5-24-2012

Bioinformatics for High-throughput Virus Detection and Discovery

Adam Allred
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational Biology

Dissertation Examination Committee
David Wang, Chair
Michael Brent
Barak Cohen
Michael Diamond
Jeffrey Gordon
Henry Huang
Gary Stormo

Bioinformatics for High-throughput Virus Detection and Discovery

by

Adam Forrest Allred

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2012

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Bioinformatics for High-throughput Virus Detection and Discovery

by

Adam Forrest Allred

Pathogen detection is a challenging problem given that any given specimen may contain one or more of many different microbes. Additionally, a specimen may contain microbes that have yet to be discovered. Traditional diagnostics are ill-equipped to address these challenges because they are focused on the detection of a single agent or panel of agents.

I have developed three innovative computational approaches for analyzing high-throughput genomic assays capable of detecting many microbes in a parallel and unbiased fashion. The first is a metagenomic sequence analysis pipeline that was initially applied to 12 pediatric diarrhea specimens in order to give the first ever look at the diarrhea virome. Metagenomic sequencing and subsequent analysis revealed a spectrum of viruses in these specimens including known and highly divergent viruses. This metagenomic survey serves as a basis for future investigations about the possible role of these viruses in disease.

The second tool I developed is a novel algorithm for diagnostic microarray analysis called VIPR (**V**iral **I**dentification with a **PR**obabilistic algorithm). The main advantage of VIPR relative to other published methods for diagnostic microarray analysis is that it relies on a training set of empirical hybridizations of known viruses to guide future predictions. VIPR uses a Bayesian statistical framework in order to accomplish this. A set of hemorrhagic fever viruses and their relatives were hybridized to a total of

110 microarrays in order to test the performance of VIPR. VIPR achieved an accuracy of 94% and outperformed existing approaches for this dataset.

The third tool I developed for pathogen detection is called VIPR HMM. VIPR HMM expands upon VIPR's previous implementation by incorporating a hidden Markov model (HMM) in order to detect recombinant viruses. VIPR HMM correctly identified 95% of inter-species breakpoints for a set of recombinant alphaviruses and flaviviruses

Mass sequencing and diagnostic microarrays require robust computational tools in order to make predictions regarding the presence of microbes in specimens of interest. High-throughput diagnostic assays coupled with powerful analysis tools have the potential to increase the efficacy with which we detect pathogens and treat disease as these technologies play more prominent roles in clinical laboratories.

Acknowledgments

I would like to thank our collaborators who made valuable scientific contributions to this work: Robert Tesh, Scott Weaver, Michael Holbrook and Kael Fischer. I would also like to thank my committee for their guidance and direction: Michael Brent, Barak Cohen, Michael Diamond, Jeffrey Gordon, Henry Huang and Gary Stormo. I would like to acknowledge generous support from the Sondra Schlesinger Graduate Student Fellowship and from the Schlesinger–Olivo Student Travel Award.

I have been incredibly fortunate to work with some amazing people in the Wang lab and in the Pathogen Discovery Facility: Irma Bauer, Kevin Chen, Stacy Finkbeiner, Carl Franz, Anne Gaynor, Lori Hotlz, Hongbing Jiang, Yanfang Jiang, Jade Le, Efrem Lim, Kathie Mihindukulasuriya, Nang Nguyen, Hilary Renshaw, Erica Siebrasse, Guang Wu, Tuya Wulan, Guoyan Zhao, Megan Bower, Lindsay Droit, Song Cao, Collin Todd and Andrew Schroeder. I would especially like to thank my thesis advisor and mentor, David Wang who has provided generous measures of inspired direction and scrupulous attention to my development as a scientist, for which I am most grateful.

My beloved family and friends have been tremendously supportive of my education and in countless other ways: Thank you to my parents Gary and Lynn Allred; to my siblings Aaron (and Becca), Jill (and Jace), Casey (and Dan), Courtney (and Adam), Abbey, David and Danny; and to my dear friends Brien Ashdown, Stacy Finkbeiner and Susan Pyles.

Table of Contents

List of Tables and Figures

CHAPTER 1:

Introduction

The need for high-throughput assays for pathogen detection

Identifying the pathogenic agent or agents in a clinical specimen from the vast diversity of microbial agents known to cause disease presents a significant challenge for diagnostic microbiology. While a particular disease phenotype may implicate certain microbes over others, identifying microbes associated with poly-etiologic disease still requires testing for the presence of many potential agents. For example, in the case of encephalitis, there are at least 75 viruses that are known to cause disease [1, 2].

