1), NS4A (YP_009047206.1), NS4B (YP_009047207.1), NS5 (YP_009047208.1), ORF8b protein (YP_009047212.1), and NSP11 (YP_009047203.1; 4378-4391) are not included in this study due to <10 BLAST hits.

Multiple sequence alignments were constructed for the selected BLAST hits using MAFFT (Katoh et al. 2002). Phylogenetic trees were constructed using MrBayes 3.2.2 with a four category gamma distribution and the mixed model for amino acid substitution (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Each tree ran for five million generations, with a sample frequency of 100. The final tree was constructed from the last 75% of samples, discarding the first 25% of samples as the default burnin, and using the half-compatible parameter, to avoid weakly supported nodes (i.e., with a posterior probability <0.5). All trees were midpoint rooted.

For every protein family, the amino acid substitution rate per site in its multiple sequence alignment was calculated using empirical Bayesian estimation as implemented in Rate4Site (Mayrose et al. 2004). Substitution rates were calculated using 16 gamma categories, the JTT substitution matrix (Jones et al. 1992), and the reconstructed phylogenies. The rates were normalized per protein family with an average across all sites equal to zero and SD equal to 1. This means that sites with a rate <0 are evolving slower than average, whereas sites with a rate >0 are evolving faster than average.

Prediction of Intrinsic Disorder Propensity and Secondary Structure

Intrinsic disorder propensity was inferred using two different predictors: IUPred (default settings; "long" option) (Dosztányi et al. 2005a, 2005b) and DISOPRED2 (Ward et al. 2004) for all proteins. For IUPred, the site-specific continuous disorder propensities for each protein were mapped onto their corresponding position in the multiple sequence alignment as raw disorder propensities and as binary states, order or disorder, using two cutoffs of 0.4 and 0.5. Disorder propensities below the cutoff were assigned order and disorder propensities at the cutoff or above were assigned disorder. For the DISOPRED2 predictions that were inferred using the nr database, the continuous disorder propensities for every site in a protein were mapped onto their corresponding position in the multiple sequence alignment as raw disorder propensities and as binary states, order or disorder, using a cutoff of 5. Consequently, for every protein family (a multiple sequence alignment and its corresponding phylogenetic tree), two continuous matrices and three binary matrices resulted: IUPred 0.4, IUPred 0.5, and DISOPRED2. An additional matrix was generated to indicate sites where the binary order and disorder assignments differ between IUPred 0.4 and DISOPRED2.

A similar methodology was employed to analyze secondary structure predicted by PSIPRED (McGuffin et al. 2000) and JPred (Drozdetskiy et al. 2015). For both predictors, the uniref90 database was used and sites were classified as loops, alpha helices, or beta strands and mapped back onto their corresponding sites in the multiple sequence alignment. This resulted in two three-state matrices for each protein family alignment, one for each predictor, and two binary matrices displaying secondary structure elements (alpha helix and beta strand) or loops. An additional matrix was generated to indicate sites where the secondary structure assignments differ between PSIPRED and JPred.

For every protein family, the binary matrices resulting from the different disorder predictions and from the different secondary structure predictions were analyzed in the corresponding evolutionary context using GLOOME. GLOOME (Gain-Loss Mapping Engine) analyzes binary presence and absence patterns in a phylogenetic context (Cohen et al. 2010). In this study, the Rate4Site option in GLOOME was used to analyze the binary matrices (IUPred 0.4, IUPred 0.5, DISOPRED2, PSIPRED, and JPred) with the corresponding phylogenetic trees to map change of state across sites in each individual protein phylogeny (Cohen and Pupko 2010; Cohen et al. 2010). GLOOME was run with 16 gamma categories and a substitution matrix set to equal rates within each state and transitions between states treated equally. From the binary disorder and order matrices, transition rates between disorder and order or vice versa (DOT) were estimated. From the binary structure and loop matrices, transition rates between structure and loop or vice versa (SLT) were estimated. Similar to Rate4Site, the rates were normalized per protein family with an average across all sites equal to zero and SD equal to 1. This means that sites with a rate <0 are evolving slower than average, while sites with a rate >0 are evolving faster than average.

Protein Family Visualization

Protein families were visualized in an integrative manner with a phylogenetic tree, any matrix (multiple sequence alignment

or predictor based) displayed as a heatmap, and site-specific sequence transition rates using Python packages ETE3 (Huerta-Cepas et al. 2016) and Matplotlib (Hunter 2007).

Statistical Analysis of Amino Acid Evolutionary Rate Distributions

Amino acid evolutionary rates (SEQ) for all sites across all alignments were aggregated and binned into four possible categories characterized by the distribution of PSIPRED predicted secondary structure at each site. Sites predicted to have a loop across all sequences are "conserved loops; C(L)" and sites predicted to have a helix across all sequences or a strand across all sequences are "conserved helix-strand; C(HS)" (table 3). Sites predicted to have all three states (helix, strand, and loop) or any combination of loop and one other state are "non-conserved helix, loop, strand; NC(HLS)" and sites predicted to have a mixture of helix and strand are "non-conserved helix-strand; NC(HS)" (table 3). In all cases, gaps were ignored when classifying combinations of secondary structure at a site or if secondary structure conservation exists at a particular site.

Results

Phylogenies

Phylogenies were built for all protein products encoded in the MERS-CoV single-stranded RNA genome, except for NSP1, NSP2, NS3, NS4A, NS4B, ORF8b protein, and NSP11, all of which had insufficient sequence data (<10 sequence hits with BLAST).

NSP12 is often used as a measure for newly identified coronaviruses. According to the International Committee of Taxonomy of Viruses, a major criterion in determining if a coronavirus is considered novel is pairwise sequence identity below 90% for NSP12 in all comparisons to previously known coronaviruses (Bermingham et al. 2012). Four main clades, alphacoronavirus, betacoronavirus, gammacoronavirus, and deltacoronavirus (fig. 1), are identified in agreement with the taxonomic classifications described by the ICTV (International Committee on Taxonomy of Viruses 2015). Coronaviruses not listed by the ICTV are assumed to be a part of the clade in which representatives with known classifications are situated in our NSP12 phylogeny.

The MERS clade and SARS clade are sister clades in the NSP12 phylogeny. The HKU1 clade and EQU clade are also sister clades. Together these four clades form the Betacoronavirus clade, in accordance with the ICTV classification (International Committee on Taxonomy of Viruses 2015). Betacoronavirus is represented in all phylogenies although the order of the individual subclades varies. Alphacoronavirus is often found as the sister clade or outgroup to betacoronavirus. Deltacoronavirus or gammacoronavirus are the most distantly related to the betacoronavirus. In the nucleocapsid phylogeny, gammacoronavirus is the first outgroup clade to betacoronavirus, and alphacoronavirus is the most distant outgroup. Most NSP trees exhibit some unresolved nodes at junctures immediately preceding terminal nodes. As an effect of the 50% majority rule, most of the 546 resolved nodes are well supported with posterior probability >0.9 for 82% and >0.99 for 68% (supplementary fig. S1, Supplementary Material online). Most trees follow the NSP12 topology for the main clades, with minor clade rearrangements. It should be noted that for NSP5, the entire alphacoronavirus clade is placed within the betacoronavirus clade, as a sister clade to the MERS clade (supplementary fig. S1, Supplementary Material online). This may be due to increased sequence divergence rates or due to recombination. Recombination events are rather frequent in coronaviruses (Su et al. 2016), and the MERS clade potentially underwent multiple recombination events as part of the host change (Zhang et al. 2016).

