JOURNAL OF CLINICAL MICROBIOLOGY, Apr. 2005, p. 1807–1817 Vol. 43, No. 4
0095-1137/05/$08.00+0 doi:10.1128/JCM.43.4.1807–1817.2005

# System To Assess Genome Sequencing Needs for Viral Protein Diagnostics and Therapeutics

Shea N. Gardner,* Thomas A. Kuczmarski, Carol E. Zhou, Marisa W. Lam, and Tom R. Slezak

*Lawrence Livermore National Laboratory, P.O. Box 808, L-174, Livermore, California 94551*

Received 23 February 2004/Returned for modification 8 May 2004/Accepted 15 December 2004

**Computational analyses of genome sequences may elucidate protein signatures unique to a target pathogen. We constructed a Protein Signature Pipeline to guide the selection of short peptide sequences to serve as targets for detection and therapeutics. In silico identification of good target peptides that are conserved among strains and unique compared to other species generates a list of peptides. These peptides may be developed in the laboratory as targets of antibody, peptide, and ligand binding for detection assays and therapeutics or as targets for vaccine development. In this paper, we assess how the amount of sequence data affects our ability to identify conserved, unique protein signature candidates. To determine the amount of sequence data required to select good protein signature candidates, we have built a computationally intensive system called the Sequencing Analysis Pipeline (SAP). The SAP performs thousands of Monte Carlo simulations, each calling the Protein Signature Pipeline, to assess how the amount of sequence data for a target organism affects the ability to predict peptide signature candidates. Viral species differ substantially in the number of genomes required to predict protein signature targets. Patterns do not appear based on genome structure. There are more protein than DNA signatures due to greater intraspecific conservation at the protein than at the nucleotide level. We conclude that it is necessary to use the SAP as a dynamic system to assess the need for continued sequencing for each species individually and to update predictions with each additional genome that is sequenced.**

Protein-based assays for pathogen detection complement DNA-based assays, as they provide orthogonal detection capabilities to prevent system-wide false positives or negatives, they may be easier to use in field-portable devices, and they may be less expensive per assay (9, 13). Protein signatures may be composed of a peptide sequence, a domain, or an entire protein. Since protein sequences are more conserved than are DNA sequences, protein-based detection may be important for highly divergent RNA viruses for which development of conserved DNA-based signatures has been problematic. In addition, protein-based assays may facilitate the detection of virulence proteins or proteins expressed from genes deliberately engineered to escape nucleotide detection via the use of alternative codons for several amino acids. Protein-based signatures must also be used to detect toxins, for which no nucleic acid may be present. Finally, peptide signatures may serve as targets for therapeutics and vaccines (14, 16).

We have built a Protein Signature Pipeline that may accept as input either protein sequence data (single proteins) or annotated DNA sequence data (whole genomes) from one or many strains of a target species. From these genomes, examination of the alignment of multiple sequences illuminates amino acid sequence fragments that are conserved among all strains of the target species. These conserved fragments are then compared to the NCBI GenBank nonredundant (nr) database of amino acid sequences, unveiling peptides that are unique to the target species (2). There may be many conserved and unique peptides on the same and on different proteins. All of the processes described above are fully automated on a 24 CPU Sun server, from multiple sequence alignment and determination of conserved fragments, to calculation of unique fragment peptides.

The resulting conserved, unique peptides that are at least 6 amino acids long are considered to be protein signature candidates. These protein or peptide signatures are short amino acid sequences from open reading frames that are at least 6 amino acids in length and that extend as far as possible before (i) the end of the protein, (ii) an intraspecifically nonconserved amino acid is reached, or (iii) a nonunique 6-mer (relative to all current sequence data available in the NCBI nr protein database) is contained within the signature region.

If a subset of these signatures is to be developed empirically as a target for antibody or ligand binding, then this subset is subjected to additional analyses. These analyses include, but are not limited to, assessment of surface accessibility of the peptides within the protein, cellular location and expression of the proteins on which the peptides are located, protein stability, biochemical properties, posttranslational modifications, and antigenicity. When possible, three-dimensional structural models are built. These additional criteria currently require various levels of manual input to perform the analyses and/or to collate the results. The signatures that pass this rigorous scrutiny may be used to generate sets of antibodies or synthetic ligands that selectively bind to these protein signatures and not to proteins produced by near or distant phylogenetic neighbors. Since the signature regions are highly conserved within a species, it is likely that they are functionally important to the organism's survival or reproduction. Those signatures that land

* Corresponding author. Mailing address: Lawrence Livermore National Laboratory, P.O. Box 808, L-174, Livermore, CA 94551. Phone: (925) 422-4317. Fax: (925) 423-6437. E-mail: gardner26@llnl.gov.

on or near protein-active sites may be developed into therapeutics, since antibody or ligand binding may interfere with protein function. Signature regions may even be considered as vaccine targets, since these unique peptides may educe a highly specific response in the host (4, 15).

The SAP software calls only the fully automated portion of the pipeline, in which conservation and uniqueness are determined. These are the aspects relevant to analyses of sequencing needs, since additional sequence data alter the regions indicated to be conserved and unique but do not modify conclusions regarding protein expression, posttranslational modifications, protein structure, and so on.

## MATERIALS AND METHODS

For the SAP analyses, we start with a pool of $T$ target genomes, and from that, we randomly select $s$ samples of size $t$, where $t$ ranges from 1 to $T$, sampling without replacement, so that no genome is duplicated in a given sample. Each sample must contain a high-quality reference genome with annotation to delineate protein-coding regions. The remaining genomes may be finished or draft sequences. Thus, in cases where there is only one finished, annotated genome available, that genome is included in every sample, and the remainder of the genomes are randomly chosen.

Second, on each sample of genomes, we perform a nucleic sequence alignment with the alignment program Whole Genome Alignment through Scalable Algorithms, developed at Lawrence Livermore National Laboratory by David Hysom and Chuck Baldwin. This new software is the only tool currently available that enables us to align multiple finished or draft genomes with one or more finished genomes and can align large bacterial genomes in minutes.

In addition, a set of gene pairs (start, end) for both the plus and minus strands relative to the reference genome is required. This implies that coding frames for the translation of nucleic acid codons into amino acids for each protein of the target organism's genome have been correctly determined.

Next, we determine amino acid conservation among the target genomes within a given sample based on their nucleic acid sequence alignment. For each gene pair (start, end), we move through the corresponding gene sequence in the alignment, noting amino acids where the many-to-one map from codons to an amino acid specifies the same amino acid in each of the aligned sequences. We record each peptide that is composed of a series of six or more contiguous amino acids that are the same in all the target sequences. There may be multiple conserved peptides in each protein delineated by the gene pairs (start, end). The software does not generate output from sections of an multiple sequence alignment that contain insertions or deletions. However, it continues to scan input, and if it finds another region without insertions or deletions, it will recover in the correct coding frame and continue processing. If codons that map to STOP are found in the same place and correct coding frame in each genome, the software will terminate processing of that gene and proceed to the next gene pair (start, end). The software is coded to handle overlapping gene pairs (start, end). The output of this portion of the software is a FASTA-formatted list of each conserved peptide.

This target conservation FASTA file for the sample under consideration is then fed into the uniqueness verification part of our Protein Signature Pipeline, as outlined above. The inputs to this process are (i) the NCBI GenBank nr database; (ii) a list of GenBank gi (genome identification) numbers corresponding to all nr entries that are found in the target organism; and (iii) the FASTA file of peptides conserved among target strains, as described above. First, entries in the list of gi numbers that are found in the GenBank nr database are removed from a copy of the database that we call nr_minus. Thus, nr_minus contains no entries from the target organism, and we aim to find peptides from the target conservation FASTA that are unique relative to anything in nr_minus. To do so, we use suffix tree algorithms (8) to eliminate all peptides from our target conservation FASTA that match any peptide with a length of six or greater in nr_minus. Suffix tree algorithms serve as the most efficient and scalable method that we have found for comparing query sequences to large sequence databases (3, 19).