Another challenge to pathogen detection is the possibility that any given specimen may contain novel microbes. The fact that, for some disease syndromes, there are specimens for which no etiologic agent can be found suggests that there may be some microbial pathogens that have yet to be discovered. This appears to be a distinct possibility in the case of diarrhea, for which no etiologic agent can be detected in as many as 40% of cases [3-7].

Traditional diagnostics suffer from several limitations and are ill-equipped to address these challenges. Both PCR and antigen detection-based methods such as ELISA are dependent upon the availability of specialized reagents (primers and antibodies, respectively) which must be applied serially in separate assays in order to test for the presence of multiple agents. Multiplex PCR can test for the presence of a handful of agents simultaneously, but sensitivity is hindered with the inclusion of additional primers. Without an assay that is capable of detecting many potential agents in parallel, exhaustive testing of all microbial agents associated with a poly-etiologic disease such as encephalitis is not feasible.

Another limitation of traditional diagnostics is that they are focused on the detection of a single locus or protein in a candidate agent. The presence of the entire microbial genome is then inferred from this isolated result. This diagnostic paradigm does not accommodate the detection of recombination in candidate microbes. Unless other genomic loci or protein components are specifically assayed, recombination in target microbes would go undetected. There is a great need in diagnostic microbiology for high-throughput assays capable of detecting multiple loci from many agents in parallel and in an unbiased fashion. Metagenomic sequencing and microarrays offer significant advantages over traditional diagnostics in this regard.

Metagenomic sequencing

The vast majority of microbial species cannot be readily cultured [8]. Thus, our ability to characterize the makeup of microbial communities using cultivation-dependent methods is severely limited. Metagenomic sequencing was developed as a way to circumvent the need to culture microbes in order to explore the structure of microbial communities.

While traditional sequencing approaches are designed to characterize genomes of a single species of interest, metagenomic approaches transcend species boundaries in order to explore the makeup of communities of microorganisms [9]. The first culture-independent investigations of environmental sequences were performed in the mid-1980s and involved direct sequencing of 5S and 16S rRNA sequences to identify microbial species [10]. Later, environmental sequence was cloned into vector libraries for sequencing [11]. Among the earliest metagenomes studied using this methodology were those found in seawater samples [12]. Since that time, metagnomic sequencing has been

used to explore the structure of communities in many different environments including soil, acid mine drainage, and public restroom surfaces [13-15].

The fact that the microbial cells that inhabit the human body outnumber human cells ten to one gives credence to the classification of human beings as walking ecosystems or "superorganisms" and offers ample motivation for applying metagenomic sequencing to understanding microbial colonization and infection of humans. Among the first human-derived microbial communities to be studied were those present in stool [16, 17]. The ongoing Human Microbiome Project began in 2008 and has as its goal the characterization of microorganisms from many different sites on the human body.

Metagenomic sequencing of viruses

Metagenomic sequencing of viral communities presents a unique challenge because of the lack of a universal 16S rRNA available for phylotyping of sequenced species [8]. In fact, there is no single gene common to all virus genomes. Analyses of viral metagenomes must take into account the incredible functional diversity among virus species. Because of this, viral metagnomics typically involves sequence-independent amplification of extracted nucleic acids followed by sequencing and *in silico* similarity searches against databases containing known virus sequences [18]. One way in which metagenomic sequencing of viruses can be applied is to interrogate the community structure of virus populations. Several studies have focused on the analysis of phage for this purpose [16, 19-21].

Another way in which metagenomic sequencing is incredibly useful is in the discovery of novel virus species [22]. Novel viruses which were discovered through

sequence-independent amplification and mass sequencing include human bocavirus, WU polyomavirus, KI polyomavirus, Merkel cell polyomavirus and seal picornavirus 1 [23-27]. Sanger sequencing and more recently 454 pyrosequencing yield long reads which are desirable for phylotyping of potentially highly divergent sequences. In order to identify novel species from metagenomic sequence, high-quality sequences must be identified. In addition, it is helpful to identify a set of non-redundant sequences prior to similarity search and taxonomic assignment. Short-read platforms such Solexa (Illumina) and SOLiD have been developed more recently and yield more sequence than previous platforms, increasing the sensitivity with which viruses present in low concentrations can be detected, but require assembly of small reads prior to phylotyping in order to accurately assess species origin, and the assembly of short reads in a metagenomic context has proven difficult. I have performed an analysis of Sanger sequencing data in order to get a first-ever look at the diarrhea virome. Diarrhea offers a promising opportunity to apply metagenomic sequencing for virus detection and discovery since an etiologic agent cannot be identified in as many as 40% of cases of diarrhea [3-7].