The phylogenies for membrane protein, spike protein, NSP5, and NSP8–NSP16 demonstrate (with the given BLAST cutoffs) recoverable protein homologs such that all coronaviruses are represented (i.e., all coronaviruses represented in the NSP12 phylogeny). Nucleocapsid, NSP4, and NSP7 have recoverable homologs in all clades except deltacoronavirus. NSP3 and NSP6 homologs are too divergent in deltacoronavirus and/or gammacoronavirus relative to MERS-CoV. Envelope appears specific to betacoronavirus (fig. 1), but it is a short protein that has been found to diverge rapidly and is likely present outside betacoronavirus (Fehr and Perlman 2015). Because different protein families yield slightly different phylogenies, for the remaining evolutionary analyses, every protein family was analyzed in the context of its own phylogeny.

Intrinsic Disorder Is Rarely Conserved

For all protein families, structural disorder propensities were predicted using IUPred (Dosztányi et al. 2005a, 2005b) and DISOPRED2 (Ward et al. 2004). To verify the robustness of the binary IUPred and DISOPRED2 predictions, the binary assignments were compared on a site-by-site basis (table 1). When converted to binary (i.e., two states per site disordered or ordered) IUPred 0.4 and IUPred 0.5 are in good agreement with the larger differences seen for NSP8, NSP9, and nucleocapsid (7.5%, 6.5%, and 19.0%, respectively) (table 1). Comparing IUPred 0.4 or IUPred 0.5 to DISOPRED2, large differences are in particular seen for nucleocapsid (38.7% and 29.7% respectively) and NSP8 (23.5% and 25.9%, respectively) (table 1). For nucleocapsid, regions that are found to be disordered by IUPred 0.4 are found to be ordered by IUPred 0.5 and DISOPRED2 (fig. 2 and supplementary fig. S2, Supplementary Material online). For NSP8, regions that are only slightly disordered in a few sequences according to

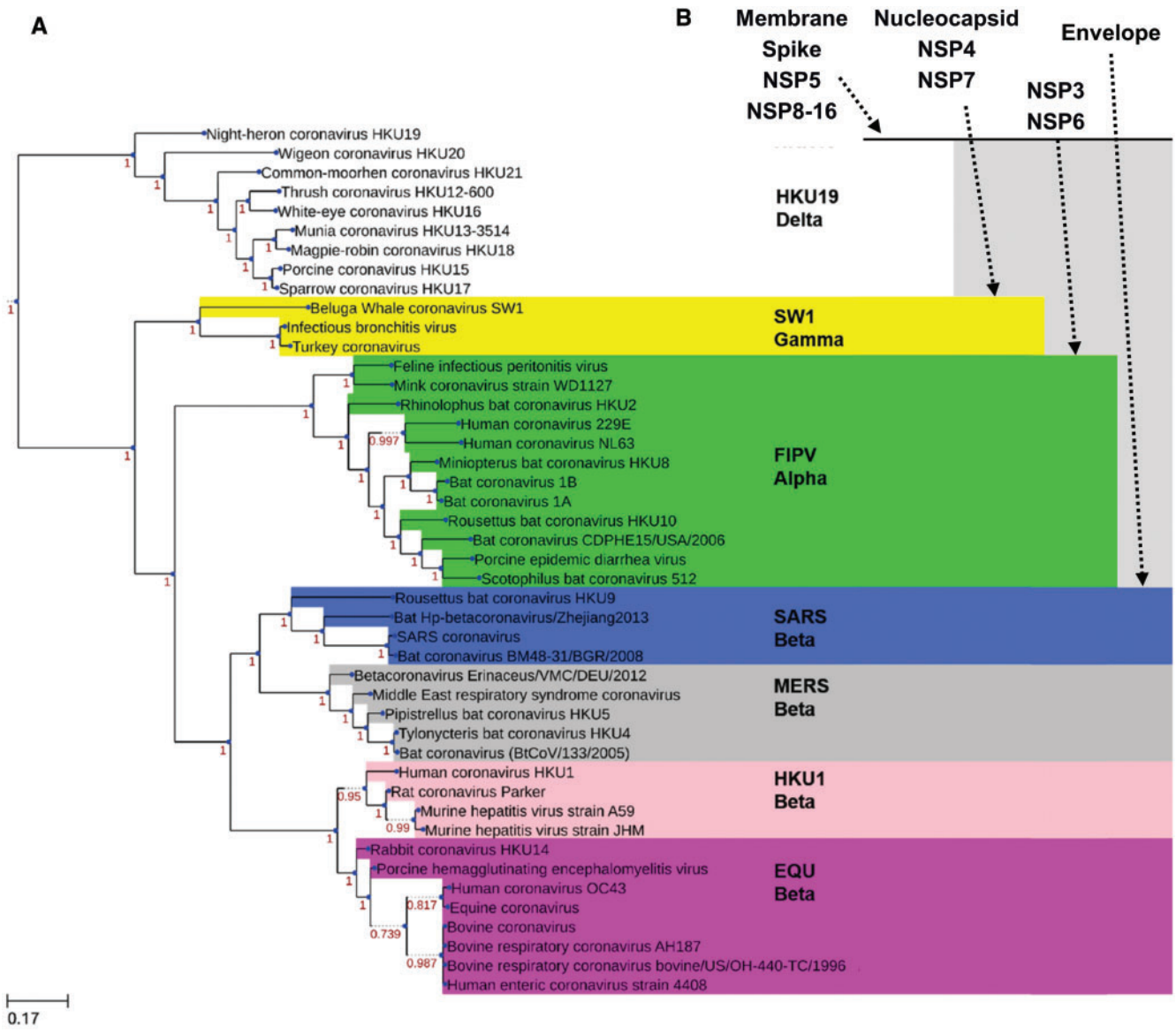


Fig. 1.—CoV representative phylogeny. (*A*) NSP12 is a representative for the CoV protein phylogenies, colored by clade (alphacoronavirus or FIPV, green; betacoronavirus has four different subclades: SARS, blue; MERS, gray; HKU1, pink; EQU, purple; gammacoronavirus or SW1, yellow; deltacoronavirus or HKU19, white.) Posterior probability indicating node support is shown in red. (*B*) Protein family distribution across coronavirus based on the given cutoff (>30% sequence identity and >50% coverage relative to MERS-CoV sequence query). Clade color applied throughout the remaining figures. Areas shaded in gray with an arrow indicate that the protein family is not identified for that clade with the given cutoffs, but is found from the arrow tip. See supplementary fig. S1, Supplementary Material online, for the remaining phylogenetic trees.