These analyses yield a computationally predicted list of peptides that are conserved among target strains (based on nucleic acid sequences) and unique relative to any nontarget proteins in the nr database. For the SAP analyses, we use the scalar statistic of $y$ = the number of protein signatures for a given sample of input target genomes, and we do not perform additional protein signature annotation. We examine the range of $y$ for all $s$ samples of size $t$ target input genomes and plot the range and its quantiles for each value of $t$ using range plots. For these analyses, $s = 10$, a constraint set by the time required to run each call of the Protein Signature Pipeline (approximately 20 min) and the total number of Monte Carlo simulations completed (1,500) for the results presented in this paper.

Range plots illustrate the span of predictions generated by different random samples of genomes (see results in Fig. 2). The number of target strains $t$ is represented along the $y$ axis. The numbers of peptide signatures are plotted along the $x$ axis as a horizontal line spanning the range of predicted values for the $s$ random samples. The median, 75%, and 90% quantiles of the random samples are indicated with three vertical short lines along each horizontal range line. If a sample of $t$ target strains were sequenced, there would be a 90% chance that the number of protein signatures for that sample would be less than or equal to the 90% quantile mark. The expected outcome is a reduction in the number of signatures that are generated as nonconserved candidates are eliminated with increases in the number of target sequences used to predict the signatures. If the number of signatures predicted using all $T$ targets in the pool is $c$, then we arbitrarily chose a threshold value for the 75% quantile of $c + 20$ as an objective goal for sequencing efforts. That is, for a target sample size $t$, if the 75% quantile landed within 20 of the number of signatures predicted using the full data set, then at least $t$ genome sequences would be desired for this species for the purposes of protein signature prediction. These range plots enable us to examine the entire span of outcomes on a relatively simple graph and to rapidly determine the value of additional target sequences. They were created using the R statistical language (10).

DNA signatures and SAP results were computed as described previously (5, 7). Briefly, DNA signatures were generated as follows. Conserved regions of the genomes of a target species were determined using multiple sequence alignment. Unique regions relative to sequence in a 1-Gb database of nontarget bacterial and viral species were identified using suffix tree algorithms developed by S. Kurtz and colleagues (http://www.zbh.uni-hamburg.de/research/GI/software/vmatch/). From the conserved, unique regions, primers and probes suitable for TaqMan assays were selected. These may be in either coding or intergenic regions. The SAP analyses for DNA signatures were performed using Monte Carlo sampling from the pool of target genomes, as described above for proteins, except that DNA rather than protein signatures were computed for each random sample.

Our DNA SAP analyses examined the number of target sequences as well as the number of near-neighbor sequences required (Monte Carlo simulations with sample sizes of up to 10 near neighbors), but our protein SAP analyses investigated only the number of target sequences required. The reason is that composing the lists of near-neighbor proteins for random, temporary exclusion from the protein database (to estimate the value of that near-neighbor sequence data) would be difficult to automate for rapid, high-throughput computations. Thus, we compared the target proteins to all the proteins in NR, regardless of their phylogenetic relationship to the target. This was comparable to DNA SAP results using all available near-neighbor data.

Statistical analyses of results were performed using Microsoft Excel and JMP of the SAS Institute, Inc. In order to determine the contribution of variation among strains in codon usage to our finding that there are more conserved protein than DNA signatures, we performed the following analyses. For all amino acids that were conserved among the sequenced isolates of a given target species (or type), the number of times that a different codon was used by any isolate for a conserved amino acid was tabulated (nucleotide sequence divergence), as was the total number of times each amino acid was conserved (protein sequence conservation). The ratio of these two numbers, representing the fraction of times that a different nucleotide sequence coded for a conserved amino acid, was plotted using the JMP statistical package.

## RESULTS

For most organisms, sequencing 1 to 4 target genomes will narrow the selection of TaqMan DNA signature candidates down to within 20 of the number using the full data set (Table 1) (7). The numbers of genomes needed to narrow the list of protein signatures to within 20 of that predicted with the full data set is highly variable, from 1 to over 20, and does not appear to be related to genome structure (e.g., single- or double-stranded RNA or DNA), genome length, or the fraction of

TABLE 1. Fully informed predicted numbers of nucleotide and protein signatures, and number of sequences required to approximate fully informed predictions[a]

| Genome structure | Virus | No. of genomes in target pool | Approx genome length (1,000 bp) | Conserved and unique fraction of target genome with full data set (%) | Fraction conserved and unique X genome length | *x* | *y* | *t* such that 75% quantile is within 20 of *x* (no. of target genomes) | *t* such that 75% quantile is within 20 of *y* (no. of target genomes) |
|---|---|---|---|---|---|---|---|---|---|
| dsDNA virus | Human adenovirus B | 6 | 35 | 67 | 23.8 | 3 | 18 | 3 | 5 |
| | Human papillomavirus type 16 | 8 | 8 | 65 | 5.04 | 11 | 37 | 1 | 2 |
| | JC | 210 | 5 | 68 | 3.5 | 1 | 31 | 10 | 1 |
| | Vaccinia | 6 | 194 | 5 | 9.7 | 0 | 52 | 1 | 3 |
| | Variola | 14 | 186 | 5 | 9.3 | <20 | 90 | 1 | 6 |
| ssDNA virus | Maize streak | 32 | 2.7 | 52 | 1.4 | 0 | 3 | 1 | 5 |
| Retroid virus | Hepatitis B | 379 | 3 | 20 | 0.5 | 0 | 0 | 1 | 1 |
| ssRNA negative-strand nonsegmented virus | Marburg | 6 | 19 | 56 | 15.8 | 0 | 113 | 4 | 6 |
| | Ebola Zaire | 5 | 19 | 80 | 10.6 | 167 | 119 | 1 | 1 |
| | Mumps | 13 | 15 | 85 | 12.8 | 4 | 65 | 6 | 9 |
| | Vesicular stomatitis | 4 | 11 | 88 | 9.7 | 2 | 100 | 4 | 4 |
| ssRNA negative segmented | Lassa virus segment S | 6 | 3.4 | 13 | 0.44 | 0 | 19 | 2 | 2 |
| ssRNA positive-strand nonsegmented virus | FMDV | 19 | 8 | 33 | 2.64 | 0 | 24 | 3 | 14 |
| | Human poliovirus | 31 | 7 | 21 | 1.5 | 0 | 0 | 2 | 3 |
| | Human poliovirus 1 | 22 | 7 | 46 | 3.22 | 0 | 0 | 3 | 3 |
| | Plum pox virus | 5 | 10 | 83 | 8.3 | 14 | 138 | 3 | 4 |
| | SARS coronavirus | 40 | 30 | 78 | 23.4 | 100 | 1106 | 1 | >21 |
| | Venezuelan equine encephalitis virus | 18 | 11 | 5 | 0.6 | 0 | 15 | 2 | 8 |

[a] *x*, number of TaqMan DNA signatures with full data set; *y*, number of protein signatures with full data set; ds, double-stranded; ss, single-stranded.

the genome that is conserved and unique (Tables 1 and 2). In most cases, more sequenced genomes are required (to narrow the list of signature candidates to within 20 of our best estimate using all genome sequences currently available) for protein signatures than for TaqMan DNA signatures. Otherwise, no generalizations can be made regarding the number of sequenced genomes needed for protein signatures (Table 2). All correlations between the numbers of genomes needed to narrow the list of protein signatures to within 20 of that predicted are weak, using the full data set and any of the other factors (last row of Table 2).

