2012

Ranking viruses: Measures of positional importance within networks define core viruses for rational polyvalent vaccine development

Tavis K. Anderson
University of Wisconsin-Madison, tavis.anderson@ars.usda.gov

William W. Laegreid
University of Illinois at Urbana-Champaign, laegreid@uiuc.edu

Francesco Cerutti
University of Torino, Italy, francesco.cerutti@unito.it

Fernando A. Osorio
University of Nebraska - Lincoln, fosorio1@unl.edu

Eric A. Nelson
South Dakota State University, Eric.Nelson@SDSTATE.EDU

See next page for additional authors

Published in *Bioinformatics* (2012); doi: 10.1093/bioinformatics/bts181 Published online in Advance Access April 11, 2012.

Ranking viruses: Measures of positional importance within networks define core viruses for rational polyvalent vaccine development

Tavis K. Anderson,[1] William W. Laegreid,[2] Francesco Cerutti,[3] Fernando A. Osorio,[4] Eric A. Nelson,[5] Jane Christopher-Hennings,[5] and Tony L. Goldberg [1]

1. Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.
2. Department of Veterinary Pathobiology, University of Illinois, Urbana, IL 61802, U.S.A.
3. Department of Animal Production, Epidemiology and Ecology, Faculty of Veterinary Medicine, University of Torino, 10095 Gruglasco (TO), Italy
4. Department of Veterinary and Biomedical Sciences and Nebraska Center for Virology, University of Nebraska, Lincoln, NE 68583, U.S.A.
5. Department of Veterinary and Biomedical Sciences, South Dakota State University, Brookings, SD 57071, U.S.A.

Corresponding author — T. L. Goldberg, tgoldberg@vetmed.wisc.edu

Abstract

Motivation: The extraordinary genetic and antigenic variability of RNA viruses is arguably the greatest challenge to the development of broadly effective vaccines. No single viral variant can induce sufficiently broad immunity, and incorporating all known naturally circulating variants into one multivalent vaccine is not feasible. Further, no objective strategies currently exist to select actual viral variants that should be included or excluded in polyvalent vaccines.

Results: To address this problem, we demonstrate a method based upon graph theory that quantifies the relative importance of viral variants. We demonstrate our method through application to the envelope glycoprotein gene of a particularly diverse RNA virus of pigs: porcine reproductive and respiratory syndrome virus (PRRSV). Using distance matrices derived from sequence nucleotide difference, amino acid difference, and evolutionary distance we constructed viral networks and used common network statistics to assign each sequence an objective ranking of relative "importance." To validate our approach, we use an independent published algorithm to score our top-ranked wild-type variants for coverage of putative T-cell epitopes across the 9383 sequences in our dataset. Top-ranked viruses achieve significantly higher coverage than lower-ranked viruses, and top-ranked viruses achieve nearly equal coverage as a synthetic mosaic protein constructed *in silico* from the same set of 9383 sequences.

Conclusion: Our approach relies on the network structure of PRRSV, but applies to any diverse RNA virus because it identifies subsets of viral variants that are the most important to overall viral diversity. We suggest that this method, through the objective quantification of variant importance, provides criteria for choosing viral variants for further characterization, diagnostics, surveillance, and ultimately polyvalent vaccine development.

1 Introduction

Of all infectious threats to global health, viruses with RNA genomes have proven to be the most difficult to control (Cleaveland *et al.*, 2001; Jones *et al.*, 2008). Viruses such as HIV, influenza virus, hepatitis C virus, and Dengue virus, have defied control in part as a result of their extraordinary genetic and antigenic variation (Katze *et al.*, 2008), and their consequent ability to evolve rapidly (Belshaw *et al.*, 2008; Domingo, 2002; Drake and Holland, 1999; Holmes, 2009). HIV, for example, can evolve into diverse strains within an individual patient over the course of days or weeks, confounding the development of vaccines and antivirals and frustrating diagnostics and surveillance (Crandall, 1999; Gaschen *et al.*, 2002). Thus, the efficacy of even the most immunogenic vaccines is questionable in a field setting where multiple variants may circulate simultaneously, and where novel variants are continuously evolving. Vaccine studies almost universally acknowledge this challenge but have yet to develop objective criteria selecting vaccine candidates that adequately, yet parsimoniously, capture overall viral diversity (Barouch, 2008; Korber *et al.*, 2009).

In an effort to develop vaccines that maximize the representation of antigenic features present in diverse viral populations, a series of computational strategies have been proposed. Driven largely by HIV vaccine research, the approaches have included concatenating commonly recognized T-cell epitopes (Palker *et al.*, 1989), creating pseudoprotein strings of T-cell epitopes (De Groot *et al.*, 2005), and generating consensus overlapping peptide sets from proteins (Thomson *et al.*, 2005). Evolutionary approaches such as the use of consensus sequences (Gao *et al.*, 2004; Gao *et al.*, 2005; Gaschen *et al.*, 2002), and the most recent common ancestor (MRCA) of viral populations, have also been proposed with the assumption that these approaches capture viral diversity (Gaschen *et al.*, 2002). Unfortunately, experimental studies in animal models using these strategies have documented underwhelming humoral immune responses (Doria-Rose *et al.*, 2005; Gao *et al.*, 2005).

The most widely studied approach uses a genetic algorithm to generate, select, and recombine (*in silico*) potential T-cell epitopes into full-length "mosaic" protein sequences that can provide greater coverage of global viral variants than could any single wild-type protein (Fischer *et al.*, 2007). This mosaic protein approach was able to achieve between 74% and 87% coverage of HIV-1 Gag sequences, compared to only between 37% and 67% using a single natural Gag protein (Fischer *et al.*, 2007). Subsequent experimental studies have demonstrated that these computationally designed vaccines augment the breadth and depth of antigen-specific T-cell responses as compared with consensus or natural sequence HIV-1 antigens (Barouch *et al.*, 2010; Santra *et al.*, 2010). However, this approach yields synthetic peptides with unknown biological properties that may not be able to be incorporated into a recombinant live vaccine.

Given the need for an objective strategy to select wild-type viral variants that might be adapted for polyvalent vaccines, we propose a method based on network analysis. In recent years, interest in the analysis and modeling of networks has surged outside the traditional fields of social science (Wasserman and Faust, 1994; Watts, 2004), mathematics and computer science (Albert and Barabasi, 2002; Butts, 2009). Network statistics have proven to be powerful tools for studying diverse phenomena such as social interactions, connections amongst neurons, and the structure of the Internet, and for optimizing solutions to engineering problems (Albert and Barabasi, 2002; Watts, 2004). For our purpose, it has long been recognized that biological molecules interact directly and indirectly (Yamada and Bork, 2009), and network statistics have proven fruitful in drug discovery studies (Azmi *et al.*, 2010; Zhao and Li, 2010), human disease classification (Barabasi *et al.*, 2011; Loscalzo *et al.*, 2007), and in functional analyses of gene regulatory networks (Yu *et al.*, 2007). As such, it is possible to construct viral networks that explicitly define the positional importance of individual viral variants in the network based on relationships between biomolecules (e.g. Allesina and Pascual, 2008; Salathe *et al.*, 2010; Yamada and Bork, 2009).