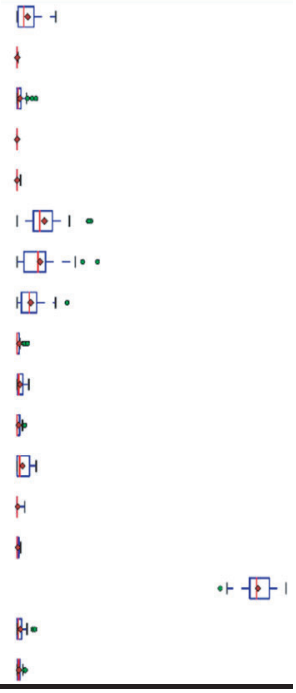
IUPred 0.4 and IUPred 0.5, DISOPRED2 predicts disorder to be conserved for all sequences (fig. 3).

To quantify the fraction of disordered sites per protein family, we report the IUPred 0.4 results only for simplicity (table 1). In general, IUPred 0.4 predicts more disorder than DISOPRED2, but several protein families have almost no disordered sites. NSP3 and NSP8-10 have some variation in disorder content for different viruses. Based on the fraction of disorder, nucleocapsid is the only highly disordered protein among the CoVs in this study, even if NSPs 8–10 have outliers that are >20% disordered.

To compare the disorder-to-order transition rates (DOT) for all protein families where the binary matrices of disorder and order include both states, the quadrant count ratio (QCR) was estimated as a measure of association in assigning slower than average vs. faster than average transition rates. For IUPred 0.4 vs. IUPred 0.5, for IUPred 0.5 vs DISOPRED2, and for IUPred 0.4 vs. DISOPRED2, the QCRs

Table 1
Protein Family Wide Disagreement of Disorder and Secondary Structure Predictions

Predictor / Protein family	IUPred 0.4 vs. IUPred 0.5 %	IUPred 0.4 vs. DISOPRED2 %	IUPred 0.5 vs. DISOPRED2 %	PSIPRED vs. JPred %	Disorder fraction[1] IUPred 0.4
NSP3	2.1	4.6	3.2	21.2	
NSP4	0.1	0.9	0.8	22.6	
NSP5	0.8	3.1	2.4	13.7	
NSP6	0.0	0.8	0.8	18.6	
NSP7	0.0	5.4	5.5	8.1	
NSP8	7.5	23.5	25.9	12.1	
NSP9	6.5	11.0	7.2	20.4	
NSP10	3.9	12.1	9.8	24.6	
NSP12	0.5	1.4	1.0	20.2	
NSP13	1.0	8.8	8.5	25.4	
NSP14	0.5	0.8	0.3	16.7	
NSP15	0.9	2.0	1.1	16.3	
NSP16	0.2	1.4	1.3	14.1	
ENVELOPE	0.3	10.0	10.4	29.7	
NUCLEOCAPSID	19.0	38.7	29.7	10.2	
MEMBRANE	1.0	9.5	10.2	20.4	
SPIKE	0.6	2.6	2.1	17.7	
TOTAL DISAGREEMENT	1.9	5.6	4.7	18.8	



[1]Tukey boxplot constructed using the IUPred 0.4 predicated disorder fraction (number of disordered sites/total sites) per sequence per protein. Green dots represent outliers; red diamond are the mean and red lines are the median values.

were 0.76, 0.69, and 0.63, respectively. This shows a strong positive association for site-specific DOT for all methods and cutoffs, with IUPred 0.4 vs. IUPred 0.5 being the strongest (table 2). For nucleocapsid and NSP8, the positive associations are weaker, suggesting that many sites have IUPred disorder propensity in the 0.4 to 0.5 range and large differences between IUPred and DISOPRED2, in accordance with the large disagreement between the binary assignment of these predictors (tables 1 and 2).

Secondary Structure Prediction and Structure-to-Loop Transitions

For all protein families, secondary structure elements were predicted using PSIPRED (McGuffin et al. 2000) and JPred (Drozdetskiy et al. 2015). For most protein families, the disagreement between secondary structure predictors is greater than for the disorder predictors (table 1). In fact, 15 of the 17 protein families compared disagree at more than 10% of alignment sites, and two of these disagree at more than 20% of sites. To compare the binary structure-to-loop transitions (SLT), QCR was estimated as a measure of association for SLT based on the different predictors. In general, there is a moderate positive association between SLT for PSIPRED vs. SLT for JPred that is weaker than for the different DOT comparisons (table 2). It should be noted that SLT does not differentiate between alpha helix and beta strand, but considers both as "structure." This is a correct assumption if protein structure is conserved and consistently predicted, but for some protein families that is not the case.

Four protein families (NSP3, NSP12, NSP13, and SPIKE) have more than 40% of their sites found within the NC(HLS) category with non-conserved helix, strand, and loop (two or three states present at the same site) (table