Our analyses predict that substantially more protein signatures than TaqMan DNA signatures exist that are conserved among all the strains of a species (Table 1). This is predicted

TABLE 2. Pairwise correlation coefficients between the variables in Table 1, excluding the outlying data points for SARS[a]

| Parameter | No. of genomes in target pool | Approx genome length (1,000 bp) | Conserved and unique fraction of target genome with full data set (%) | Fraction conserved and unique X genome length | *x* | *y* | *t* such that 75% quantile is within 20 of *x* (no. of target genomes) | *t* such that 75% quantile is within 20 of *y* (no. of target genomes) |
|---|---|---|---|---|---|---|---|---|
| No. of genomes in target pool | 1.00 | | | | | | | |
| Approx genome length (1,000 bp) | −0.19 | 1.00 | | | | | | |
| Conserved and unique fraction of target genome with full data set (%) | −0.13 | −0.47 | 1.00 | | | | | |
| Fraction conserved and unique X genome length | −0.38 | 0.35 | 0.47 | 1.00 | | | | |
| *x* | −0.14 | −0.06 | 0.35 | 0.28 | 1.00 | | | |
| *y* | −0.37 | 0.22 | 0.53 | 0.83 | 0.43 | 1.00 | | |
| *t* such that 75% quantile is within 20 of *x* | 0.20 | −0.29 | 0.46 | 0.17 | −0.20 | 0.08 | 1.00 | |
| *t* such that 75% quantile is within 20 of *y* | −0.35 | 0.01 | −0.09 | 0.11 | −0.28 | 0.01 | 0.07 | 1.00 |

[a] *x*, number of TaqMan DNA signatures with full data set; *y*, number of protein signatures with full data set.

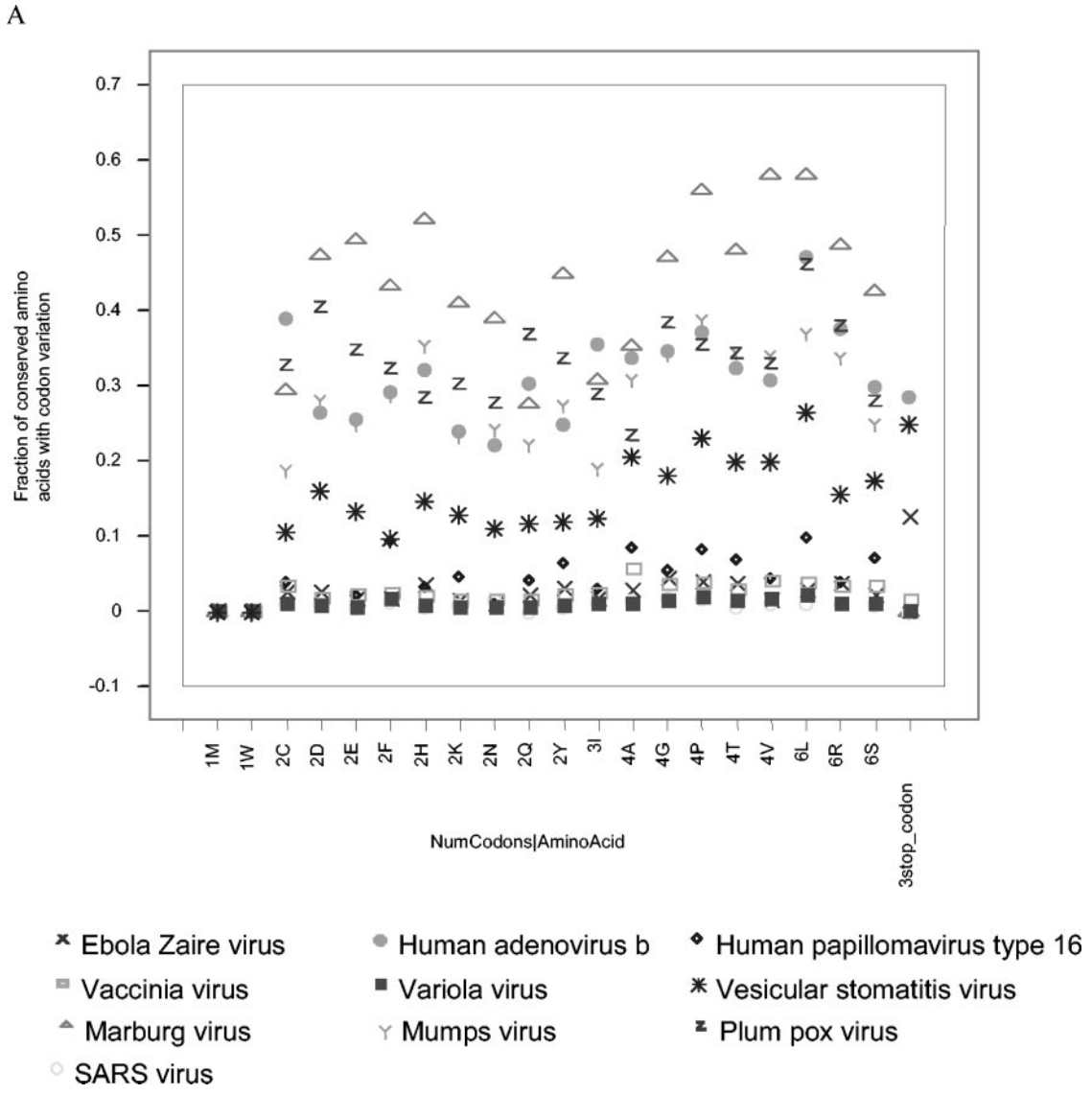



FIG. 1. Fraction of conserved amino acids for which there is variation in nucleotide sequence across strains (that is, alternative codons used for the same amino acid in a give location in the proteome). The amino acids are listed along the *y* axis, sorted by the number of codon options for each amino acid (indicated by the number immediately preceding the one-letter amino acid abbreviation). (A) The less- to moderately divergent species, and (B) the moderately to more-divergent species (Marburg virus, JC virus, and maize streak virus could have been included in either plot).

despite the fact that protein signatures are limited to coding regions of a genome, while DNA signatures may occur in either coding or intergenic regions. To a large extent, this stems from the fact that amino acid sequences are more conserved than are nucleotide sequences due to the wobble, usually in the third base position, of many codons (Fig. 1). There is a large difference among viruses, with Lassa, human polio, Venezuelan equine encephalitis, foot-and-mouth disease, and hepatitis B viruses showing high nucleotide divergence coding for conserved amino acids. Human adenovirus B and JC, maize streak, mumps, Marburg, plum pox, and vesicular stomatitis viruses display intermediate levels of nucleotide variation. Human papillomavirus type 16 and severe acute respiratory syndrome (SARS), Ebola Zaire, vaccinia, and variola viruses show very low levels of nucleotide variation in codon use among sequenced isolates. Although one to six possible codons may code for an amino acid, codon variation differences among amino acids do not show a pattern relating to the number of codon options.