In the present paper, we use a highly diverse and particularly problematic RNA virus of pigs as a case study: porcine reproductive and respiratory syndrome virus (PRRSV). This exercise provides "proof of concept," as well as a novel strategy for the control of this virus of critical importance to agriculture. We present a detailed topological network analysis that considers the importance of every individual viral variant *i* by considering every possible interaction (direct and indirect) between variant *i* and *j*. We use network indices, from local to global, including node degree, a series of centrality indices and PageRank, the algorithm at the heart of Google. We then examine how these indices relate to each other, and discuss how to choose the most appropriate metric for selecting viral variants for study, and to help guide the development of polyvalent vaccines, diagnostic testing strategies, and epidemiological surveillance. Our study makes minimal initial assumptions about evolutionary processes or viral biology and can complement other approaches such as the mosaic protein approach (Fischer *et al.*, 2007) to polyvalent vaccine design.

2 Methods

2.1 Porcine reproductive and respiratory syndrome virus

Porcine reproductive and respiratory syndrome virus (PRRSV) is the causative agent of porcine reproductive and respiratory syndrome (PRRS), an economically damaging disease of domestic swine (Tian *et al.*, 2007; Zimmerman *et al.*, 1997). PRRSV is in the order *Nidovirales*, family *Arteriviridae*, along with equine arteritis virus, lactate dehydrogenase elevating virus, and simian hemorrhagic fever virus (Conzelmann *et al.*, 1993; Meulenberg *et al.*, 1993). The virus has a short (approximately 15 kb) single stranded, positive-sense RNA genome that encodes at least nine open reading frames (ORFs) (Conzelmann *et al.*, 1993; Meulenberg *et al.*, 1993). Despite significant research, including a focus on the virus's genetic diversity, it has remained difficult to control (Shi *et al.*, 2010; Shi *et al.*, 2010). Vaccination, in particular, has variable efficacy and has failed to reduce transmission or to reduce clinical signs (see Thanawongnuwech and Suradhat, 2010 and references therein).

One likely cause for the difficulty in controlling PRRSV is its genetic and antigenic variability. Type 1 PRRSV (or European origin) differs at approximately 50% of nucleic acid positions from Type 2 PRRSV (of North American origin), implying a high degree of divergence between continents (Forsberg *et al.*, 2001; Forsberg *et al.*, 2002; Murtaugh *et al.*, 2010). Further, within these viral types, PRRSV exists as a spectrum of genotypes, displaying considerable heterogeneity with very little geographic population substructure (Shi *et al.*, 2010; Stadejek *et al.*, 2006). Additionally, PRRSV varies between farms, herds, among animals within herds, and even within individual animals (Goldberg *et al.*, 2000; Goldberg *et al.*, 2003). Extensive nucleotide sequence data are available for the viral ORF5 gene, which is used commonly for phylogenetic analyses and source tracking (e.g. Shi *et al.*, 2010). Consequently, ORF5 diversity acts as a marker of PRRSV variability.

2.2 Creation of a core PRRSV sequence dataset

All PRRSV ORF5 sequences from isolates in Genbank, the National Center for Biotechnology Information's on-line sequence repository (http://www.ncbi.nlm.nih.gov/genbank/), and the now defunct PRRSV Database were downloaded in December 2010 (see Supplementary Material for sequence files). These sequences were comprised of 25,265 worldwide samples, three live-attenuated vaccine strains (MLV, ATP, and PrimePac), and one laboratory attenuated strain (Abst-1). Most viruses were from the United States of America, but viruses from Canada, Mexico, China, Korea, Japan, Thailand, Austria, Denmark, Italy and Poland were also present.

As a first "cleaning" step to ensure we only included wild-type variants, we removed sequences that were described with any of the following terms: patent, vaccine, attenuated, attenuation, clone. The remaining sequences were then aligned using default settings in MUSCLE v3.6 (Edgar, 2004), with manual correction in Mesquite (Maddison and Maddison, 2009). At the same time, we removed all type 1 variants, thus restricting our analyses to type 2 PRRSV. The remaining aligned sequences underwent a redundancy analysis within the computer program Mothur (Schloss *et al.*, 2009), and identical sequences were removed. In addition, to avoid bias in subsequent network and phylogenetic analyses (Lemmon *et al.*, 2009), we removed sequences with any nucleotide base ambiguities. The next filtering step was to remove poor quality data using three criteria: first, sequences were removed if they did not have a start and a stop codon; second, sequences were removed if more than 1% of the gene sequence was missing; and third, a sequence was removed if it produced an insertion or deletion in the alignment that was biologically implausible (e.g. causing a frame shift and introducing premature stop codons) or occurred in fewer than 0.005% of sequences. The alignment was then screened for evidence of recombination using 3Seq (Boni *et al.*, 2010; Boni *et al.*, 2007), with all putative recombinants subject to secondary screening and validation using RDP3 (Martin *et al.*, 2010): if identified by three or more methods within the software, sequences were removed from

Table 1. Network ranking of select porcine reproductive and respiratory syndrome virus reference sequences and the top 5 ranked sequences out of 9383.

	Nucleotide network					Amino acid network					Evolutionary distance network				
	k	*CC*	*B*	*w*	PR	*k*	*CC*	*B*	*w*	PR	*k*	*CC*	*B*	*w*	PR
Reference variants:															
VR-2332 (PRU87392)	2564	1991	2845	3423	3361	3121	2650	2951	2714	3439	4099	1524	727	4163	3625
RespPRRS (AF066183)	3915	2857	4478	4279	4875	5665	4627	5447	3814	5624	5406	2807	1520	4441	4989
JA-142 (AY424271)	595	571	530	780	554	31	39	91	79	31	445	511	743	485	534
MN-184 (EF488739)	3605	4619	1881	6497	1367	3352	4303	4484	6080	3097	3107	1259	997	6327	1459
PrimePac (AF066384)	586	489	533	701	583	202	221	206	468	133	136	136	46	482	94
China "atypical" (EF112446)	2732	1909	3387	2467	4248	5322	4120	3153	3295	5317	1448	2354	2866	1447	1911
Top 5 ranked variants[1]:															
DQ475317	1	1	5	1	1	12	12	59	4	11	3	3	6	1	4
DQ477864	2	2	26	2	2	16	13	81	7	15	10	10	26	6	15
PRU66382	6	6	75	4	11	20	20	138	11	19	4	4	10	2	5
DQ477862	25	22	174	18	33	21	21	141	12	21	7	7	16	4	10
AB175723	24	20	147	16	28	25	24	149	17	26	5	5	3	3	7

Ranking statistics include degree (*k*), closeness centrality (*CC*), betweenness (*B*), eigenvector centrality (*w*), and PageRank (PR). Variants are referred to using common name or GenBank accession numbers: rankings are on a scale of 1 to 9383.