DNA microarrays as a diagnostic tool

DNA microarrays were developed in the mid-1990s as a powerful tool for the global quantification of gene expression. The inherently parallel nature of microarrays lends itself well to multi-locus interrogation of a single agent as well as simultaneous detection of numerous pathogens [22]. The first microarray for virus detection, called the ViroChip, was developed in 2002 and included 1600 70-mer oligonucleotide probes representing 140 different virus species [28]. The inclusion of both conserved as well as

virus-specific probes on the microarray made it useful for the detection of unknown viruses with similarity to sequenced viruses as well as for the detection known viruses. Since the development of the ViroChip, a handful of diagnostic microarray designs have been proposed [28-37]. Diagnostic microarrays have proven to be effective tools in detecting known viruses in clinical specimens with high sensitivity and specificity [38]. In addition, several viruses have been discovered using pan-viral microarrays, perhaps the most notable of which is SARS coronavirus [29]. Other viruses discovered through the use of diagnostic microarrays include the cardiovirus HTC-UC1 [39], a beluga whale coronavirus [40], an avian bornavirus [41] and a gammaretrovirus (XMRV) [42] which was identified in patients with prostate cancer.

Diagnostic microarray analysis

Early analyses of diagnostic microarray data were based on visual inspection and the use of standard microarray analysis tools such as hierarchical clustering [22]. However, analyzing the many hybridizations generated from screening clinical specimens necessitated the development of robust computational strategies for diagnostic microarray analysis. Regardeless of the chosen platform or probe design strategy, an objective algorithm is required to make use of the wealth of data that comes from each hybridization in order to score and rank potential candidates and to identify infecting viruses. Such an algorithm must account for challenges traditionally associated with microarray analysis including cross-hybridization, probe saturation and sample variation.

The first algorithm expressly designed for interpretation of viral microarrays is E-Predict [43]. E-Predict uses a theoretical energy matrix to compute correlations between

experimental hybridizations and genome-derived energy vectors. Free energies are determined using BLAST to identify significant alignments between probes and viral genomes, followed by a free energy calculation using a nearest-neighbor approach. E-Predict offers different options for normalization of hybridization and intensity vectors including sum, unit-vector and quadratic normalizations. In addition, several options are available for the correlation metric. These include Pearson correlation, Spearman rank correlation and Euclidean distance. One advantage of E-Predict is that it offers an iterative option for detecting multiple viruses. One disadvantage is that accurate p-value calculations are dependent on the accumulation of many previous hybridizations in order to define a null distribution of scores.

DetectiV is a software package for diagnostic microarray analysis that enables visualization, normalization and significance testing for microbial detection [44]. It was developed in and is executed using the R programming environment. Significance testing in DetectiV is based on the comparison of the probe intensities from an experimental hybridization to a user-selected control which can be one of three things: an array's median value for all probes, the mean value of a set of designated control probes, or a control array. Advantages of DetectiV include its accommodation of several different control types and its visualization capabilities. One disadvantage is that DetectiV does not readily deal with the issue of cross-hybridization of probes to similar species.

PhyloDetect is an algorithm that groups candidate targets into a nested hierarchy based on probe-to-genome binding predictions [45]. Predictions are binary such that '1' indicates predicted binding of probe to target and '0' indicates predicted lack of binding. For analysis of an array, hybridization intensities are also made binary. PhyloDetect

computes a likelihood-based test statistic that reflects the number of probes that have ‘0’ indicators for a given group in the hierarchy as well as a user-selected false positive rate parameter. Phylodetect uses the resulting statistic to test whether an organism in the group is present (the null hypothesis), or whether no organism in the group is present (the alternative hypothesis). An advantage of PhyloDetect is its explicit recognition of target similarity by collapsing similar profiles [46]. A possible disadvantage is information loss is its reduction of intensities to binary values.

CLiMax, which stands for Composite Likelihood Maximization, is based on a biophysical model of probe-target hybridization [46]. One feature of CLiMax is that it seeks to identify a set of targets that best explains the observed intensities, explicitly accounting for the possibility of co-infection. It does this using a greedy method to identify a locally optimal set of targets. Probe intensities are modeled with a logisitic regression. Individual probes contribute additively to a likelihood score which is possible under the assumption of independence between probes. One advantage of CLiMax is that it uses probe sequence to identify low-complexity probes which may be less specific and more prone to cross-hybridization. Although CLiMax is capable of recognizing multiple infection, predictions of additional viruses can be suspect [46].