The number of protein signatures is correlated with the number of conserved and unique DNA bases (Table 2, correlation coefficient of 0.83), excluding the outlying data points for SARS. The correlation between the number of protein signatures and the number of TaqMan DNA signatures is weak (correlation coefficient = 0.43). In an analysis of variance using the number of protein signature candidates as the dependent variable and with the three model effects of (i) genome structure, (ii) the number of genomes, and (iii) the number of

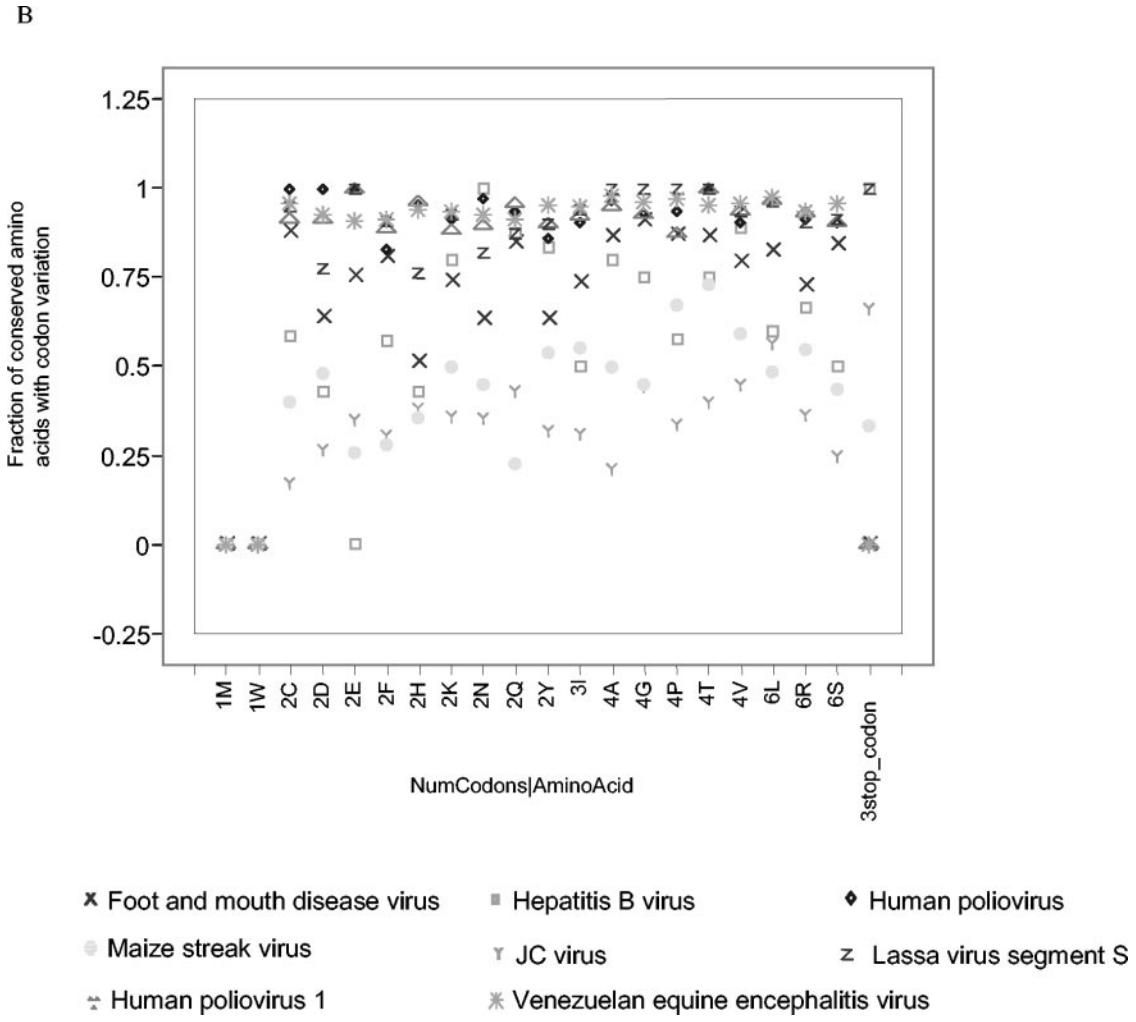



FIG. 1—*Continued.*

conserved and unique DNA bases, only the number of conserved and unique DNA bases had a significant effect, with $P = 0.046$. Results for each species are shown in Fig. 2.

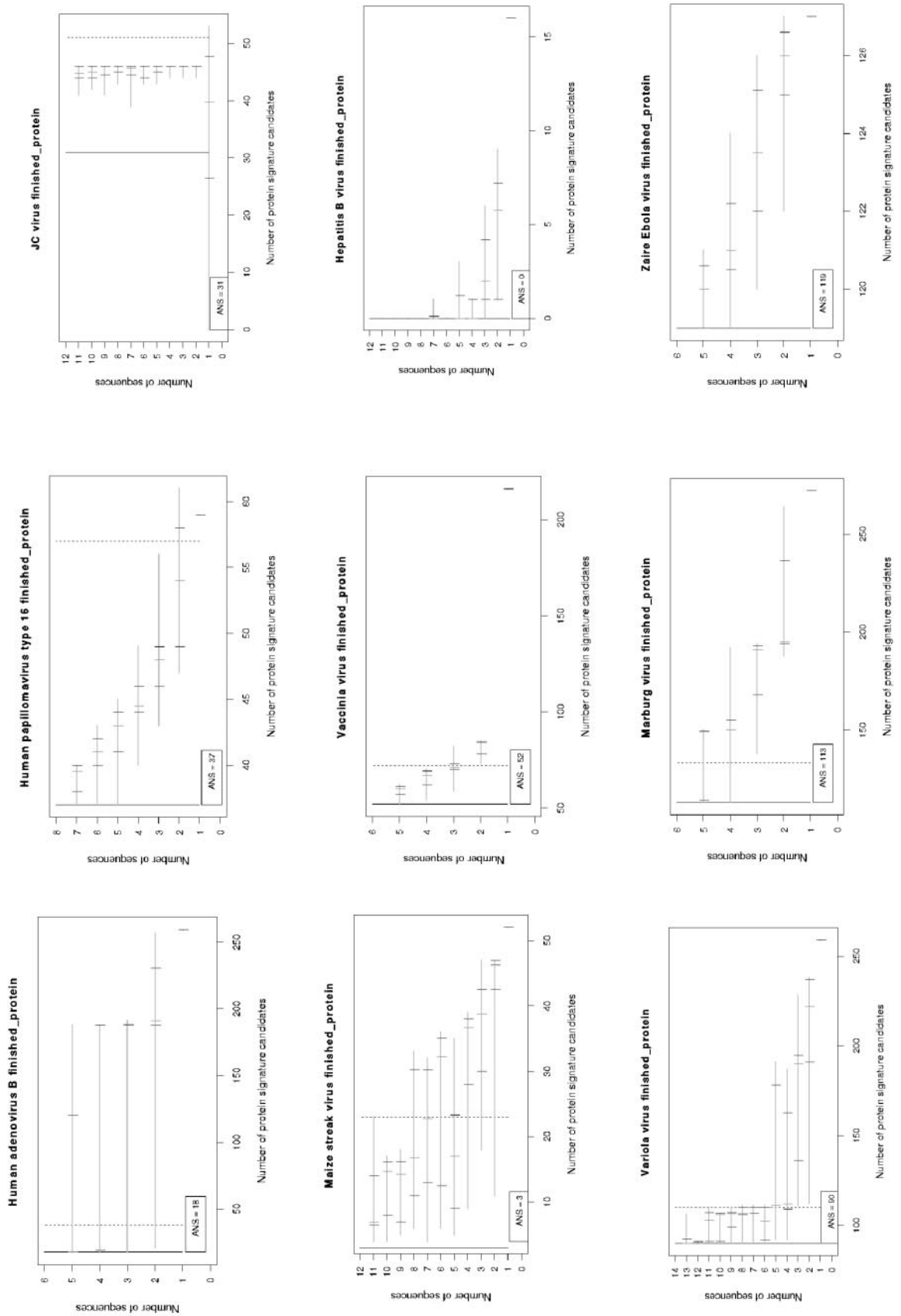
**Human adenovirus B.** Human adenovirus B appears to have one strain that is more divergent from the others, so if this strain and any one of the four more closely related strains were sequenced, adequate predictions could have been made with only two sequences. For a random selection of strains, however, it is necessary to use five genomes in order to have a 75% chance of predicting the set of signatures observed using all six strains.