1. The top 5 sequences are a selection of those variants that were found ranked in the top 100 in all three networks using degree (*k*) as the ranking metric.

further analysis. For reference, even though some were eliminated in our filtering process, we included several common reference strains: VR2332, RespPRRS, China "atypical," JA-142, MN-184, PrimePac, and ATP viral variants. This process resulted in a set of 9383 non-identical ORF5 sequences that represent the full extent of known genetic diversity of wildtype type 2 PRRSV.

2.3 Construction of PRRSV networks

We constructed three undirected, unweighted networks using adjacency matrices derived from 1) nucleotide sequence differences, 2) amino acid sequence differences, and 3) evolutionary distances. We generated the nucleotide difference adjacency matrix using PAUP* 4.0 (Swofford, 2003), with cells containing values representing uncorrected numbers of nucleotide differences between pairs of aligned sequences. Similarly, we generated a second adjacency matrix by calculating the uncorrected number of changes between amino acid sequences using program R (Team, 2010) with the APE (Paradis, *et al.*, 2004) and SeqinR packages (Charif and Lobry, 2007). Our third adjacency matrix represented evolutionary distances. Values in cells were patristic distances (distances along the branches of a phylogenetic tree) calculated by PAUP* 4.0 (Swofford, 2003): we built the phylogeny used to calculate these distances using RAxML 7.2.8 (Stamatakis, 2006; Stamatakis *et al.*, 2008) maximum likelihood analyses on the CIPRES Science Gateway (Miller *et al.*, 2010). We implemented a General-Time-Reversible nucleotide substitution model with 4 categories of gamma distributed rate heterogeneity and a proportion of invariant sites (GTR + Γ + I) model of molecular evolution.

The three matrices represent different types of biological relationships among viral sequences. The first makes the fewest assumption about the structure or function of each variant, the second acknowledges that immunological cross-protection and virulence operate at the level of proteins, and the third acknowledges that evolutionary history might be an important constraint on viral immunobiology. Although these matrices contain quantitative information on relationships between viral variants, to simplify theory and measurement, we dichotomized each using the sna package in R (Butts, 2008). This approach is appropriate given that the relationship under study, in this instance genetic distance, is stable and the values taken across pairs are constrained (Butts, 2009).

2.4 Topological indices used to describe positional importance

To rank sequences, we used a range of indices that are dependent on the characteristics of the focal node but also include information about the overall topology of the viral network.

The most fundamental metric, the degree (or connectivity; *k*) of a sequence, describes the number of links a sequence makes with other sequences. Those sequences with high degrees are those that are connected highly with all other sequences. This index is useful for identifying viral variants that have the most direct connections to other variants. Other, secondary, clustering techniques further detail the relative importance of each sequence to "sequence diversity space" and include: closeness; betweenness; eigenvector centrality; and PageRank.

Closeness provides a measure that describes the position, in terms of distance, from a focal sequence to all other sequences and is measured as:

$$\left(CC(v) = \frac{1}{\sum_{u \in V} d(v,u)} \right) \quad (1)$$

Intuitively, closeness provides an index of the extent to which a given sequence has short paths to all other sequences. This is a measure of whether a sequence is in the "middle" of our defined sequence space; smaller values can be interpreted to indicate sequences of greater importance (*i.e.* the removal of such a sequence will affect a majority of other sequences).

Betweenness, first introduced by Freeman (Freeman, 1977), is a measure that describes the centrality of a sequence, provided as a frequency with which a sequence is located on the shortest path between all other sequences. First, let $\delta_{st}(v)$ denote the pairwise dependency, or the fraction of shortest paths between s and t that pass through v. Following this, the betweenness of virus v is defined as:

$$B(v) = \sum_{s \neq v \neq t \in V} \delta_{st}(v) \quad (2)$$

Conceptually, high-betweenness sequences lie on a large number of nonredundant shortest paths between other vertices; they may be thought of as "bridges." This metric has been used to assess node importance in other biological networks (Yamada and Bork, 2009), to describe sexual contact networks (Liljeros *et al.*, 2001), and to describe the structure of the worldwide air transportation network (Guimera *et al.*, 2005).

Eigenvector centrality, as described by Bonacich (Bonacich, 1972), determines node scores that correspond to the leading eigenvector of the sequence adjacency matrix. These scores arise from a reciprocal process in which the centrality of each sequence is proportional to the sum of the centralities of those sequences to which it is connected. The defining equation is:

$$\lambda w = \boldsymbol{A} w \quad (3)$$

where $\boldsymbol{A}$ is the adjacency matrix of the graph, lambda is the constant (the eigenvalue), and w is the eigenvector. In general, sequences that have high eigenvector centralities are those that are connected to many other sequences that are, in turn, connected to many others. This metric has been widely used to identify key network components, including connectivity within the human brain (Lohmann *et al.*, 2010).

PageRank, in the context in which we use it, rates viral variants as important if they receive links from other viruses that are in turn also rated as important. Each virus i is assigned an importance, and each link a_{ij} (exiting virus i to enter virus j) carries an equal fraction of the importance value: the importance of the virus is then the sum of the importance assigned to its incoming connections. This recursive problem is solved by building a matrix $\boldsymbol{S}$ in which each element represents the fraction of importance assigned to a link: subsequently, importance is solved by computing the eigenvector associated with the dominant eigenvalue of the matrix (Bryan and Leise, 2006).

Finally, we analyze the community structure of the networks by computing the modularity Q given by (Clauset *et al.*, 2004):

$$Q = \sum_{i=1}^{m} (e_{ii} - a_i^2) \quad (4)$$

where m is the number of modules inside the network, e_{ii} is the fraction of links in the network connecting nodes of the same community i, and a_i is the fraction of links that have one or two ends inside community i. The larger the fraction of links inside each community, the higher the value of Q. In this way, modularity Q can be taken as a reference parameter in order to find the optimal community divisions based upon the topology of the network. These algorithms have been used extensively to address interaction patterns within varied systems, including but not limited to American football (Girvan and Newman, 2002), and plant-pollinator dynamics (Olesen *et al.*, 2007).

The critical concept to note in the above analytical approach is that the position a sequence takes within "sequence diversity space" is determined by its similarity to other sequences, and the similarity values of the other sequences to which it is connected. In using these algorithms, we account for the underlying structure of the genetic relationships between sequences and can objectively rank each sequence and its contribution to PRRSV genetic diversity.

2.5 Relationship between indices

We calculated the indices described above for every individual viral variant in each PRRSV network. We ranked the importance of all variants according to their values of a particular index. Thus, in total there were 5 rankings corresponding to 5 different indices (*i.e.*, *k*, *CC*, *B*, *w*, PR) for each of three networks. We also examined whether or not different indices provided the same information about the relative importance of viral variants. First, to show the relationship between indices, we conducted hierarchical clustering on the rankings of the viral variants by the values of the 5 indices: first, we standardized the values within indices, and then we used similarity measures appropriate for interval data. Second, we used Kendall rank correlations between all pairs of indices; and third, we applied a Kendall Concordance test to determine whether the observed ranks were significantly different from those expected by chance when independently ranking 9383 nodes by 5 indices.