A probabilistic approach for diagnostic microarray analysis

None of the previously published algorithms for analyzing diagnostic microarrays accommodates the use of training data for learning probe-specific behaviors. Training data are empirical observations that can be used to leverage future predictions. The incorporation of training data is a hallmark of machine learning approaches, of which

probabilistic inference is one important example. Probabilistic approaches rely on probabilities which factor into the calculation of statistics and govern the parameterization of a predictive model. Bayesian statistics provide a powerful framework for dealing with parameter uncertainty within a probabilistic model. In a Bayesian approach, what is desired is a posterior probability i.e. *P(Model|Data)* but what in fact what is available is the inverse of that i.e. *P(Data|Model)*. Bayes' rule is the formula that allows a posterior probability to be computed from the available conditional probability. In addition, Bayes' rule involves marginalizing over uncertain parameters. Bayesian inference relies on the use of a prior which defines the pre-existing distribution associated with a given model before any data are observed.

Hidden Markov models (HMMs) allow probabilities to be multiplied together to calculate a probability score for a particular series of events. These events and their associated probabilities are strung together using states connected by transitions. The probabilities associated with each state are called emission probabilities, while the probabilities associated with moving from one state to another are called transition probabilities. HMM state emissions can be derived either from a discrete distribution, as is the case for HMMs whose states emit nucleotides, or they can be derived from a continuous distribution, as would be the case for an HMM whose states emit hybridization intensities. A first-order HMM is one in which the next state in a path is dependent only on the current state.

Dynamic programming algorithms are a class of algorithms that go hand-in-hand with HMMs as an efficient means of fishing out optimal "hidden" paths from the many possible paths defined by an HMM. Dynamic programming algorithms accomplish this

through the use of a matrix in which cumulative probabilities representing various possible paths through an HMM are stored. The four steps in the classic implementation of a dynamic programming algorithm are: 1. Initialization (fills in the first column of the matrix), 2. Iteration (fills in all the remaining columns but the last), 3. Termination (fills in the last column) and 4. Traceback (reconstructs the most likely path).

Probabilistic modeling has had a major impact in numerous applications in sequence analysis including gene finding, profile searches, multiple sequence alignment and regulatory site identification [47]. Applications of probabilistic models to array data include identification of differential gene expression, detection of copy number variation and motif discovery [48-50]. No probabilistic model has been described for the analysis of microarray data for microbial detection. I have developed a Bayesian probabilistic model, VIPR, for analyzing diagnostic microarrays that accommodates the use of training data in the form of hybridizations of known viruses in order to learn probe-specific behaviors and improve detection. I have applied VIPR to a set of hybridizations of hemorrhagic fever viruses in order to assess performance. Moreover, I have adapted VIPR to the detection of recombinant microbes by developing a hidden Markov model. This algorithm, VIPR HMM, was tested using a set of viral encephalitis vaccine strains and represents the first diagnostic microarray analysis tool capable of detecting recombinants.

CHAPTER 2:

Metagenomic Analysis of Human Diarrhea: Viral Detection and Discovery

This work is published in PLoS Pathogens (2008) Feb 29; 4(2): e1000011

Stacy R. Finkbeiner[1†], Adam F. Allred[1†], Phillip I. Tarr[2], Eileen J. Klein[3], Carl D. Kirkwood[4], David Wang[1]

[†]These authors contributed equally to this work.

[1] Departments of Molecular Microbiology and Pathology & Immunology, Washington University School of Medicine, St. Louis, MO USA

[2]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO USA

[3]Department of Emergency Medicine, Children's Hospital and Regional Medical Center, Seattle, Washington, USA

[4] Enteric Virus Research Group, Murdoch Childrens Research Institute, Royal Children's Hospital, Victoria, Australia.

ABSTRACT

Worldwide, approximately 1.8 million children die from diarrhea annually, and millions more suffer multiple episodes of nonfatal diarrhea. On average, in up to 40% of cases, no etiologic agent can be identified. The advent of metagenomic sequencing has enabled systematic and unbiased characterization of microbial populations; thus, metagenomic approaches have the potential to define the spectrum of viruses, including novel viruses, present in stool during episodes of acute diarrhea. The detection of novel or unexpected viruses would then enable investigations to assess whether these agents play a causal role in human diarrhea. In this study, we characterized the eukaryotic viral communities present in diarrhea specimens from 12 children by employing a strategy of 'micro-mass sequencing' that entails minimal starting sample quantity (<100 mg stool), minimal sample purification and limited sequencing (384 reads per sample). Using this methodology we detected known enteric viruses as well as multiple sequences from putatively novel viruses with only limited sequence similarity to viruses in Genbank.