**Human papillomavirus.** Numbers of protein signatures for human papillomavirus type 16 continue to decline even when as many as six or seven sequences are used to generate the predictions, suggesting that additional sequences might continue to eliminate nonconserved candidates. However, the overall number of protein signatures is fairly low, so other annotation analyses, regarding expression levels, surface accessibility, and so on, of existing signatures should be considered, as this might be a more productive investment to narrow a list for laboratory study than continued sequencing.

**JC polyomavirus.** Since 210 genomes of JC virus have already been sequenced, it is unlikely that additional genomic sequencing is required for the prediction of peptide signature candidates. Only 31 peptide signatures stand up to computational screening for conservation and uniqueness, a feasible number for additional annotation and empirical investigations. Combinations of 2 to 11 of the 210 genomes produce 40 to 45 signature candidates, indicating that a wise selection of a few of the most distantly related strains of JC virus for sequencing would have been sufficient to predict a manageable list of peptide signature regions.

**Vaccinia virus.** Results suggest that there are adequate numbers of vaccinia sequences to predict peptide signature candidates, since the number of candidates appears to be approaching a plateau around 50. The range plot indicates that additional sequencing is unlikely to reduce the number of candidates much below 52. Additional annotation of the current 52 targets is feasible, followed by lab screening of the most promising.

**Variola virus.** For variola virus, the range plot indicates that additional sequencing of stored isolates from infections during


JC virus finished_protein
Hepatitis B virus finished_protein
Zaire Ebola virus finished_protein
Human papillomavirus type 16 finished_protein
Vaccinia virus finished_protein
Marburg virus finished_protein
Human adenovirus B finished_protein
Maize streak virus finished_protein
Variola virus finished_protein
Number of sequences
Number of protein signature candidates
ANS = 31
ANS = 0
ANS = 119
ANS = 37
ANS = 52
ANS = 113
ANS = 18
ANS = 3
ANS = 90

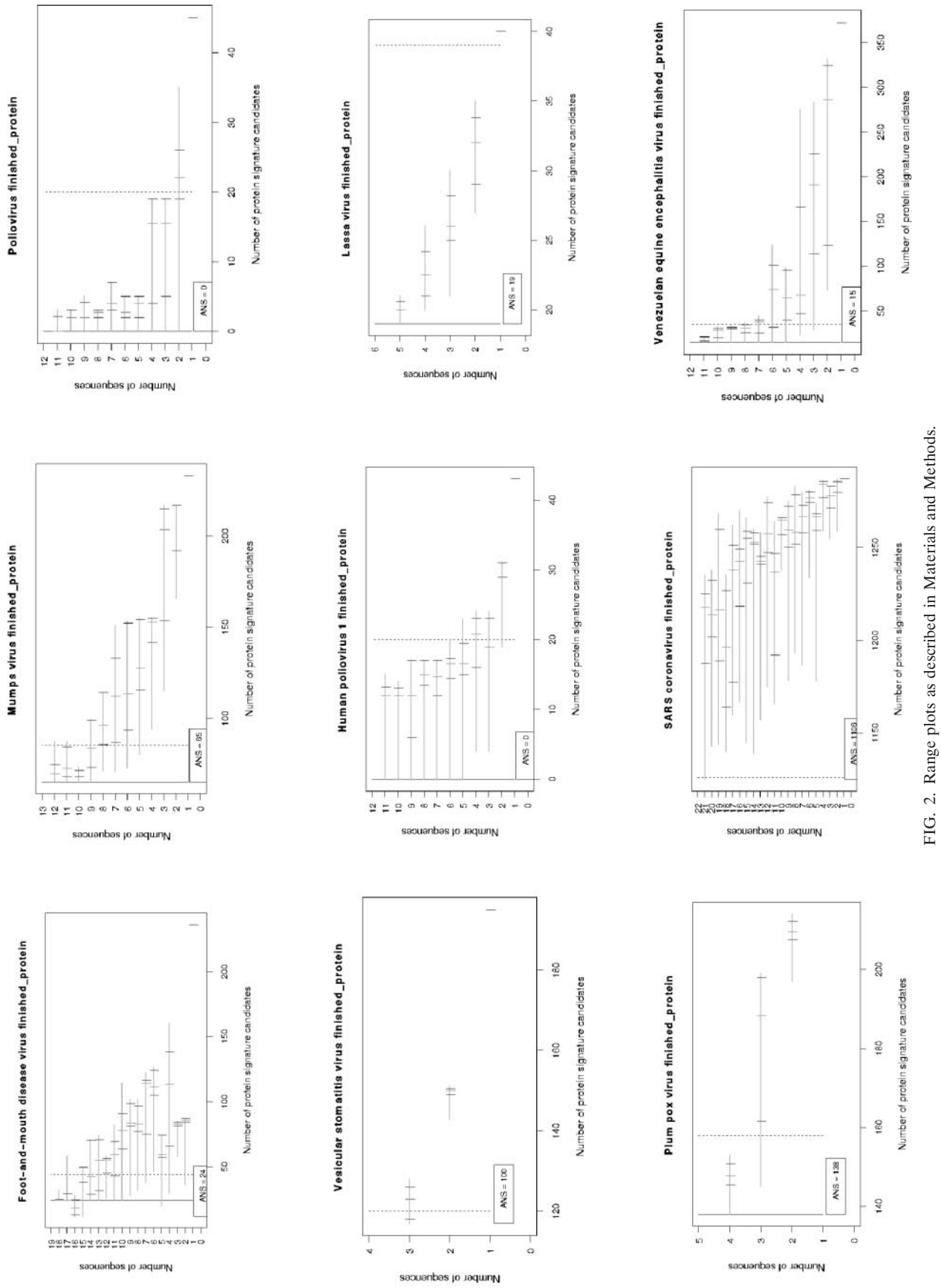



FIG. 2. Range plots as described in Materials and Methods.

the 1900s is unlikely to narrow the list of protein signature candidates. The median and lower bound of the range for the number of candidates lies within 20 of the number predicted using the full data set, with as few as four genomes.

**Maize streak virus.** Maize streak virus has only three peptide signatures using all 32 sequences available at the time of our analyses, although most combinations of 9 or more sequences would have been adequate to narrow the number of candidates to close to this. Thus, additional genome sequencing for the purpose of protein signature development is not recommended.

**Hepatitis B virus.** Hepatitis B virus is so heterogeneous that not a single peptide candidate can be found that is conserved among all sequenced strains. Only four genome sequences could have provided this information, so sequencing could have stopped there if the only aim of sequencing were to discover a single, conserved peptide target. However, hepatitis B results highlight the fact that continued sequencing may be desired to identify all of the variant sequences in a divergent species. For protein diagnostic signatures, it may be necessary to subdivide species with divergent isolates or strains, such as hepatitis B, into different clusters or clades and to develop clade-specific peptide signatures rather than species-specific signatures. This will enable signature peptides that are conserved within the clade to be identified if there are no peptides conserved across all members of the species. In this case, one would perform SAP analyses on the different clades to determine when a sufficient number of isolates had been sequenced from that clade.