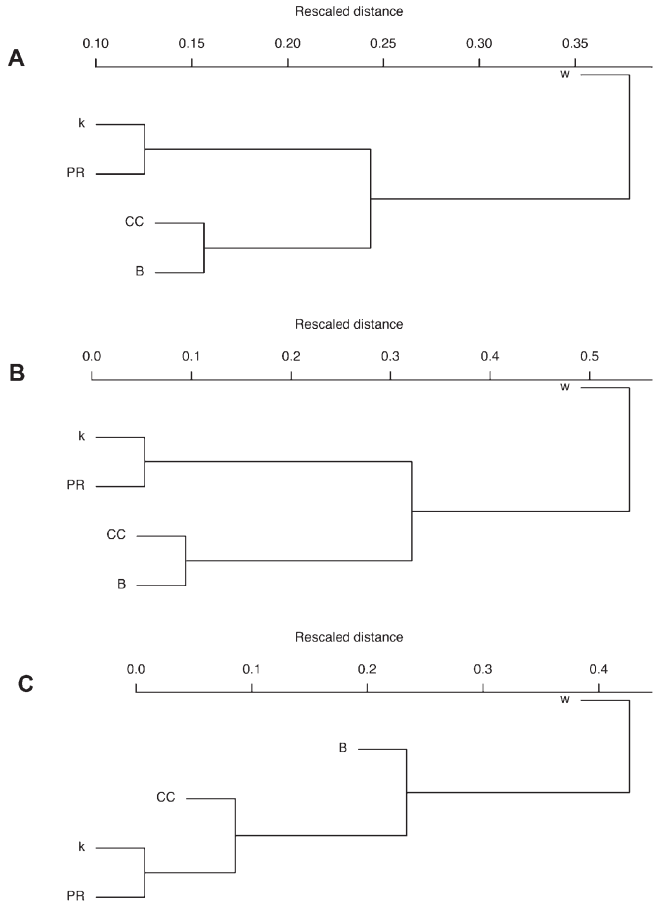

Figure 1. Hierarchical clustering dendrogram showing the similarities between rankings of nodes by the 5 network indices. The indices include, degree (*k*), closeness centrality (*CC*), betweenness (*B*), eigenvector centrality (*w*), and PageRank (PR), and measure various aspects of the relative importance of viral variants. We show clustering of rankings based upon the network derived from nucleotide differences (A), amino acid difference (B), and evolutionary distance (C).

2.6 Design and evaluation of T-cell vaccine candidates

We compared the putative epitope coverage of known PRRSV variants by generating vaccine "mosaic" proteins using a promising design algorithm from the HIV literature (Fischer *et al.*, 2007). To assess whether high-ranking variants achieved better epitope coverage than low-ranking variants, we used the Vaccine Epitope Coverage Assessment online tool (Epicover; Thurmond *et al.*, 2008) to compute how well the top and bottom ranked proteins from each network covered putative epitopes across the full dataset of 9383 wild-type PRRSV sequences. To further explore the association between network ranking and epitope coverage, we ranked sequences according to each centrality metric, selected every 20th ranked sequence and calculated its epitope coverage. We quantified the strength of association using MIC (Reshef *et al.*, 2011), and Pearson's correlation coefficient (ρ). To

compare our method to the mosaic protein approach (Thurmond *et al.*, 2008), we used the Mosaic Vaccine Designer online tool to generate *in silico* a single mosaic protein from the 9383 wild-type sequences. We then used Epicover to measure the coverage of this single mosaic across the 9383 wild-type sequences.

3 Results

The viral network derived from nucleotide differences included 9383 variants, and had 88,040,689 potential links of which 11,044,560 were realized, resulting in a connectance of 0.125. The amino acid derived network included 7889 variants, and had 62,236,321 potential links of which 8,636,124 were realized, resulting in a connectance of 0.139. The remaining network, derived from evolutionary distances, included 9383 viral variants with 11,005,086 realized links, with a resultant value of connectance of 0.125. Measured in this way, connectance is the average fraction of viral variants to which an individual variant in the network is connected, based upon similarity. Twenty-two viral variants were ranked in the top 100 sequences in all three networks, and 51 variants were ranked in the bottom 100 sequences in all three networks.

All three viral networks displayed cumulative degree distributions that were different from what would be expected if the link distribution were random. Each network had data consistent with an exponential degree distribution ($P(k) \sim \exp(-\gamma k)$: nucleotide $AICc$ = –8449.9; evolutionary distance $AICc$ = –7637.8; amino acid $AICc$ = –6191.2). The data were not well represented by the power-law ($P(k) \sim k^{-\gamma}$: nucleotide $AICc$ = –1246.5; evolutionary distance $AICc$ = –1417.4; amino acid $AICc$ = –816.1), or truncated power-law ($P(k) \sim k^{-\gamma} \exp(-k/k_x)$: nucleotide $AICc$ = –8171.8; evolutionary distance $AICc$ = –1410.9; amino acid $AICc$ = – 5847.8). The identity of the best-fit model is secondary to our data departing from a power-law distribution; this suggests that very highly connected variants are more rare than would be expected if the networks were built using a scale-free distribution to describe the number of interactions per virus.

There were differences in the rankings of viral variants by the five different indices (for select sequence rankings, see Table 1). Qualitatively, there were differences in the information provided by the indices rankings of variants based upon the hierarchical clustering dendrograms (Figure 1). These dendrograms were insensitive to the clustering method. The more similar indices are in their ranking of the variants, the shorter the distance between the branch representing them, and the first shared branching point. Nevertheless, the rank orderings of variants were statistically significantly correlated between indices (*i.e.*, a variant ranked highly by degree (k) tended also to receive a high ranking as measured by eigenvector centrality; Table 2). This was further validated using Kendall's W, which showed that the ranking provided by each index is significantly similar across each viral network: nucleotide differences ($W = 0.82$, $\chi^2 = 38421$, $P < 0.01$); amino acid differences ($W = 0.85$, $\chi^2 = 33495$, $P < 0.01$); and evolutionary distance ($W = 0.798$, $\chi^2 = 37447$, $P < 0.01$).

Our ranking approach identified a single wild-type sequence that achieved coverage of putative epitopes essentially equal to that of a "mosaic" protein generated using all 9383 sequences (Figure 2). Our top ranked variant (DQ475317) within the nucleotide network achieved 48.5% coverage, whereas the mosaic protein achieved 49%. Top ranking variants consistently covered more putative epitopes than those that were ranked poorly: the top ranking amino acid variant achieved 40% coverage; and the most important variant in the evolutionary distance network achieved 41% coverage (Figure 2). The 22 variants ranked in the top 100 of all three networks achieved, on average, 43% (± 1.7% standard deviation) coverage of putative epitopes derived from known PRRSV ORF5 variants. By contrast, the 51 variants ranked in the bottom 100 of all three networks achieved, on average, only 26% (± 3.5% standard deviation). This difference was statistically significant ($t = 28.20$; 69 degrees of freedom; $P < 0.0001$). There was a significant ($P < 0.0001$) relationship between network ranking and putative epitope coverage when variants were ranked by: degree (Figure 3: $\rho = -0.76$; MIC = 0.63); closeness ($\rho = -0.78$; MIC = 0.61); betweenness ($\rho = -0.66$; MIC = 0.52); PageRank ($\rho = -0.65$; MIC = 0.47); and eigenvector centrality ($\rho = -0.59$; MIC = 0.53).