INTRODUCTION

While traditional sequencing approaches are designed to characterize genomes of a single species of interest, metagenomic approaches, such as mass sequencing, transcend species boundaries allowing one to explore the makeup of microbial communities. Such methods provide a holistic look at microbial diversity within a given sample, completely bypassing the need for culturing [21, 51-54]. Previous efforts in this field have explored the structure of virus communities in ecosystems as diverse as the ocean [21, 55] and the human gut [16, 17]. To date, the reported metagenomic studies of human stool have been

limited to analysis of 4 specimens collected from 3 healthy patients [16, 17]. To our knowledge, no metagenomic investigation of the viral diversity found in human diarrhea has previously been described. Human diarrhea is the third leading cause of infectious deaths worldwide and is responsible for ~ 1.8 million deaths in children under age five each year [56]. Bacteria, protozoa and viruses have all been implicated as causal agents. Chief among the known etiologic agents are rotaviruses, noroviruses, astroviruses, and adenoviruses [57]. However, it is estimated that on average up to 40% of diarrhea cases are of unknown etiology, suggesting that unrecognized infectious agents, including viruses, remain to be discovered [3-7]. Mass sequencing affords an opportunity to explore the viral diversity (including both known and novel viruses) present in stool during acute episodes of diarrhea in a systematic and unbiased fashion, thereby laying the foundation for future studies aimed at assessing whether any novel or unexpected viruses detected play a causal role in human diarrhea.

In this study, mass sequencing was applied to explore specifically the viral communities present in pediatric patients suffering from diarrhea. We anticipated that the viral communities would vary significantly from specimen to specimen and that it would be desirable to sample broadly from multiple patients to obtain an overall perspective on the diversity of viruses that might be present. Toward this end, a simple yet robust experimental strategy was developed that circumvented certain technical and economic limitations of conventional mass sequencing. In both previous viral metagenomic studies of the human gut, large quantities of fecal matter (~500g) were collected from adults and then extensively purified to enrich for viral particles [16, 17]. In contrast, pediatric samples provide considerably smaller volumes of stool; therefore, our strategy was

designed to minimize the number of physical purification steps so that as little as 30 mg of archived fecal matter could be analyzed. Here we present data generated by performing what we refer to as 'micro-mass sequencing' of several hundred sequence reads per sample from 12 different patients with acute diarrhea. This analysis provides evidence for the detection of known enteric viruses, viral co-infections, and novel viruses.

RESULTS

Aggregate library analysis

Metagenomic analysis was carried out on fecal samples collected from 12 distinct pediatric patients suffering from acute diarrhea. Patient characteristics are shown in Table 2.1. A sequence independent PCR strategy was employed to amplify the extracted nucleic acids from each sample [29]. 384 clones were sequenced for each sample library. Because the goal of this project was to define the diversity of viruses present in the clinical specimens regardless of their relative abundance, nearly identical sequence reads were clustered to generate a set of non-redundant sequence reads. Unique, high quality sequence reads were then classified into broad taxonomic groups based on the taxonomy of the most frequent top scoring BLAST matches for each sequence. A total of 4,608 sequences were generated, of which 3,169 passed through a quality filter and 2,013 of those contained unique sequence information. Of the unique sequences passing through the filter, 1,457 (72%) could be identified by similarity to sequences in the Genbank nr database based on tBLASTx (E-value $\leq 10^{-5}$) alignments. The remaining 556 (28%) sequences had no significant similarity to any sequences in the nr database and were

Table 2.1. Sample Information					
Sample	**Year Collected**	**Age of Patient**	**# of high quality sequence reads**	**# of unique reads**	**Average unique read length (bp)**
D01	2005	14 mo	365	166	526
D02	1998	10 mo	193	87	536
D03	1984	NA	302	281	506
D04	1984	4 mo	311	154	626
D05	1980	NA	243	168	563
D06	2003	11 mo	153	132	393
D07	1999	23 mo	352	186	617
D08	1999	35 mo	302	167	255
D09	1981	NA	302	294	491
D10	1983	20 mo	195	146	447
D11	1978	NA	253	103	367
D12	2005	8 mo	198	129	300

therefore categorized as being of 'unknown' origin. The 1,457 identifiable sequences were further classified into categories based on their proposed origin (Fig. 2.1). 519 (35.6%) were most similar to eukaryotic viruses, 25 (1.7%) to phage, 857 (58.8%) to bacteria, 3 (0.2%) to fungi, and 20 (1.4%) to human sequences. The remaining 33 (2.3%) were most similar to sequences that did not fall into the other previous categories and were consequently labeled as 'other'. For example, some of the sequences had significant hits to mouse, fish, and plant genomes.