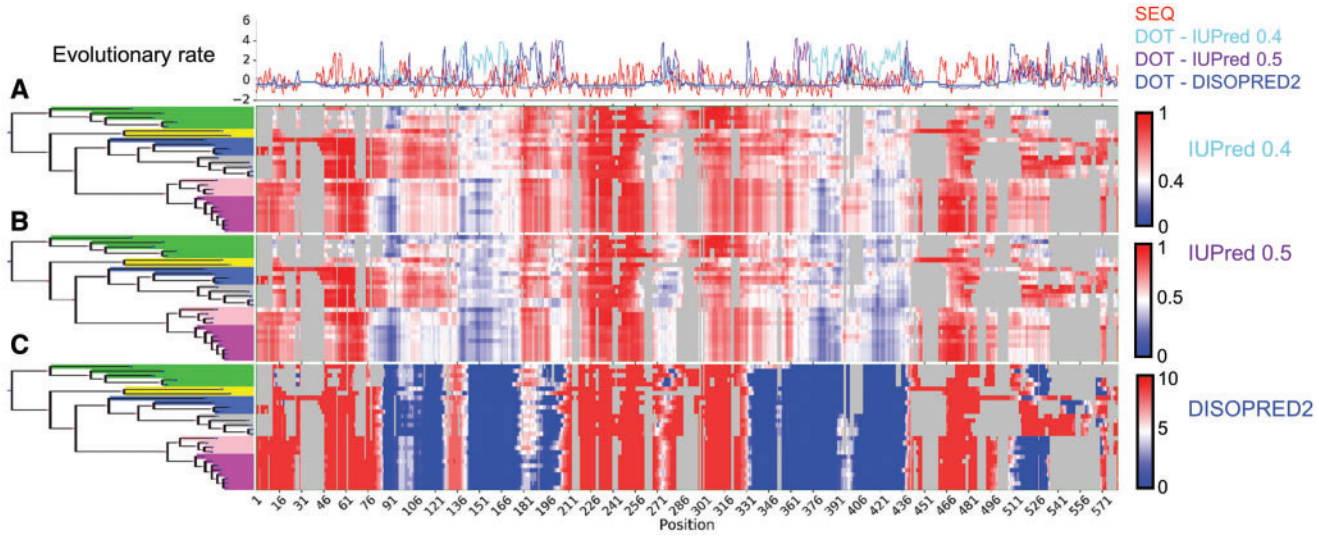


Fig. 2.—The evolutionary context of intrinsic disorder in nucleocapsid. The phylogenetic tree was built using the multiple sequence alignments for nucleocapsid. Here, the multiple sequence alignment is colored by disorder propensity (with gaps in gray): (*A*) IUPred 0.4, blue-to-white-to-red shows disorder propensity according to the scale for IUPred 0.4. (*B*) IUPred 0.5, blue-to-white-to-red shows disorder propensity according to the scale for IUPred 0.5. (*C*) DISOPRED2, blue-to-white-to-red shows disorder propensity according to the scale for DISOPRED2. Above the heat maps, the normalized evolutionary rates per site for amino acid substitution (SEQ) and the DOT for the binary transformations of *A*–*C* are shown. Heat maps visualized with the Python packages ETE3 (Huerta-Cepas et al. 2016) and Matplotlib (Hunter 2007).

3). For NSP13, JPred predicts 72% of all sites to be a mixture of helix, strand, and loop, or any combination of loop and one other structural element (fig. 4). Envelope and NSP6 have 13% and 12% of their respective sites in the NC(HS) category. Considering only the PSIPRED predictions, the NC(HS) category has 245 sites across all 17 protein families. That is one-tenth the size of the next smallest set which is C(HS) with 2275 sites. Next, C(L) has 3344 sites, and the largest category is NC(HLS) with 4257 sites. Comparing the evolutionary sequence rates for the sites in the different categories, based on PSIPRED predictions only, reveals that sites in the C(HS) category are evolving at a slower rate than all other categories. NC(HS) is only just significantly different (P = 4.62E−03) from C(HS), and is not significantly different from NC(HLS) and C(L) (P = 1.85E−02 and P = 8.33E−01, respectively). However, NC(HLS) and C(L) are significantly different from each other, and both are significantly different from C(HS) (P = 1.82E−46 and P = 2.33E−21, respectively) (fig. 5).

Identifying Target Sites

For regions with five or more consecutive sites that were 100% conserved in sequence across 1) all CoV or 2) across the MERS and SARS clades, the information of structural disorder prediction from IUPred and DISOPRED2 was used to identify all ungapped sites that were consistently predicted to have 100% conserved order. Next, the information of secondary structure prediction from PSIPRED and JPred was used to narrow down this list further by only including sites that are not changing their predicted secondary structure state for both predictors. Applying the aforementioned filters to the initial 10,000 sites resulted in one (1) region of five residues or more conserved across all CoV within the N-terminal domain of NSP12: DNQDL (table 4). Interestingly, this region is in the vicinity of sites found important for nucleotidylating activity across the order *Nidovirales* (Lehmann et al. 2015).

Considering only the sequences in the SARS and MERS clades, 21 sequence regions of five residues or more were found in seven protein families (table 4). For NSP5, NSP7, and NSP14, experimentally determined structures show that most regions are surface accessible (fig. 6). Some of the identified target sites are known for their functional importance. For instance, C145 in the middle of GSCGS in NSP5 is part of the catalytic dyad in the NSP5 protease (Yang et al. 2003). For NSP12 and NSP13, which have the majority of all sites, no structures are available. The sites adjacent to DNQDL are also conserved in the SARS and MERS clades, and five additional target sites, conserved for the SARS and MERS clades, are found in the C-terminal direction relative to the DNQDL motif (table 4). Continuing into the RNA-dependent RNA polymerase domain (RdRP) in NSP12, four additional regions of target sites are found, and the last three regions are found in the C-terminal part. Importantly, in RdRP and in the C-terminal part are sites that are also conserved across all CoVs in this study. NSP13 has four regions of target sites distributed across the protein.