Table 2. Kendall rank correlation coefficients, t, between the rank ordering by the 5 indices of the 9383 nodes in the porcine reproductive and respiratory syndrome virus network.

		k	*CC*	*B*	*w*	PR
Nucleotide	*k*	1.0000	0.7084	0.5974	0.7273	0.7339
	CC		1.000	0.6655	0.6719	0.5633
	B			1.0000	0.4531	0.6993
	w				1.0000	0.4690
	PR					1.0000
		k	*CC*	*B*	*w*	PR
Amino acid	*k*	1.0000	0.8151	0.6348	0. 5876	0. 9427
	CC		1.0000	0. 7139	0. 6589	0. 7872
	B			1.0000	0. 5005	0. 6464
	w				1.0000	0. 5342
	PR					1.0000
		k	*CC*	*B*	*w*	PR
Evolutionary	*k*	1.0000	0.6656	0.5719	0.6224	0.8268
	CC		1.0000	0.7552	0.5612	0.6957
	B			1.0000	0.4087	0.6571
	w				1.0000	0.4576
	PR					1.0000

Correlation coefficients for the nucleotide derived network (Nucleotide), amino acid similarity network (Amino acid), and evolutionary distance network (Evolutionary).
All correlations were significant at $P < 0.01$ level.

4 Discussion

Our results demonstrate that a dataset of 9383 sequences could be ranked according to three matrices and five network centrality metrics. Using interaction networks derived from sequence nucleotide difference, amino acid difference, and evolutionary distance we were able to score individual PRRSV sequences based on their relative importance to the viral network. This yielded 22 sequences that were represented in the top 100 sequences in all matrices, and provided objective criteria for selecting viral variants for polyvalent vaccine design. We identified a single wild-type sequence that covered 48% of

putative epitopes derived from the known diversity of PRRSV. Further, we document a relationship between our ranked variants and putative epitope coverage. In summary, we have demonstrated that it is possible to use network statistics to rank viruses with respect to their overall "importance" within a network, and that the resulting rankings can help guide the selection of viruses for inclusion in vaccines.

The key insight provided by our work is that it is possible to construct *de novo* a virus network from sequence data and objectively quantify the positional importance of viral variants within such a network. For vaccine development, it is clear that choosing a single variant as a type-strain is problematic, and vaccination with all known viral variants is not technically feasible. Consequently, there is need for techniques that "filter" highly diverse genomic datasets, and select representative sequences for further study. We suggest that the calculation of network indices that describe the relative importance of viral variants may inform decisions about type-strain selection and polyvalent vaccine development. Though this approach is flexible, and can be implemented using small and less diverse genomic data, more traditional approaches may be more appropriate in these situations. Of the indices we measured, eigenvector centrality provided a robust prediction of variant importance. This metric provides a measure of a viral variant's global position, rather than its sheer number of connections (i.e., *k*), and describes sequences that are highly connected and fall within densely populated substructures of the viral network. More generally, our case study of the PRRSV system serves as an illustration of the potential of these techniques in the objective selection of important viral variants for additional study.

We demonstrate that it is possible to select only highly ranked sequences and achieve equal or better coverage of putative T-cell epitopes than using low ranked sequences, and that top-ranked sequences achieve nearly as good T-cell epitope coverage as artificial mosaic proteins derived computationally from the full set or sequences. The latter conclusion is important because mosaic proteins derived from the aforementioned algorithm may, when synthesized, have biological properties that render them unsuitable for inclusion in actual vaccines (Fischer *et al.*, 2007; Nickle *et al.*, 2007; Rolland *et al.*, 2007). Our method achieves nearly equal coverage by identifying the "most important" wild-type viruses. These viruses, if they can be found in reference collections, would make excellent candidates for subsequent biological studies, with the goal of modifying their immunomodulatory properties and virulence to the point that they would be suitable vaccine candidates. Our results predict that such vaccines would be biologically viable while also conferring broadly cross-protective immunity.

Our approach is likely to be useful given the extraordinarily large amounts of data currently available as a result of next-generation sequencing technologies (Holmes and Grenfell, 2009). Although existing computational methods to account for viral diversity are powerful when investigating small numbers of sequences, the size of potential tree-space and computational complexity increases faster than exponentially with the number of sequences (Holmes and Grenfell, 2009; Rambaut *et al.*, 2008). This forces use of biology-independent filters (e.g. limiting analysis to geographic regions, considering ease-of-access to samples, or using well characterized genes) to produce a dataset that is manageable. At the

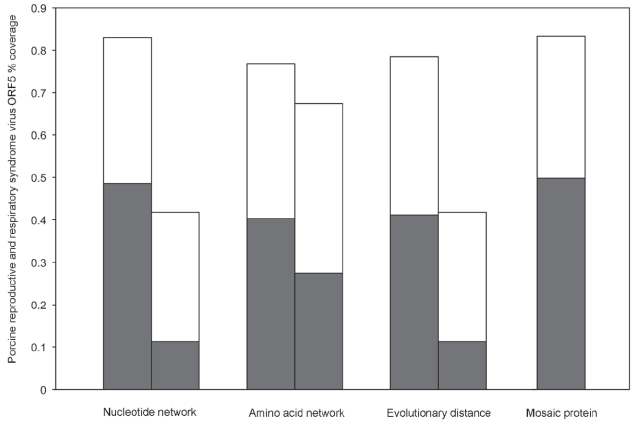

Figure 2. Overall coverage of T-cell epitopes by vaccine candidates. Coverage of nine-mers in porcine reproductive and respiratory virus using either the top-ranked (left bar) or bottom-ranked (right bar) viral variant using degree (*k*) as the ranking metric. Exact (gray-shaded), and 8 of 9 (one-off; white-shaded) coverage was computed for four test situations: a) the top- and bottom-ranked sequences from the nucleotide difference network; b) the top- and bottom-ranked sequences from the amino acid difference network; c) the top- and bottom-ranked sequences from the evolutionary distance network; and d) a mosaic protein generated from all 9383 sequences. The "mosaic" sequence was generated by the genetic algorithm described in (Fischer *et al.*, 2007); all sequences, including the mosaic protein, were compared against all nine-mers generated from 9383 unique PRRSV variants.