**Marburg virus.** There are 113 peptide signature candidates for Marburg virus using the six target genomes currently available to us. Since the number of candidates has declined from using four or five targets, indications are that the point of diminishing returns has not been reached and that additional sequencing may be desired to further narrow the selection of candidates for testing.

**Ebola Zaire virus.** Strains of Ebola Zaire virus are so similar that additional sequencing of the isolates from recent outbreaks is not required for developing protein signatures. Although there are too many signatures (119) to test all of them, additional sequencing is unlikely to eliminate nonconserved candidates at this time, since so little strain divergence has occurred for this emergent pathogen. If a geographically separate or symptomatically different outbreak occurs, then additional sequencing may be warranted.

**Mumps virus.** Sequencing the first nine strains of the mumps virus led to better prediction of conserved protein signatures. After 10 or more sequences, however, little improvement occurred, and indications are that no further sequencing of isolates from clades or outbreaks already represented by sequencing is required for the prediction of conserved protein signatures.

**Vesicular stomatitis virus.** With 100 protein signature candidates predicted for vesicular stomatitis virus and declines from each strain added up to the four genomes currently available, additional sequencing will likely narrow the selection and improve the quality of protein signature candidates.

**Lassa virus.** Lassa virus segment S is the only segment of Lassa virus (and the only segmented virus) with sufficient available sequence data to generate informative SAP range plots. Only 19 protein signatures are predicted to be conserved across all sequenced strains, and gains from sequencing more than two or three strains are minimal.

**FMDV.** With 19 genomes of foot-and-mouth disease virus (FMDV) publicly available at the time of our analyses, 24 conserved, unique protein signatures can be predicted. Some judiciously chosen combinations of 10 or fewer genomes could have winnowed the candidate list to approximately this level, so it appears that no additional sequencing of FMDV isolates from already-sequenced outbreaks is required for protein signature prediction.

**Human poliovirus.** Human poliovirus (types 1, 2, and 3) is very heterogeneous (like hepatitis B virus), yielding no protein signatures that are conserved among strains. In fact, some combinations of only two strains generated not a single conserved peptide. Thus, the aim of continued sequencing is to identify all variants, and this is useful to identify subgroupings of isolates for which protein signatures might be developed.

Because human poliovirus is so heterogeneous, we also did a SAP run to look for protein signatures that were unique to poliovirus (types 1, 2, or 3) and conserved only among the 22 available genomes of poliovirus type 1. Still, poliovirus type 1 was too variable for a single protein target, with as few as five genomes.

We looked in more detail at a multiple sequence alignment, and it was evident that the strain (gi 30908795 gb AY278553.1 Human poliovirus 1 isolate P1W/Bar65) collected in Byelorussia in 1963 to 1966 was very different from the other sequences, all of which were collected from 1990 onward (Fig. 3). This strain was as different from the other isolates collected in Russia during 1996 and 1999 as it was from isolates collected in China or Haiti since 1991. Running the protein signature pipeline with only the other 21 genomes, excluding the isolate collected in 1963 to 1966, yielded 10 peptide signature candidates that were conserved and unique relative to everything in nr except poliovirus types 1, 2, and 3. This highlights the contribution of temporal separation to viral heterogeneity and the importance of sampling across time as well as across spatial dimensions.

**Plum pox virus.** Results suggest that additional plum pox virus sequencing may improve the quality and reduce the quantity of protein signature candidates. Since only a subset of the 139 current signatures could feasibly be screened in a laboratory, narrowing the candidate pool will be necessary.

**SARS virus.** The 40 sequences of SARS virus available at the time of our analyses are so conserved that our analyses predict over a thousand signatures. Near-neighbor sequence data may be more valuable to eliminate nonunique candidates than more SARS sequences from the outbreak already represented.

**Venezuelan equine encephalitis virus.** Venezuelan equine encephalitis virus is extremely variable at the DNA level, and it is not possible to identify a single TaqMan DNA signature (Table 1) (6). At the protein level, in contrast, 15 peptide signatures are conserved in all 18 available genomes. The list of 15 or so candidates remains fairly constant whether eight or more genomes are used in the analyses, indicating that no further sequencing of currently known isolates for outbreaks already represented by sequencing efforts is warranted for the purpose of protein signature prediction.

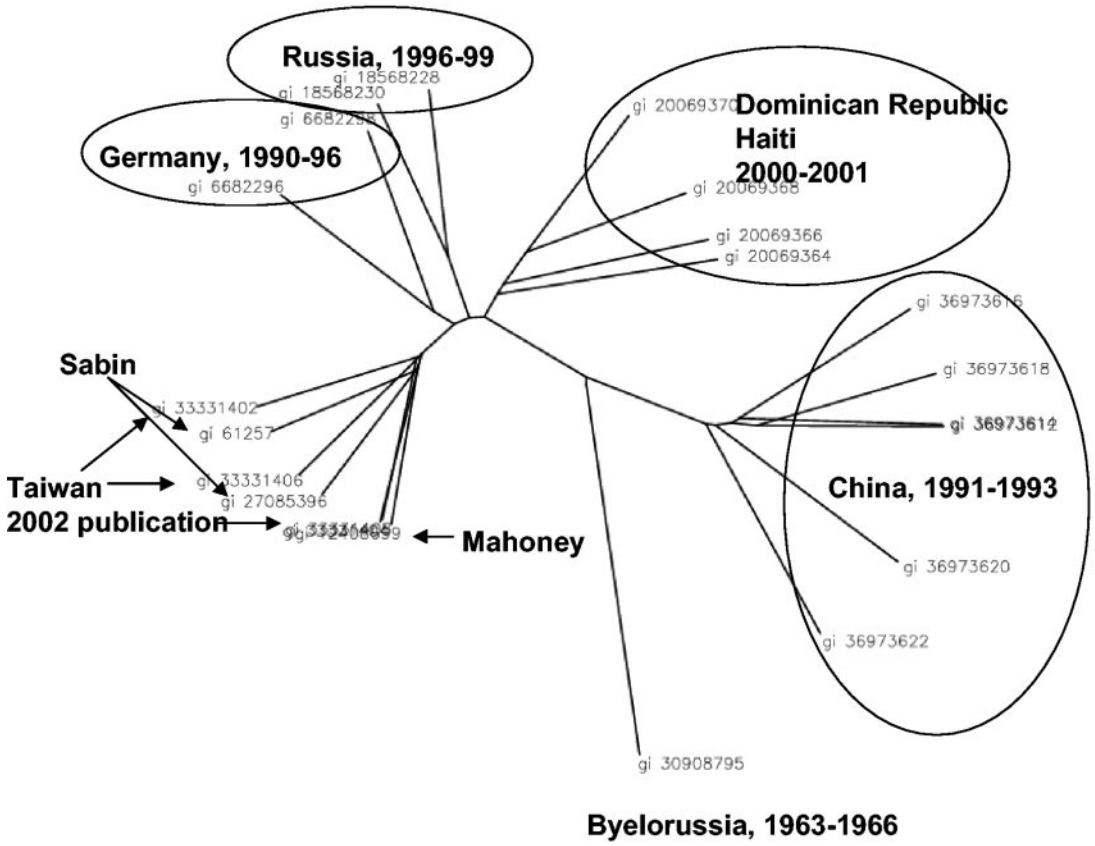



FIG. 3. Unrooted phylogenetic tree of human poliovirus type 1 genomes constructed by applying the unweighted pair group method with arithmetic mean for clustering to the DiAlign similarity scores computed using DiAlign. The tree was drawn using PHYLIP (http://evolution.genetics.washington.edu/phylip.html). The origin of each isolate is indicated. We were unable to find the collection date of Taiwanese isolates (gi 33331402, gi 33331404, gi 33331406, and gi 33331408) from a human immunodeficiency virus patient specified in the source publication. When the outlier from Byelorussia, collected during the 1960s, is excluded from the calculations for amino acid conservation, 10 protein signatures that are conserved among all the other genomes are identified.