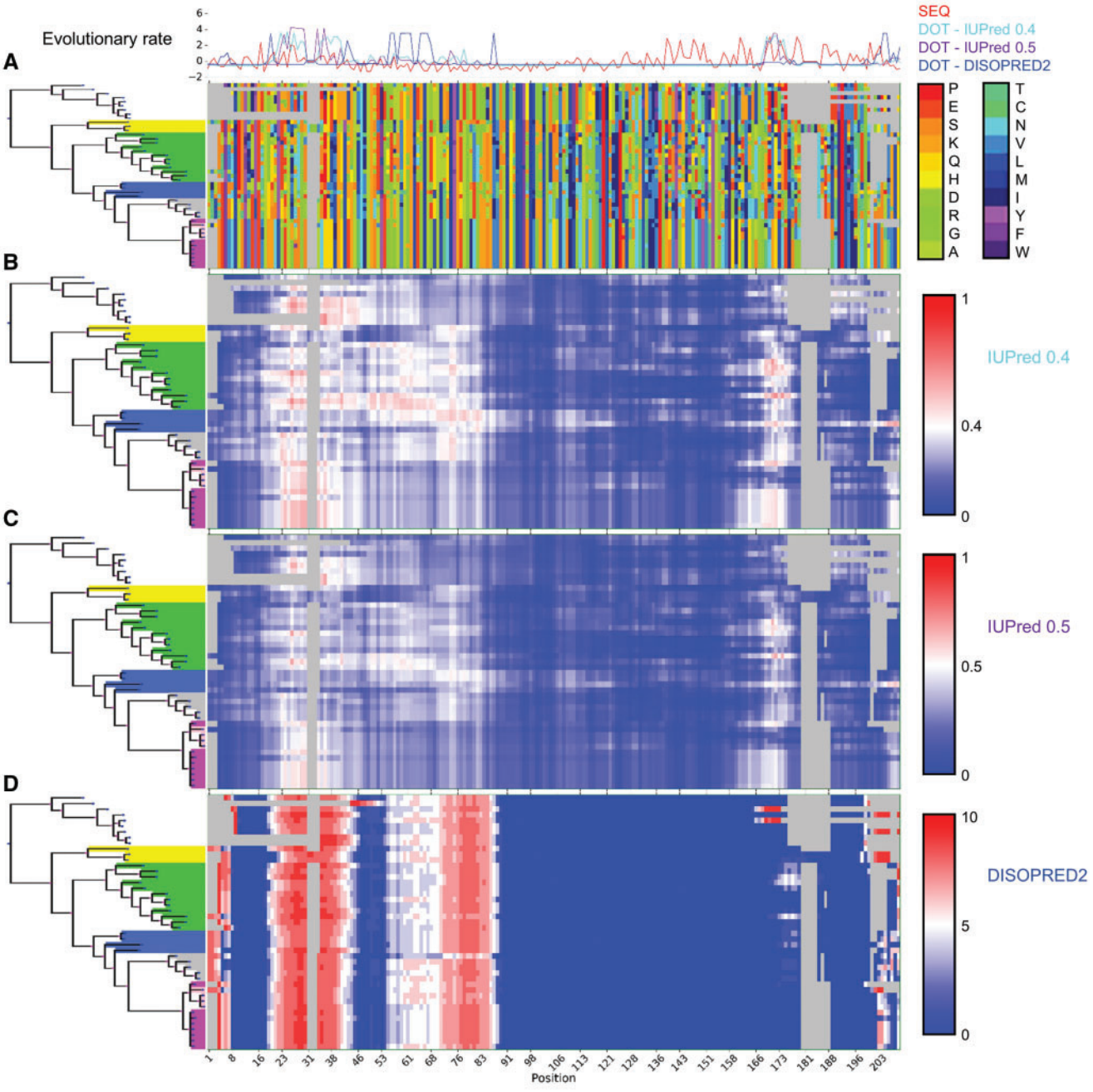


Fig. 3.—The evolutionary context of intrinsic disorder in NSP8. The phylogenetic tree was built using the multiple sequence alignments for NSP8. (*A*) The multiple sequence alignment is colored by amino acid according to scale, arranged based on TOP-IDP disorder promoting propensity of the amino acids (Campen et al. 2008), and gray denotes gaps. (*B*) IUPred disorder propensity per site in the multiple sequence alignment. Blue-to-white-to-red shows disorder propensity according to the scale for IUPred 0.4. (*C*) IUPred disorder propensity per site in the multiple sequence alignment. Blue-to-white-to-red shows disorder propensity according to the scale for IUPred 0.5. (*D*) DISOPRED2 disorder propensity per site in the multiple sequence alignment. Blue-to-white-to-red shows disorder propensity according to the scale. Above the multiple sequence alignment, the normalized evolutionary rates per site for amino acid substitution (SEQ) and the DOT for the binary transformations of *B*–*D* are shown. Heat maps visualized with the Python packages ETE3 (Huerta-Cepas et al. 2016) and Matplotlib (Hunter 2007). See supplementary figures S2 and S4, Supplementary Material online for additional graphics for every protein family.

Table 2
QCR[a] for DOT and SLT

Protein family	Rate			
	DOT	DOT	DOT	SLT
	IUPred 0.4 vs. IUPred 0.5	IUPred 0.5 vs. DISOPRED2	IUPred 0.4 vs. DISOPRED2	PSIPRED vs. JPred
NSP3	0.75	0.68	0.61	0.51
NSP4	N/A[b]	N/A	0.58	0.61
NSP5	0.72	0.84	0.73	0.65
NSP6	N/A	N/A	N/A	0.70
NSP7	N/A	N/A	0.9	0.66
NSP8	0.86	0.43	0.38	0.67
NSP9	0.76	0.55	0.62	0.67
NSP10	0.93	0.67	0.68	0.42
NSP12	0.96	0.89	0.86	0.57
NSP13	0.76	0.70	0.59	0.36
NSP14	0.87	0.85	0.8	0.58
NSP15	0.82	0.85	0.67	0.53
NSP16	0.93	0.86	0.85	0.59
Envelope	N/A	N/A	0.58	0.53
Membrane	0.54	0.57	0.67	0.55
Nucleocapsid	0.33	0.43	0.31	0.61
Spike	0.81	0.62	0.49	0.55
All	0.76	0.69	0.63	0.55

[a]QCR: Quadrant Count Ratio measures the association for the same site-specific rate with different predictors or cutoffs.
[b]N/A: at least one of the rates in the comparison could not be estimated due to the lack of any disordered state in the binary state matrix (supplementary fig. S5, Supplementary Material online).

Discussion

We have analyzed the protein evolution of the genetic components that make up the MERS-CoV proteome. As previously established, MERS-CoV has the same genomic makeup as HKU4-CoV and HKU5-CoV in the MERS clade (Woo et al. 2012). Some protein products are only found in the MERS clade, and these were excluded from this study due to insufficient data. Furthermore, for other protein products, some clades may not be represented in our protein families if their proteins were too divergent. This was an important factor in determining the applied BLAST hit cutoffs, as relaxing cutoffs produced alignments with more gaps and increasing stringency reduced the representative pool. Because alignment quality is important due to the sensitivity of both Rate4Site and for phylogenetic reconstruction, the chosen cutoffs are suitable. We note some clade-specific differences in recoverable homologs between different CoV, but many components are shared among them (fig. 1).

Viral proteins often possess multifunctionality, mediated by a conformational change in response to environment-specific factors (Xue et al. 2014). Although conformational flexibility is important for function, it also offers flexibility in what sequence motifs are on display. If these sequences are rapidly diverging, different sequence motifs will be displayed, reinforcing the notion that flexible regions are potentially important in rewiring protein–protein interactions between virus and host (Gitlin et al. 2014). Although most CoV proteins have almost no intrinsic disorder, several CoV protein families have homologous sites that display loop in some sequences, helix in others and strands in some (table 3, supplementary fig. S3, Supplementary Material online). These sites are not necessarily disordered but they may be conformationally flexible in real-time (with secondary structure transitions in the same sequence, making them difficult to predict) or on evolutionary time-scales (so that different secondary structure elements actually are present in different sequences). The C(HS) and C(L) sites make up approximately 50–80% of most multiple sequence alignments. With the common expectation that protein structure is more conserved than sequence these numbers are surprisingly low. Neither PSIPRED nor JPred consistently predicts the same state for 20–50% of all sites in these multiple sequence alignments.

The accuracy of PSIPRED and JPred's secondary structure predictions are about 80% (Bryson et al. 2005; Drozdetskiy et al. 2015). PSIPRED has been found to rarely predict an alpha helix instead of a beta strand and vice versa, and most of the PSIPRED errors are due to secondary structure not being predicted (Li et al. 2014). When secondary structure is not conserved for the same site in a multiple sequence alignment, it suggests that the secondary structure prediction may be 1) inaccurate, 2) not predicted with high confidence, or 3) the