Individual library statistics

384 clones were sequenced for each individual sample. The proportion of high quality sequences for each sample varied between 40% and 95% of the total clones (Table 2.1). The percentages of unique sequences per sample ranged from 41% to 97% of the high quality reads (Table 2.1). The average length of the unique, high quality sequences ranged from 255 to 626 bp. Viral sequences constituted between 0-100% of the reads in

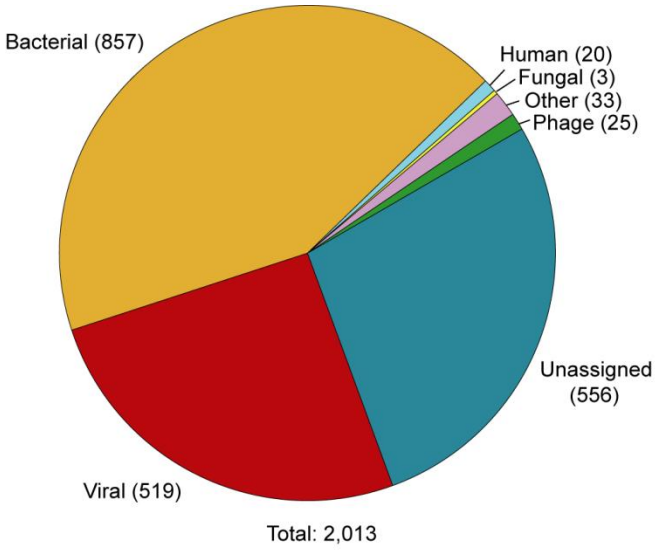

Fig. 2.1. Compositie analysis of all sequences. Sequences from all 12 libraries were categorized based on the best tBLASTX scores (E-value: $<10^{-5}$) as viral, phage, bacterial, human, fungal, other, or unassigned. Numbers in parenthesis represent the number of sequences in each category.

each library (Fig. 2.2). Some libraries (e.g., D01 and D05) were predominantly composed of viral sequences (64% and 95% respectively), whereas others consisted largely of bacterial (e.g., D08 and D12) or unassigned (e.g., D03 and D07) sequences. Based on the initial BLAST classification criteria, sequences with similarity to viruses from 7 different viral families and three unclassified genera (picobirnavirus, anellovirus and mimivirus) were detected in the 12 different samples (Fig. 2.2). Five of the samples (D03, D05, D06, D08, and D12) contained sequences from at least two different virus families known to infect humans.

Detection of known viruses

The first specimen analyzed was a positive control stool specimen that had tested positive for rotavirus (D01) by enzyme immunoassay. It was our expectation that this sample would yield sequences derived from the infecting rotavirus. In this library, 107 non-redundant sequence reads were identified as viral in origin, almost all of which possessed ≥ 90% amino acid (aa) BLAST identity to known rotavirus sequences in Genbank. The