## DISCUSSION

These analyses indicate that there are more protein than TaqMan DNA signatures for virtually all of the organisms examined. This is a result of mainly the following factors: First, proteins are more conserved than are nucleotide sequences, due to the nucleotide wobble often in the third base position for many amino acids. Second, our definition of a protein signature requires a minimum of only six conserved, unique consecutive amino acids, while TaqMan DNA signatures require conserved, unique nucleotides for two primers and a probe, each of at least 18 base pairs. Third, strict limitations on sequences deemed suitable for TaqMan PCRs (18) (e.g., amplicon length, no self complementarity, $T_m$, etc.) eliminate many sequence regions that are conserved and unique. In contrast, for the protein signature counts that we reported here, we did not consider other limitations besides conservation and uniqueness (e.g., expression, surface accessibility, etc.) that would further reduce the number of protein signature candidates in preparing a list to go to the laboratory for experimental development.

Our analyses indicate that the key reason for the higher frequency of protein than nucleotide signatures is protein sequence conservation through the existence of multiple codons for the same amino acid. For organisms with many protein signatures, such as the emerging viruses SARS and Ebola Zaire, less than 1 to 5% of the conserved amino acids have variable nucleotide codons. For viruses with an intermediate number of protein and DNA signatures, such as human adenovirus B and vesicular stomatitis, approximately 10 to 40% of the conserved amino acids have variable nucleotide codons. Very divergent viruses with few or no signatures, such as Venezuelan equine encephalitis, Lassa, and polio, display 90 to 100% codon variation in the conserved amino acids. Such high levels of nucleotide variation in regions of protein conservation make a case in favor of protein detection assays over nucleotide assays for these viruses.

Regardless of the considerations above, the fact that there are more protein than DNA signature candidates is particularly notable for highly variable viruses of biothreat concern. For example, for Marburg virus, Venezuelan equine encephalitis virus, and FMDV, there is not a single TaqMan DNA signature that is conserved among all strains, but there are multiple protein signatures. Nucleotide sequence conservation among strains is so low for some single-stranded RNA viruses that there are no regions long enough from which to select a single stretch of 18 conserved bases on which to locate a

primer. Thus, the fact that we can identify highly conserved, species-specific peptides indicates that these peptides, or the proteins on which they reside, may be important targets for therapeutics and vaccines.

SARS and, to a lesser extent, Ebola Zaire are outliers because there was no near-neighbor sequence data in GenBank at the time of our analyses to narrow the list of protein or DNA signature candidates. Most of the genomes of these two viruses are conserved and unique (~80%), and could be mined for signatures. Due to the recent emergence of these viruses, little divergence has occurred between isolates (12), yielding a wide selection of candidate signatures conserved among all strains. Although a single genome sequence would have been sufficient to generate a good list of TaqMan DNA signatures for SARS, dozens of sequences are necessary to narrow the list of protein signatures. Even so, with a total of 40 sequenced SARS genomes available at the time of our protein analyses, there are over 1,100 protein signature candidates, far too many to develop empirically for diagnostics, vaccines, or therapeutics. These results coincide with our conclusions regarding sequencing for TaqMan DNA signatures (7). It would be far more efficient to sequence near-neighbor species to eliminate nonunique regions of the genome than to continue sequencing additional SARS genomes. If SARS near neighbors can be sequenced, and if they follow the same patterns as the other single-stranded RNA viruses that we have examined, one might expect an order of magnitude reduction in the number of protein signature candidates. This would help to eliminate signatures that are likely to yield false positive results from close relatives. Judging by the high levels of divergence for other single-stranded RNA viruses that have been circulating for a longer period of time, however, we can predict that SARS and Ebola Zaire viruses will also diverge given time.

Our results indicate that when selecting the first isolates of a species to sequence, researchers should attempt to sequence the least similar isolates first to identify the most divergent proteins/peptides. The least similar isolates may be chosen based on spatial or temporal separation, lack of gene flow between populations, or those that present the most divergent symptoms or pathology. In some cases, as for human adenovirus B, the sequences of only two strains, if they are appropriately selected, would be sufficient to predict a list of high-quality protein signatures likely to be conserved among additional strains. However, if subsamples of strains for sequencing are randomly rather than carefully chosen, the sequencing of five strains of human adenovirus B is predicted to be necessary to narrow the list of protein signatures to those of the highest quality.

The lack of peptide signatures for poliovirus may be a consequence of a relatively high evolutionary rate for this virus. Poliovirus type 1 has been shown to have a particularly high rate of evolution on a per year basis of $9.7 \times 10^{-3}$ substitutions per year per nucleotide (1). This compares to more typical values an order of magnitude lower, $1 \times 10^{-3}$ substitutions per year per nucleotide, for most viruses. Even slower rates of evolution have been measured for others, ranging from $1 \times 10^{-6}$ to $1 \times 10^{-3}$, for viruses such as measles virus, influenza virus C, and GB virus C (11). However, regardless of the rate of viral evolution, and excluding the strain collected 4 decades ago, we were still able to discover 10 protein signatures that were conserved among all the other polio type 1 genomes available in GenBank at the time of our analyses.

Hepatitis B virus also appears to be a highly divergent virus, in terms of both nucleotide sequences and amino acid sequences. Hepatitis B, a retrovirus, lacks proofreading during viral transcription, introducing a high frequency of mutations into the copied sequence (17, 20). A clade-level analysis of the 379 genomes available at the time of our analyses would likely yield protein signatures for different subtypes, as different types are known to have different geographical distributions (20).

The paucity of generalizations that can be made regarding the number of genome sequences required to predict high-quality protein signatures argues in favor of using our SAP as a system, rather than simply for one-time analyses with which one attempts to extrapolate to other species. As additional genome sequences become available, new SAP calculations should be performed and used to evaluate whether additional sequencing is required or if the point of diminishing returns has been reached. If the number of signature candidates remains approximately constant with the addition of new sequence data, then no more genomic sequencing of the target species may be required in order to predict conserved peptide signatures (e.g., variola virus, maize streak virus, hepatitis B virus, mumps virus, foot-and-mouth disease virus, poliovirus, Venezuelan equine encephalitis virus, and JC virus). Similarly, if the number of candidates declines by only a small amount, then the cost of laboratory work to empirically eliminate poor signature candidates might be less than the cost of additional target isolate sequencing to eliminate targets computationally (e.g., vaccinia virus, Ebola Zaire virus, Lassa virus, human adenovirus B, and human papillomavirus type 16). In these cases, the decision may depend on the length of the organism, since this affects sequencing costs, versus the ease of culturing or working with the organism in the laboratory, particularly a biosafety level 3 or 4 laboratory. Otherwise, additional sequencing could be continued to eliminate regions of poor conservation from consideration (plum pox virus, vesicular stomatitis virus, and Marburg virus).