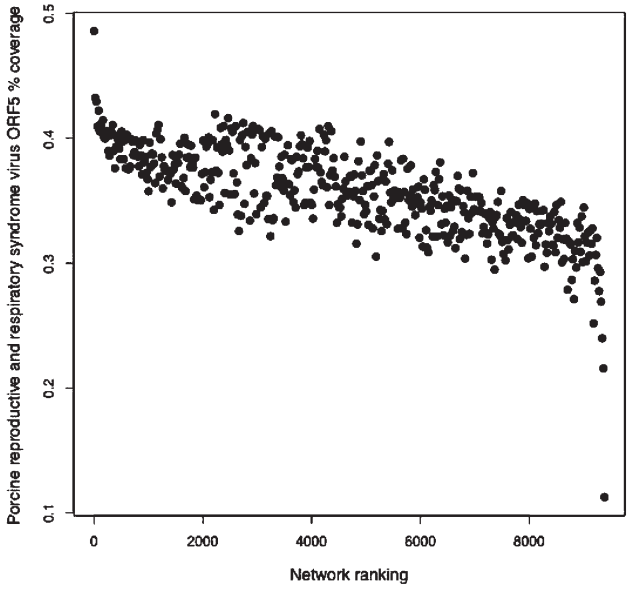

Figure 3. Association between network ranking and coverage of T-cell epitopes. Coverage of nine-mers in porcine reproductive and respiratory syndrome virus by a ranked subset of ORF5 viral variants. Viral variants were ranked using degree (*k*), derived from the nucleotide difference network. The putative epitope coverage of every 20th sequence was assessed using the Vaccine Epitope Coverage Assessment online tool (Thurmond *et al.*, 2008).

other extreme, there is a growing recognition that the use of consensus sequences, though computationally efficient, may not capture the evolutionary dynamics or phenotypic variation necessary to generate immunogenetic and broadly effective vaccines (Aaskov *et al.*, 2006; Wahala *et al.*, 2010). Our approach provides a way of "distilling" this overwhelming diversity and focusing subsequent analyses on a small and objectively chosen subset of viral variants. Though biases may be introduced to the genetic networks due to limitations in source data, our approach is fully adaptable to sample sets with different or no biases, such as geographically restricted populations, or to virulent populations of viruses. Notably, our network analyses were conducted and completed in 72 hours on a standard 2 GHz desktop computer with 8 GB of RAM. This suggests that our framework is not prohibitive from a computational perspective.

We suggest that network centrality measures, in differentiating viral variants according to how influential they are, can inform the selection of type-strains for vaccine design. The degree, k, of a node is the first indication of its centrality; intuitively a well-connected sequence with a high degree will be traversed by a proportionally larger number of short paths than those sequences with low degrees. However, it is important to distinguish this metric as a local index; it does not take in to account the importance of the neighboring nodes. This suggests that a variant, though well connected, may fall within a relatively sparse area of the network, and may not account for a large amount of genetic information. To overcome this restriction, the other centrality indices quantify positional importance using global measures accounting for genetic distance (*i.e.*, the number of edges on the shortest path connecting i and j). Eigenvector centrality is particularly useful; previous research has suggested that nodes characterized by higher eigenvector centrality play a more influential role in the structure and dynamics of the networks in which they are nested (Allesina and Pascual, 2008; Allesina and Pascual, 2009). In our viral networks, those variants that have high eigenvector centrality scores are those that contribute disproportionately to global viral diversity, acting as core variants that are likely to prove useful in the development of vaccines, diagnostic tests, and epidemiological surveillance strategies.

Another potential metric that provides insight into variant selection is the betweenness of a sequence. One biological interpretation of this index is that it quantifies the probability that node i represents an intermediate step in the evolution between one sequence to another. In Table 2, we demonstrated a positive correlation between sequence degree and sequence betweenness, confirming the intuitive idea that sequences of high degree are those with high betweenness: the larger the number of neighbors a particular sequence has, the greater the probability that the sequence will fall on a shortest path between other sequences. Deviations from this correlation indicate "gatekeeper" sequences, *i.e.* sequences that represent unique pathways to subsections within the network. Thus, using this metric we are able to quantify a variant's importance in two ways: firstly, a variant that has a high betweenness score provides a measure of the amount of genetic information which the variant controls; and secondly, if a variant falls outside of the correlation we document, it represents a unique sequence that acts as a bridge to distinct modules within the viral network.

Our rankings of PRRSV reference variants reveal that many studies of PRRSV biology to date have focused on reference strains that are not central to the diversity of the virus. Typically, such reference strains are selected because of ease-of-access, publication history, and evaluation of whether the variant represents a particular geographic location or biological property (e.g. virulence, immunogenecity). Our analysis shows that selection of such strains may limit the generality of subsequent conclusions. Additionally, current phylogenetic methods that attempt to classify type-strains (Shi *et al.*, 2010), though analytically rigorous, carry a series of assumptions that introduce biases into variant selection. This is particularly true in PRRSV, where incomplete geographic sampling and a reliance on a gene for inference that is subject to strong immunological pressure has likely resulted in phylogenetic trees that are sub-optimal (Murtaugh *et al.*, 2010). Given these weaknesses and the failings of current classification strategies (see Murtaugh *et al.*, 2010 and references therein), we suggest that defining variant importance via network statistics is a promising and productive avenue. Indeed, this approach is best suited for PRRSV and other viruses with a large amount of genetic diversity and little phylogenetic substructure. In the case of viruses that sort into clear monophyletic clades, it may be impossible to select one variant that represents diversity across clades. In this situation, we suggest a ranking system based on specific scenarios, such as geographically restricted populations, virulent populations of viruses, or as in our case study, the selection of known subtypes. An alternate solution is to select sequences from within defined network groups using modularity-based clustering heuristics (see Newman, 2011 and references therein).

One strength of our approach is that it makes minimal assumptions about viral biology in assigning the rankings of individual sequences. Underlying matrices of amino acid similarity, for example, do not assign immunological importance to specific protein regions. However, our approach is very flexible for incorporating biological information as it is generated. For example, our approach could easily be applied to specific protein regions that future studies identify as being immunologically important. A weakness, however, is that the representational framework is restrictive given the dyadic nature of interactions; if our goal is to develop an effective vaccine, this process may omit components of what is a complex system of interactions. Indeed, research in the HIV field has revealed that induction of protective immunity is a function of robust mucosal immunity, high avidity and polyfunctional T-cells, broad neutralizing antibodies and optimized vaccine delivery methods (Wijesundara *et al.*, 2011). This is in addition to the difficulties associated with viral antigenic diversity (Fischer *et al.*, 2007). However, decades of research have been unsuccessful in the development of broadly effective vaccines for numerous highly variable pathogens. Thus, our approach to representing variability as a network, and ranking variants for further study, may prove to be a judicious strategy that aids in the development of efficacious vaccines and targeted control.