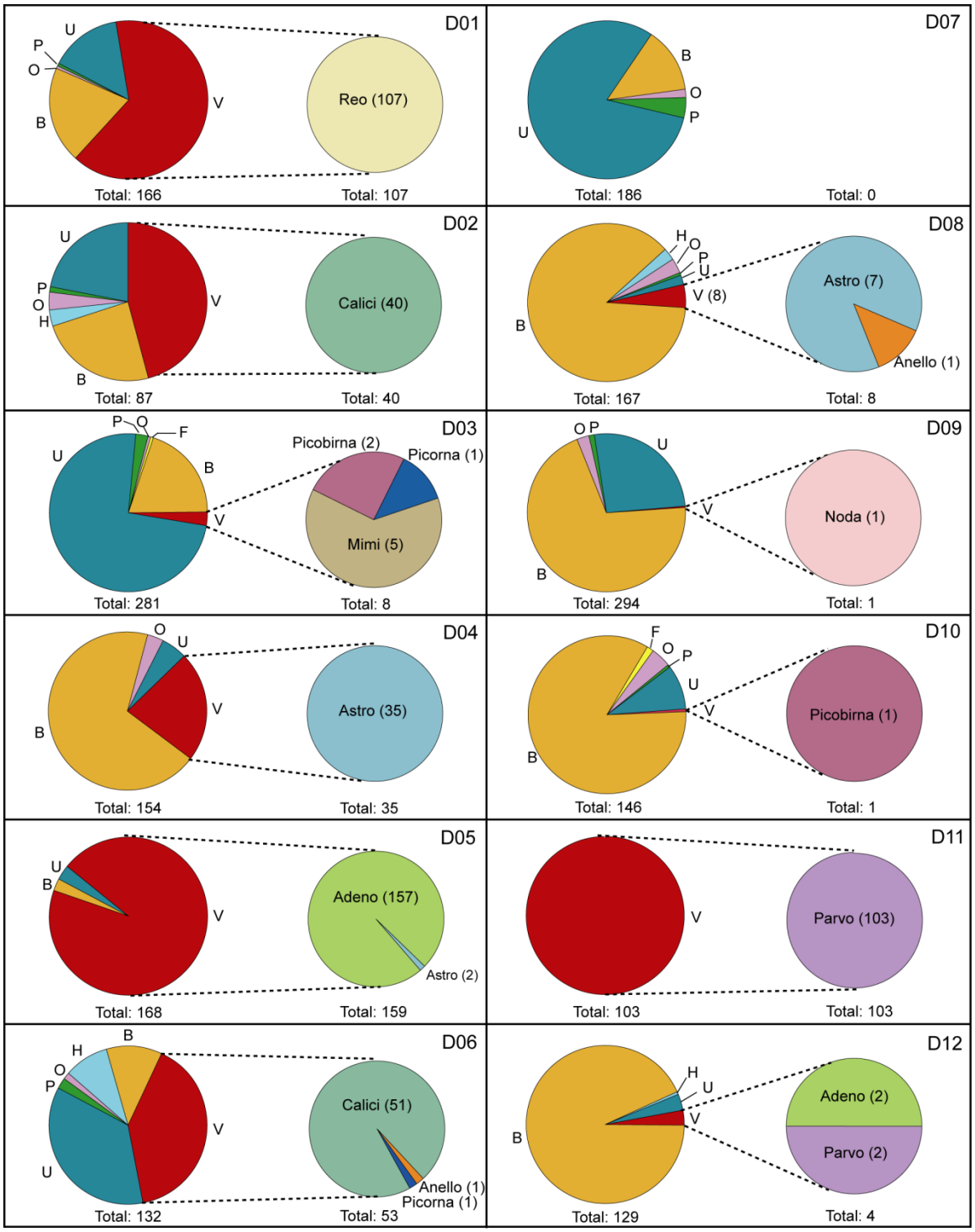

Fig. 2.2. Categorization of sequence reads based on best tBLASTX scores (E-value: $<10^{-5}$). Pies on the left side of each box depict the categorization of sequences from individual samples by phylotype: viral (V); phage (P); bacterial (B); human (H); fungal (F); other (O); and unassigned (U). Pies on the right side of each box depict further characterization of viral sequences by viral families/taxa: *Reoviridae* (Reo); *Caliciviridae* (Calici); *Astroviridae* (Astro); anellovirus (Anello); picobirnavirus (Picobirna); *Picornaviridae* (Picorna); mimivirus (Mimi); *Nodaviridae* (Noda); *Adenoviridae* (Adeno); *Parvoviridae* (Parvo). Numbers in parentheses indicate the number of sequence reads in each category.

sequence data included cloned fragments from all 11 RNA segments of the rotavirus genome.

An additional 11 stool specimens were then selected that had tested negative in conventional PCR and enzyme immunoassays for the known diarrhea viruses (rotaviruses, caliciviruses, astroviruses, and adenoviruses). Despite such screening, sequences derived from the canonical enteric viruses were detected in a number of samples. For example, calicivirus sequences were detected in D02 and D06, astrovirus sequences in D04, and adenoviruses were detected in D05 and D12. Almost all individual sequence reads in these cases possessed >90% aa identity to existing viral sequences in Genbank.

Adeno-associated virus (AAV), a member of the *Parvoviridae* family, was detected in two samples, D11 and D12. These viruses are known to infect the gastrointestinal tract, but are not thought to be enteric pathogens. For productive infections or reactivation from a latent state, AAV requires co-infection with a helper virus that is most commonly an adenovirus or less typically, a herpesvirus [58]. In D12, adenovirus sequences were detected. No additional viruses were detected in D11.