Table 3
Structural Conservation of Sites Per Protein Family

Protein Family	Conserved Sites Helix OR Strand C(HS)		Conserved Sites Loop C(L)		Non-conserved sites Helix AND Strand AND Loop NC(HLS)		Non-conserved sites Helix AND Strand NC(HS)	
	% sites per protein family alignment per predictor							
	PSIPRED	JPred	PSIPRED	JPred	PSIPRED	JPred	PSIPRED	JPred
NSP3	18.64	16.46	28.54	24.45	50.94	56.21	1.88	2.88
NSP4	29.74	46.84	29.00	22.49	34.20	29.00	7.06	1.67
NSP5	34.88	39.81	37.04	41.36	27.78	17.90	0.31	0.93
NSP6	40.81	59.81	10.59	11.21	36.45	25.86	12.15	3.12
NSP7	45.78	55.42	24.10	24.10	30.12	20.48	0.00	0.00
NSP8	49.28	55.50	21.05	24.88	27.75	18.66	1.91	0.96
NSP9	40.52	37.93	34.48	38.79	24.14	21.55	0.86	1.72
NSP10	17.12	25.34	58.22	40.41	22.60	28.08	2.05	6.16
NSP12	24.74	25.05	34.17	27.46	39.62	44.13	1.47	3.35
NSP13	16.67	7.43	33.83	17.33	47.69	72.44	1.82	2.81
NSP14	22.71	29.49	44.32	43.04	31.50	26.01	1.47	1.47
NSP15	22.68	21.22	39.76	39.02	35.37	35.37	2.20	4.39
NSP16	23.87	22.26	43.23	40.32	28.71	32.90	4.19	4.52
Envelope	31.11	46.67	16.67	16.67	38.89	28.89	13.33	7.78
Membrane	35.51	43.12	26.09	24.28	29.35	30.80	9.06	1.81
Nucleocapsid	8.78	11.19	62.82	62.99	27.71	24.96	0.69	0.86
Spike	18.35	19.52	28.85	26.15	51.98	53.16	0.81	1.17
Total number of sites	2275		3344		4257		245	

Color guide	0-10	>10-20	>20-30	>30-40	>40-50	>50-60	>60	%

regions are indeed metamorphic; they can transition from one element to another. Although (1) is difficult to address without experimentally determined structures for all sequences, (2) and (3) are not necessarily incompatible interpretations because low confidence secondary structure prediction could indicate metamorphic secondary structure regions. Metamorphic secondary structure regions have interesting consequences for conformational and functional flexibility.

It should be noted that, despite the low amount of disordered sites in most CoV proteins, several regions are not conserved in disorder propensity across all sequences, but sometimes the different predictors disagree as in the case of NSP8. Clade-specific disordered regions resulting from indel events suggest that they are not essential to the critical functions of the protein, but could cause gain-and-loss of interactions with its hosts. However, when disorder propensity is only mildly fading for a region that is present across the protein family, it may be important for the fundamental function of the protein. The virus structural proteins that interact to form the virion commonly include an envelope protein, a membrane protein, and a capsid protein that together form the machinery that encases, transports, and releases the virus. The interactions between the structural proteins are often regulated by conformational changes like VP40 in Ebola (Bornholdt et al. 2013) and Envelope protein from Dengue virus (Zheng et al. 2014). Conformational changes

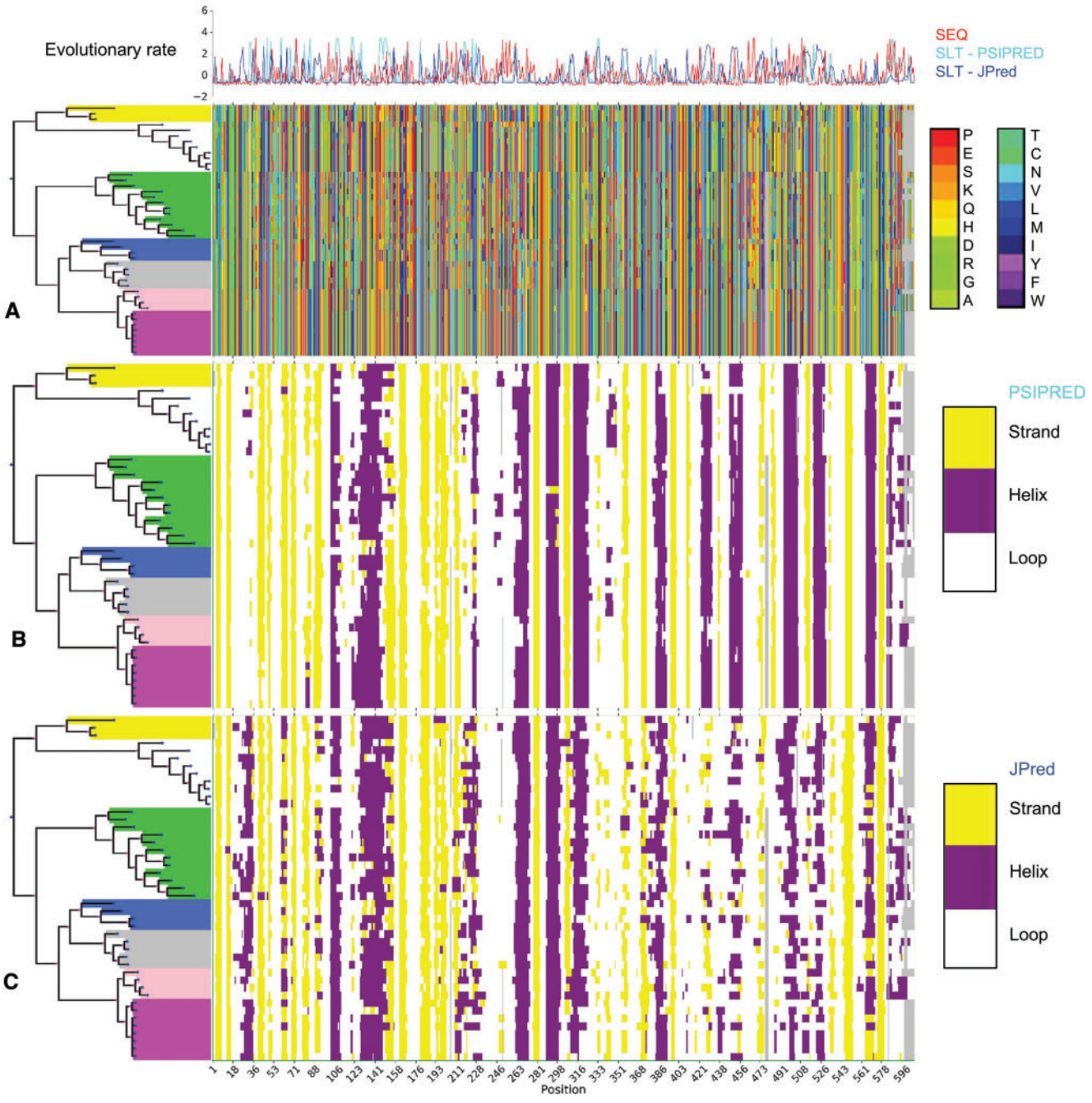


Fig. 4.—The evolutionary context of secondary structure in NSP13. The phylogenetic trees were built using the multiple sequence alignments for NSP13. (*A*) The multiple sequence alignment is colored as in fig. 3. (*B*) PSIPRED secondary structure prediction per site in the multiple sequence alignment, color coded according to the scale. (*C*) JPred secondary structure prediction per site in the multiple sequence alignment, color coded according to the scale. Above the multiple sequence alignment, the normalized evolutionary rates per site for sequence substitution (SEQ) and SLT based on the binary transformations of B-C are shown. Heat maps visualized with the Python packages ETE2 (Huerta-Cepas et al. 2016) and Matplotlib (Hunter 2007). See supplementary figs. S3 and S4, Supplementary Material online for a complete set of graphics for every protein family.