Important tasks for future studies include: (i) improving our knowledge of the biological meanings of different centrality indices; (ii) extending and improving the ranking system through comparative analyses of similar viruses; and (iii) testing whether the high ranking individual variants are better in inducing cross-protective immunity than those with

lower rankings *in vivo*. Further, a potential and as yet unrealized application of our approach is the targeted design of vaccines that modify virulence. Our case study incorporated all sequence variation, but it is possible to select for analysis only sequences of high virulence. Vaccines rarely provide full protection from disease, and those vaccines that are ineffective may unintentionally select for increased pathogen virulence (Gandon *et al.*, 2001). We posit that, given the high likelihood that any vaccine against a highly variable RNA virus such as PRRSV would offer less-than-perfect protection, an alternative goal of vaccination could be to drive viral evolution intentionally towards benignness. Our ranking method could provide rational criteria for selecting among highly virulent types to include in vaccines that may, over time, drive the population of viruses towards lower virulence (Ewald, 2004).

Acknowledgments

This work was supported by the PRRS CAP, United States Department of Agriculture NIFA Award 2008-55620- 19132. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Aaskov, J., *et al.* (2006) Long-term transmission of defective RNA viruses in humans and Aedes mosquitoes, *Science*, **311**, 236-238.

Albert, R. and Barabasi, A. (2002) Statistical mechanics of complex networks, *Rev. Mod. Phys.*, **74**, 47-97.

Allesina, S. and Pascual, M. (2008) Network structure, predator-prey modules, and stability in large food webs, *Theor. Ecol.*, **1**, 55-64.

Allesina, S. and Pascual, M. (2009) Googling Food Webs: Can an Eigenvector Measure Species' Importance for Coextinctions?, *PLoS Comput. Biol.*, **5**.

Azmi, A. S., *et al.* (2010) Proof of concept: network and systems biology approaches aid in the discovery of potent anticancer drug combinations, *Mol. Cancer Ther.*, **9**, 3137-3144.

Barabasi, A-L., Gulbahce, N. and Loscalzo, J. (2011) Network medicine: A network-based approach to human disease, *Nature Rev. Genet.*, **12**, 56-68.

Barouch, D. H. (2008) Challenges in the development of an HIV-1 vaccine, *Nature*, **455**, 613-619.

Barouch, D. H., *et al.* (2010) Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys, *Nat. Med.*, **16**, 319-323.

Belshaw, R., *et al.* (2008) Pacing a small cage: Mutation and RNA viruses, *Trends Ecol. Evol. (Amst.)*, **23**, 188-193.

Bonacich, P. (1972) Factoring and Weighting Approaches to Status Scores and Clique Identification, *J. Math. Sociol.*, **2**, 113-120.

Boni, M.F., *et al.* (2010) Guidelines for identifying homologous recombination events in influenza A virus, *PLoS ONE*, **5**, e10434.

Boni, M. F., Posada, D. and Feldman, M. W. (2007) An exact nonparametric method for inferring mosaic structure in sequence triplets, *Genetics*, **176**, 1035-1047.

Bryan, K. and Leise, T. (2006) The $25,000,000,000 eigenvector: The linear algebra behind google, *SIAM Rev.*, **48**, 569-581.

Butts, C. (2008) Social network analysis with sna, *J. Stat. Softw.*, **24**, 1-51.

Butts, C. T. (2009) Revisiting the foundations of network analysis, *Science*, **325**, 414- 416.

Charif, D. and Lobry, J. R. (2007) Seqin{R} 1.0-2: A contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. In Bastolla, U., *et al.* (eds), *Structural approaches to sequence evolution: Molecules, networks, populations*. Springer Verlag, New York, pp. 207-232.

Clauset, A., Newman, M.E.J. and Moore, C. (2004) Finding community structure in very large networks, *Phys. Rev. E*, **70**, 66111.

Cleaveland, S., Laurenson, M. K. and Taylor, L. H. (2001) Diseases of humans and their domestic mammals: Pathogen characteristics, host range and the risk of emergence, *Phil. Trans. R. Soc. B*, **356**, 991-999.

Conzelmann, K. K., *et al.* (1993) Molecular Characterization of Porcine Reproductive and Respiratory Syndrome Virus, a Member of the Arterivirus Group, *Virol.*, **193**, 329-339.

Crandall, K. A. (1999) *The evolution of HIV*. Johns Hopkins University Press, Baltimore.

De Groot, A. S., *et al.* (2005) HIV vaccine development by computer assisted design: The GAIA vaccine, *Vaccine*, **23**, 2136-2148.

Domingo, E. (2002) Quasispecies theory in virology, *J. Virol.*, **76**, 463-465.

Doria-Rose, N. A., *et al.* (2005) Human immunodeficiency virus type 1 subtype B ancestral envelope protein is functional and elicits neutralizing antibodies in rabbits similar to those elicited by a circulating subtype B envelope, *J. Virol.*, **79**, 11214-11224.

Drake, J. and Holland, J. (1999) Mutation rates among RNA viruses, *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 13910-13913.

Edgar, R. C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, **32**, 1792-1797.

Ewald, P. W. (2004) Evolution of virulence, *Infect. Dis. Clin. North Am.*, **18**, 1-15.

Fischer, W., *et al.* (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants, *Nat. Med.*, **13**, 100-106.

Forsberg, R., *et al.* (2001) A molecular clock dates the common ancestor of European-type porcine reproductive and respiratory syndrome virus at more than 10 years before the emergence of disease, *Virol.*, **289**, 174-179.

Forsberg, R., *et al.* (2002) The genetic diversity of European type PRRSV is similar to that of the North American type but is geographically skewed within Europe, *Virol.*, **299**, 38-47.

Freeman, L. (1977) A set of measures of centrality based on betweenness, *Sociometry*, **40**, 35–41.

Gandon, S., *et al.* (2001) Imperfect vaccines and the evolution of pathogen virulence, *Nature*, **414**, 751-756.

Gao, F., *et al.* (2004) Centralized immunogens as a vaccine strategy to overcome HIV-1 diversity, *Expert Rev. Vaccines*, **3**, S161-S168.

Gao, F., *et al.* (2005) Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group m consensus envelope glycoprotein., *J. Virol.*, **79**, 1154-1163.

Gaschen, B., *et al.* (2002) Diversity considerations in HIV-1 vaccine selection, *Science*, **296**, 2354-2360.

Girvan, M. and Newman, M. (2002) Community structure in social and biological networks, *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 7821-7826.

Goldberg, T. L., *et al.* (2000) Genetic, geographical and temporal variation of porcine reproductive and respiratory syndrome virus in Illinois, *J. Gen. Virol.*, **81**, 171- 179.

Goldberg, T. L., *et al.* (2003) Quasispecies variation of porcine reproductive and respiratory syndrome virus during natural infection, *Virol.*, **317**, 197-207.

Guimera, R., *et al.* (2005) The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles, *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 7794-7799.

Holmes, E. C. (2009) The Evolutionary Genetics of Emerging Viruses, *Ann. Rev. Ecol. Evol. Syst.*, **40**, 353-372.

Holmes, E. C. and Grenfell, B. T. (2009) Discovering the phylodynamics of RNA viruses, *PLoS Comput. Biol.*, **5**, e1000505.

Jones, K. E., *et al.* (2008) Global trends in emerging infectious diseases, *Nature*, **451**, 990-993.