It may be true that for any virus, a new strain that is believed to be distant spatially (lack of gene flow), temporally, or symptomatically from published genomes must be sequenced and the virus reevaluated using SAP, even if previous analyses (prior to emergence of the new strain) had indicated that no further sequencing was required. This will require biological judgment on a case-by-case basis, since in many cases, the isolates already chosen for sequencing are the most different. Thus, if the sequences of many strains, all separated in time/space/symptoms, share a set of solid protein signatures, then even a totally new outbreak is likely to have the same conserved peptides.

Our finding that genome structure (e.g., single-stranded positive-sense RNA, or double-stranded DNA) does not show a clear correspondence with the number of genome sequences required to develop good protein diagnostic signatures is consistent with results of other research regarding the lack of patterns in differing rates of evolution in RNA viruses. Jenkins and colleagues (11) found that substitution rates could not be grouped based on genome polarity and segmentation, genome length, presence of an envelope, viral persistence within indi-

vidual hosts, principal host species, and whether the proteins encoded were structural or nonstructural. The only pattern that they did find was that vector-transmitted viruses display lower substitution rates. Woelk and Holmes (21) also presented results showing that in particular, vector-borne RNA viruses have lower rates of nonsynonymous substitutions in surface structural genes than do non-vector-borne viruses. They conclude that vector-borne viruses may experience less positive (diversifying) selection than non-vector-borne viruses. Thus, it is perhaps surprising that in our analyses, vector-borne viruses (maize streak virus, vesicular stomatitis virus, plum pox virus, and Venezuelan equine encephalitis virus) did not have unusually high numbers of protein signatures compared to viruses transmitted by other means.

In conclusion, we developed a system to evaluate the value of existing sequence data and the requirement for additional sequencing for the development of high quality protein signatures. These intraspecifically conserved, species-specific peptides may be developed as targets for diagnostics, therapeutics, or vaccines. The lack of generalizations that can be made about the number of genome sequences required argues for repeated use of this system to dynamically assess the need for continued sequencing after each strain is sequenced for a given species.

## ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory, under contract no. W-7405-Eng-48. This work was supported by the Intelligence Technology Innovation Center.

We gratefully acknowledge the CDC and colleagues at Lawrence Livermore National Laboratory for the sequence data which we have used in our analyses.

## REFERENCES

1. **Bellmunt, A., G. May, R. Zell, P. Pring-Akerblom, W. Verhagen, and A. Heim.** 1999. Evolution of poliovirus type 1 during 5.5 years of prolonged enteral replication in an immunodeficient patient. Virology **265:**178–184.
2. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler.** 2000. GenBank. Nucleic Acids Res. **28:**15–18.
3. **Chain, P., S. Kurtz, E. Ohlebusch, and T. Slezak.** 2003. An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. Brief. Bioinform. **4:**105–123.
4. **Choy, W., S. Lin, P. Chan, J. Tam, Y. Lo, I. Chu, S. Tsai, M. Zhong, K. Fung, M. Waye, S. Tsui, K. Ng, Z. Shan, M. Yang, Y. Wu, Z. Lin, and S. Ngai.** 2004. Synthetic peptide studies on the severe acute respiratory syndrome (SARS) coronavirus spike glycoprotein: perspective for SARS vaccine development. Clin. Chem. **50:**1036–1042.
5. **Fitch, J. P., S. N. Gardner, T. A. Kuczmarski, S. Kurtz, R. Myers, L. L. Ott, T. R. Slezak, E. A. Vitalis, A. T. Zemla, and P. M. McCready.** 2002. Rapid development of nucleic acid diagnostics. Proc. IEEE **90:**1708–1721.
6. **Gardner, S. N., T. A. Kuczmarski, E. A. Vitalis, and T. R. Slezak.** 2003. Limitations of TaqMan PCR for detecting divergent viral pathogens illustrated by hepatitis A, B, C, and E viruses and human immunodeficiency virus. J. Clin. Microbiol. **41:**2417–2427.
7. **Gardner, S. N., M. W. Lam, N. J. Mulakken, C. L. Torres, J. R. Smith, and T. R. Slezak.** 2004. Sequencing needs for viral diagnostics. J. Clin. Microbiol. **42:**5472–5476.
8. **Giegerich, R., S. Kurtz, and J. Stoye.** 2003. Efficient implementation of lazy suffix trees. Softw. Pract. Exper. **33:**1035–1049.
9. **Hoet, A., K. Chang, and L. Saif.** 2003. Comparison of ELISA and RT-PCR versus immune electron microscopy for detection of bovine torovirus (Breda virus) in calf fecal specimens. J. Vet. Diagn. Investig. **15:**100–106.
10. **Ihaka, R., and R. Gentleman.** 1996. R: a language for data analysis and graphics. J. Comp. Graph. Stat. **5:**299–314.
11. **Jenkins, G. M., A. Rambaut, O. G. Pybus, and E. C. Holmes.** 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. **54:**156–165.
12. **Leroy, E. M., P. Rouquet, P. Formenty, S. Souquiere, A. Kilbourne, J. M. Froment, M. Bermejo, S. Smit, W. Karesh, R. Swanepoel, S. R. Zaki, and P. E. Rollin.** 2004. Multiple Ebola virus transmission events and rapid decline of Central African wildlife. Science **303:**387–390.
13. **Lopez, M., E. Bertolini, A. Olmos, P. Caruso, M. Gorris, P. Llop, R. Penyalver, and M. Cambra.** 2003. Innovative tools for detection of plant pathogenic viruses and bacteria. Int. Microbiol. **6:**233–243.
14. **Matthews, T., M. Salgo, M. Greenberg, J. Chung, R. DeMasi, and D. Bolognesi.** 2004. Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes. Nat. Rev. Drug Discov. **3:**215–225.
15. **McGaughey, G., G. Barbato, E. Bianchi, R. Freidinger, V. Garsky, W. Hurni, J. Joyce, X. Liang, M. Miller, A. Pessi, J. Shiver, and M. Bogusky.** 2004. Progress towards the development of a HIV-1 gp41-directed vaccine. Curr. HIV Res. **2:**193–204.
16. **Okkels, L. M., I. Brock, F. Follmann, E. M. Agger, S. M. Arend, T. H. Ottenhoff, F. Oftung, I. Rosenkrands, and P. Andersen.** 2003. PPE protein (Rv3873) from DNA segment RD1 of *Mycobacterium tuberculosis*: strong recognition of both specific T-cell epitopes and epitopes conserved within the PPE family. Infect. Immun. **71:**6116–6123.
17. **Park, S. G., Y. Kim, E. Park, H. M. Ryu, and G. Jung.** 2003. Fidelity of hepatitis B virus polymerase. Eur. J. Biochem. **270:**2929–2936.
18. **PE Biosystems.** Sequence detection systems quantitative assay design and optimization. PE Biosystems. [Online.] http://dna-9.int-med.uiowa.edu/RealtimePCRdocs/realtimePCRbasics.pdf.
19. **Slezak, T., T. Kuczmarski, L. Ott, C. Torres, D. Medeiros, J. Smith, B. Truitt, N. Mulakken, M. Lam, E. Vitalis, A. Zemla, C. E. Zhou, and S. Gardner.** 2003. Comparative genomics tools applied to bioterrorism defence. Brief. Bioinform. **4:**133–149.
20. **Starkman, S. E., D. M. MacDonald, J. C. M. Lewis, E. C. Holmes, and P. Simmonds.** 2003. Geographic and species association of hepatitis B virus genotypes in non-human primates. Virology **314:**381–393.
21. **Woelk, C. H., and E. C. Holmes.** 2002. Reduced positive selection in vector-borne RNA viruses. Mol. Biol. Evol. **19:**2333–2336.