in these proteins are needed for the virus life cycle. For CoVs, nucleocapsid is the only structural protein that is highly disordered. Yet, rapid evolutionary dynamics of disorder is present in nucleocapsid using two different IUPred cutoffs (0.4 and 0.5) and with DISOPRED2. Even if the different predictors and cutoffs disagree somewhat where regions with rapid evolutionary dynamics are present, these patterns suggest that nucleocapsid may be rapidly changing from one virus

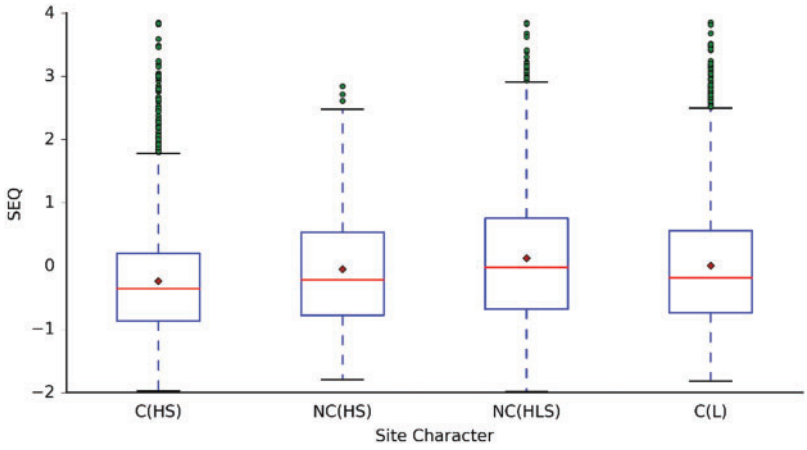


Fig. 5.—Comparison of SEQ at sites characterized by secondary structure. All pairwise rate distributions, except NC(HS) vs. NC(HLS) and NC(HS) vs. C(L), are significantly different ($P < 0.05$, after Bonferroni correction: $P < 0.008$). For a summary of the U statistic and two-tailed P values for each pairwise comparison see supplementary table S2, Supplementary Material online.

Table 4
Sites Conserved in Sequence and Structural Property

Protein family	PfamA domain	Conserved sites in the MSA[a]
NSP5	*Peptidase_C30*	149-GSCGS-153[b]
		213-AWLYAA-218[b]
NSP7	*Replicase*	7-KCTSVVLL-14[b]
		16-VLQQL-20[b]
NSP12	*RPol N-term*	228-L**DNQDL**NG-235
		239-DFGDF-243
	RdRP_1	521-DKSAG-525
		588-MTNRQ-592
		677-LANECAQVL-685
		800-**GGT**SSGD-706
	C-term	853-YPDPSR-858
		871-KTDGT-875
		889-**YPL**TK-893
NSP13	*N-term*	10-SQTSLR-15
	AAA_30	362-NALPE-366
		402-DPAQLP-407
	AAA_12	539-SSQGS-543
NSP14	*NSP11*	281-AHVAS-285[b]
		290-MTRCLA-295[b]
		438-HAFHT-442[b]
		494-CNLGG-499[b]

[a]Sites conserved across all clades in the protein family are underlined and in BOLD font. All other sites are conserved across the SARS and MERS clades.
[b]Experimentally determined structures are available in Protein Data Bank (Berman 2000)..

to another. It should also be noted that two MERS clade specific inserts around position 241 and toward the C-terminal are consistently predicted to be highly disordered. With inserts and changing structural dynamics between clades or viruses, the questions become 1) which sequence motif are displayed and 2) to what extent are these sequence motifs displayed?

Furthermore, based on the inconsistent prediction of secondary structure elements, the possibility that CoVs are more conformationally flexible than their intrinsic disorder content implies is noteworthy. Altogether, this suggests that various mechanisms for rewiring conformational and functional space are operating in the coronaviruses studied here. If regions symptomatic of conformational and functional flexibility can be avoided in order to identify broad-specificity antiviral targets with potential to be effective against coronaviruses of today and in the future, coronaviruses as a group may become more attractive drug targets for the pharmaceutical

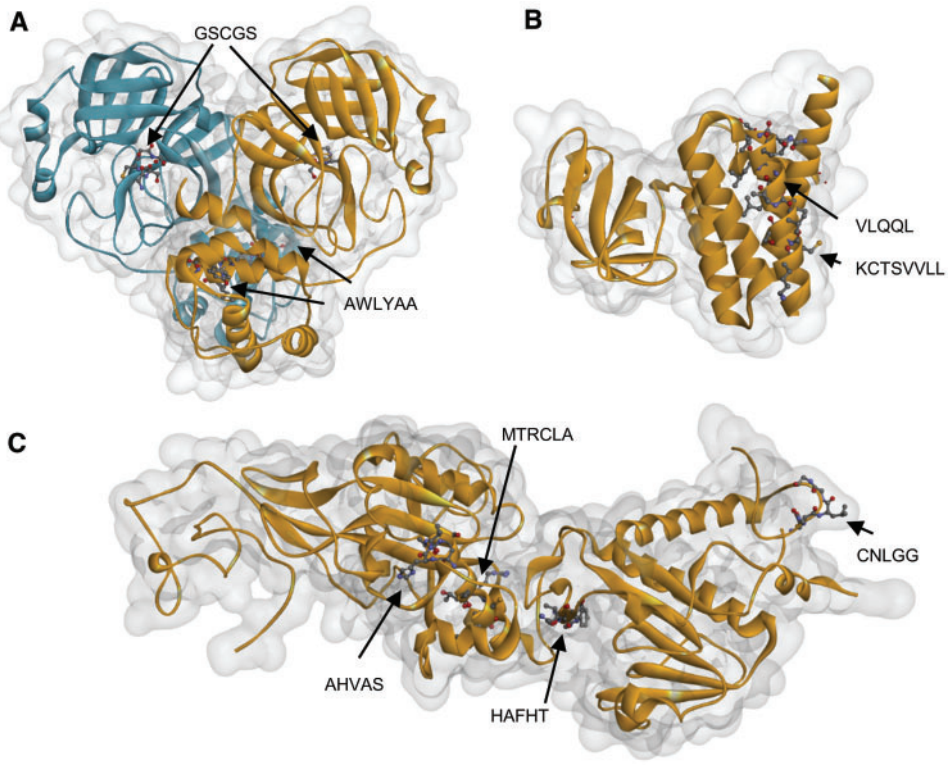


Fig. 6.—Target sites shown in 3D context. (*A*) NSP5 dimer, based on PDB id 1UK4 (Yang et al. 2003). (*B*) NSP7, based on PDB id 5F22 (unpublished). (*C*) NSP14, based on PDB id 5C8T (Ma et al. 2015). Protein structure visualized with Bioviva Discovery Studio .

industry in the event an additional coronavirus changes host to include humans or increase its virulence.