Katze, M. G., *et al.* (2008) Innate immune modulation by RNA viruses: Emerging insights from functional genomics, *Nat. Rev. Immunol.*, **8**, 644-654.

Korber, B. T., Letvin, N. L. and Haynes, B. F. (2009) T-cell vaccine strategies for human immunodeficiency virus, the virus with a thousand faces, *J. Virol.*, **83**, 8300-8314.

Lemmon, A. R., *et al.* (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference, *Syst. Biol.*, **58**, 130-145.

Liljeros, F., *et al.* (2001) The web of human sexual contacts, *Nature*, **411**, 907-908.
Lohmann, G., *et al.* (2010) Eigenvector Centrality Mapping for Analyzing Connectivity Patterns in fMRI Data of the Human Brain, *PLoS ONE*, **5**, e10232.
Loscalzo, J., Kohane, I., and Barabasi, A-L. (2007) Human disease classification in the postgenomic era: A complex systems approach to human pathobiology, *Mol. Syst. Biol.*, **3**, 124.
Maddison, W. P. and Maddison, D. R. (2009) Mesquite: A modular system for evolutionary analysis.
Martin, D. P., *et al.* (2010) RDP3: A flexible and fast computer program for analyzing recombination, *Bioinformatics*, **26**, 2462-2463.
Meulenberg, J.J.M., *et al.* (1993) Lelystad Virus, the Causative Agent of Porcine Epidemic Abortion and Respiratory Syndrome (PEARS), Is Related to Ldv and Eav, *Virol.*, **192**, 62-72.
Miller, M. A., Pfeiffer, W. and Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE)*. New Orleans, pp. 1-8.
Murtaugh, M. P., *et al.* (2010) The ever-expanding diversity of porcine reproductive and respiratory syndrome virus, *Virus Res.*, **154**, 18-30.
Newman, M.E.J. (2012) Communities, modules and large-scale structure in networks, *Nature Physics*, **8**, 25-31.
Nickle, D. C., *et al.* (2007) Coping with viral diversity in HIV vaccine design, *PLoS Comput. Biol.*, **3**, e75.
Olesen, J. M., *et al.* (2007) The modularity of pollination networks, *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 19891-19896.
Palker, T. J., *et al.* (1989) Polyvalent human immunodeficiency virus synthetic immunogen comprised of envelope gp120 T helper cell sites and B cell neutralization epitopes, *J. Immunol.*, **142**, 3612-3619.
Paradis, E., Claude, J. and Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in R language, *Bioinformatics*, **20**, 289-290.
Rambaut, A., *et al.* (2008) The genomic and epidemiological dynamics of human influenza A virus, *Nature*, **453**, 615-619.
Reshef, D. N., *et al.* (2011) Detecting novel associations in large data sets, *Science*, **334**, 1518-1524.
Rolland, M., Nickle, D. C., and Mullins, J. I. (2007) HIV-1 Group M Conserved Elements Vaccine, *PLoS Pathogens*, **3**, e157.
Salathe, M., *et al.* (2010) A high-resolution human contact network for infectious disease transmission, *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 22020-22025.
Santra, S., *et al.* (2010) Mosaic vaccines elicit CD8+ T lymphocyte responses that confer enhanced immune coverage of diverse HIV strains in monkeys, *Nat. Med.*, **16**, 324-328.
Schloss, P. D., *et al.* (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities, *Appl. Environ. Microbiol.*, **75**, 7537-7541.
Shi, M., *et al.* (2010) Molecular epidemiology of PRRSV: A phylogenetic perspective, *Virus Res.*, **154**, 7-17.
Shi, M., *et al.* (2010) A Phylogeny-based Evolutionary, Demographical and Geographical Dissection of North American Type 2 Porcine Reproductive and Respiratory Syndrome Viruses, *J. Virol.*, **84**, 8700-8711.
Stadejek, T., *et al.* (2006) Porcine reproductive and respiratory syndrome virus strains of exceptional diversity in eastern Europe support the definition of new genetic subtypes, *J. Gen. Virol.*, **87**, 1835-1841.
Stamatakis, A. (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics*, **22**, 2688- 2690.
Stamatakis, A., Hoover, P. and Rougemont, J. (2008) A Rapid Bootstrap Algorithm for the RAxML Web Servers, *Syst. Biol.*, **57**, 758-771.
Swofford, D.L. (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, Massachusetts.
Team, R.D.C. (2010) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria.
Thanawongnuwech, R. and Suradhat, S. (2010) Taming PRRSV: Revisiting the control strategies and vaccine design, *Virus Res.*, **154**, 133-140.
Thomson, S. A., *et al.* (2005) Development of a synthetic consensus sequence scrambled antigen HIV-1 vaccine designed for global use, *Vaccine*, **23**, 4647- 4657.
Thurmond, J., *et al.* (2008) Web-based design and evaluation of T-cell vaccine candidates, *Bioinformatics*, **24**, 1639-1640.
Tian, K., *et al.* (2007) Emergence of fatal PRRSV variants: Unparalleled outbreaks of atypical PRRS in China and molecular dissection of the unique hallmark, *PLoS ONE*, **2**, e526.
Wahala, W.M.P.B., *et al.* (2010) Natural Strain Variation and Antibody Neutralization of Dengue Serotype 3 Viruses, *PLoS Pathogens*, **6**, e1000821.
Wasserman, S. and Faust, K. (1994) *Social network analysis: methods and applications*. Cambridge University Press, Cambridge.
Watts, D. (2004) The "New" Science of Networks, *Ann. Rev. Sociol.*, **30**, 243-270.
Wijesundara, D.K., *et al.* (2011) Human immunodeficiency virus-1 vaccine design: Where do we go now?, *Immunol. Cell Biol.*, **89**, 367-374.
Yamada, T. and Bork, P. (2009) Evolution of biomolecular networks: Lessons from metabolic and protein interactions, *Nat. Rev. Mol. Cell. Biol.*, **10**, 791-803.
Yu, H., *et al.* (2007) The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics, *PLoS Comput. Biol.*, **3**, e59.
Zhao, S. and Li, S. (2010) Network-based relating pharmacological and genomic spaces for drug target identification, *PLoS ONE*, **5**, e11764.
Zimmerman, J. J., *et al.* (1997) General overview of PRRSV: A perspective from the United States, *Vet. Microbiol.*, **55**, 187-196.

Supplementary Information

The following three files in FASTA format are attached to the Download page:

1. "prrsv_database.fasta" (10.4 Mb) contains all PRRSV sequences that were stored on http://prrsvdb.org prior to it closing. See http://www.reeis.usda.gov/web/crisprojectpages/412855.html for description of database.
2. "prrsv_from_genbank.fasta" (12.8 Mb) contains all PRRSV sequences that were stored on NCBI Genbank http://www.ncbi.nlm.nih.gov/genbank
3. "final_cleaned_prrsv_seqs.fasta" (%.9 Mb) is the final dataset used in network and phylogenetic analyses. These sequences were obtained in accordance with methods section 2.2.