Supplementary Material

Supplementary tables S1 and S2 and figures S1–S5 are available at Genome Biology and Evolution online (http://www.gbe. oxfordjournals.org/).

Acknowledgments

We thank Joseph Ahrens, Janelle Nunez-Castilla, and Helena Gomes Dos Santos for assistance in the lab and for helpful discussions. The authors would also like to acknowledge the Instructional & Research Computing Center (IRCC) at Florida International University for providing HPC computing resources that have contributed to the research results reported within this article, web: http://ircc.fiu.edu.

Literature Cited

Anderson LJ, Tong S. 2010. Update on SARS research and other possibly zoonotic coronaviruses. Int J Antimicrob Agents 36 Suppl 1:S21–S25.
Berman HM. 2000. The Protein Data Bank. Nucleic Acids Res. 28:235–242.
Bermingham A, et al. 2012. Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012. Euro Surveill. 17:20290.
Bornholdt ZA, et al. 2013. Structural rearrangement of ebola virus VP40 begets multiple functions in the virus life cycle. Cell 154:763–774.
Le Breton M, et al. 2011. Flavivirus NS3 and NS5 proteins interaction network: a high-throughput yeast two-hybrid screen. BMC Microbiol. 11:234.
Bryson K, et al. 2005. Protein structure prediction servers at University College London. Nucleic Acids Res. 33:W36–W38.
Campen A, et al. 2008. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett. 15:956–963.
Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME: gain loss mapping engine. Bioinformatics 26:2914–2915.
Cohen O, Pupko T. 2010. Inference and characterization of horizontally transferred gene families using stochastic mapping. Mol Biol Evol. 27:703–713.
de Groot RJ, et al. 2013. Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. J Virol 87:7790–7792.
Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005a. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433–3434.
Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005b. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 347:827–839.
Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 43:W389–W394.
Fehr AR, Perlman S. 2015. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol. 1282:1–23.
Flipse J, Smit JM. 2015. The Complexity of a Dengue Vaccine: A Review of the Human Antibody Response. PLoS Negl Trop Dis. 9:e0003749.

Giles BM, Ross TM. 2011. A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets. Vaccine 29:3043–3054.
Giles BM, Ross TM. 2012. Computationally optimized antigens to overcome influenza viral diversity. Expert Rev Vaccines 11:267–269.
Gitlin L, Hagai T, LaBarbera A, Solovey M, Andino R. 2014. Rapid evolution of virus sequences in intrinsically disordered protein regions. PLoS Pathog 10:e1004529.
Gralinski LE, Baric RS. 2015. Molecular pathology of emerging coronavirus infections. J Pathol. 235:185–195
Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.
Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 33:1635–1638. doi: 10.1093/molbev/msw046.
Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng. 9:90–95.
International Committee on Taxonomy of Viruses. 2015. Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses. (2012) Ed: King, A.M.Q., Adams, M.J., Carstens, E.B. and Lefkowitz, E.J. San Diego: Elsevier Academic Press.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci 8:275–282.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Kesturu GS, et al. 2006. Minimization of genetic distances by the consensus, ancestral, and center-of-tree (COT) sequences for HIV-1 variants within an infected individual and the design of reagents to test immune reactivity. Virology 348:437–448.
Lehmann KC, et al. 2015. Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. Nucleic Acids Res. 43:8416–8434.
Li Q, Dahl DB, Vannucci M, Hyun J, Tsai JW. 2014. Bayesian model of protein primary sequence for secondary structure prediction. PLoS One 9:e109832.
Lu G, et al. 2015. Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends Microbiol. 23:468–478.
Ma Y, et al. 2015. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. Proc Natl Acad Sci U S A. 112:9436–9441.
Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol. 21:1781–1791.
McCloskey RM, Liang RH, Harrigan PR, Brumme ZL, Poon AFY. 2014. An evaluation of phylogenetic methods for reconstructing transmitted HIV variants using longitudinal clonal HIV sequence data. J Virol. 88:6181–6194.
McGuffin LJ, Bryson K, Jones DT. 2000. The PSIPRED protein structure prediction server. Bioinformatics 16:404–405.
Mokili JL, Rohwer F, Dutilh BE. 2012. Metagenomics and future perspectives in virus discovery. Curr Opin Virol. 2:63–77.
Ortiz JF, MacDonald ML, Masterson P, Uversky VN, Siltberg-Liberles J. 2013. Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses. Genome Biol Evol. 5:504–513.
Pushker R, Mooney C, Davey NE, Jacqué J-M, Shields DC. 2013. Marked variability in the extent of protein disorder within and between viral families. PLoS One 8:e60724
Reusken CB, Raj VS, Koopmans MP, Haagmans BL. 2016. Cross host transmission in the emergence of MERS coronavirus. Curr Opin Virol 16:55–62.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Rosario K, Breitbart M. 2011. Exploring the viral world through metagenomics. Curr. Opin. Virol 1:289–297.
Siltberg-Liberles J, Grahnen JA, Liberles DA. 2011. The evolution of protein structures and structural Ensembles under functional constraint. Genes (Basel) 2:748–762.
Slabinski L, et al. 2007. The challenge of protein structure determination–lessons from structural genomics. Protein Sci. 16:2472–2482.
Song H-D, et al. 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci U S A. 102:2430–2435.
Su S, et al. 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24:490–502.
van Boheemen S, et al. 2012. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. MBio 3:e00473–e00412.
van der Hoek L. 2007. Human coronaviruses: what do they cause? Antivir Ther. 12:651–658.
Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. 2004. The DISOPRED server for the prediction of protein disorder. Bioinformatics 20:2138–2139.
Woo PC, Lau SK, Li KS, Tsang AK, Yuen K-Y. 2012. Genetic relatedness of the novel human group C betacoronavirus to Tylonycteris bat coronavirus HKU4 and Pipistrellus bat coronavirus HKU5. Emerg Microbes Infect 1:e35.
World Health Organization. 2016. WHO | Middle East respiratory syndrome coronavirus (MERS-CoV). http://www.who.int/emergencies/mers-cov/en/.
Xue B, et al. 2014. Structural disorder in viral proteins. Chem. Rev 114:6880–6911.
Yang H, et al. 2003. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. Proc Natl Acad Sci U S A. 100:13190–13195.
Yu C, et al. 2016. Structure-based inhibitor design for the intrinsically disordered protein c-Myc. Sci Rep. 6:22298.
Zhang Z, Shen L, Gu X. 2016. Evolutionary dynamics of MERS-CoV: potential recombination, positive selection and transmission. Sci Rep. 6:25049.
Zheng A, Yuan F, Kleinfelter LM, Kielian M. 2014. A toggle switch controls the low pH-triggered rearrangement and maturation of the dengue virus envelope proteins. Nat Commun. 5:3877.

Associate editor: Dr. Chantal Abergel