Detection of novel virus sequences

In many of the libraries, individual sequence reads were detected that possessed ≤ 90% aa identity to their highest scoring BLAST hit (representative sequences are listed in Table 2.2) suggesting that these sequences might be derived from novel viruses. In part because BLAST alignments are based on local sequence comparisons, BLAST is not an optimal method for making taxonomic assignments. In order to more accurately and precisely

Table 2.2. Selected sequence reads with limited BLAST identity to known viruses				
Sample	**Sequence Read Accession #**	**Identity to top hit**	**Top Hit (Accession #)**	**Virus Family/Taxa**
D03	ET065742	78%	Human picobirnavirus strain 1-CHN-97 (AF246939)	Picobirnavirus
D03	ET065743	90%	Human coxsackievirus A19 (AF499641)	Picornaviridae
D06	ET067042	74%	Human enterovirus 91 (AY697476)	Picornaviridae
D06	ET067045	66%	TTV-like mini virus (AB026931)	Anellovirus
D06	ET067040	79%	Snow Mountain virus (AY134748.1)	Caliciviridae
D06	ET067041	88%	Norovirus C14 (AY845056.1)	Caliciviridae
D08	ET065575	57%	Human astrovirus 4 (AY720891)	Astroviridae
D08	ET065582	67%	Human astrovirus 5 (DQ028633)	Astroviridae
D08	ET065578	45%	TT virus (AB041963)	Anellovirus
D09	ET066010	35%	Epinephelus septemfasciatus nervous necrosis virus (AM085331)	Nodaviridae
D10	ET066456	81%	Human picobirnavirus 2-GA-91 (AF245701)	Picobirnavirus

assess the relationship of these sequences to known viruses, we generated phylogenetic trees using the maximum parsimony method [59]. In cases where more than one sequence read hit the same region of a genome, only one representative sequence read is listed in Table 2.2 and phylogenetic trees are shown for only these representative sequences (Fig. 2.3-2.4 and Fig. 2.S1-2.S4). Phylogenetic analysis revealed that many of the sequences were divergent from known sequences on the order that approximated a distinct subtype or genotype (Fig. 2.S1-2.S4). This included two libraries with picobirnaviruses (D03, D10) (Fig. 2.S1), two with picornaviruses (D03, D06) (Fig. 2.S2), two with anelloviruses (D06, D08) (Fig. 2.S3), and one with a norovirus (D06) (Fig. 2.S4).

In several instances, much more highly divergent sequences were detected that suggested that novel virus species might be present. The library generated for sample D08 included 7 unique sequence reads derived from two loci that displayed 52-67% aa identity to human astroviruses. Phylogenetic analysis of the individual sequence reads suggested that a novel astrovirus was present in D08 (Fig. 2.3). These sequence reads were assembled into two contigs, one of ~800 bp that mapped to ORF1a and one of ~500 bp that mapped to ORF1b. RT-PCR and subsequent sequencing of the amplicon

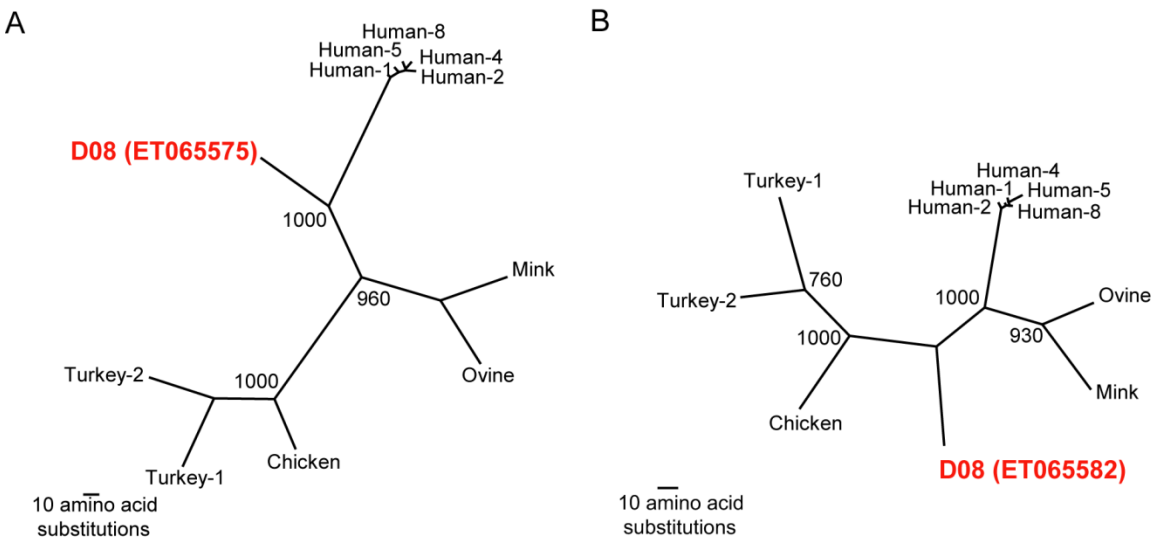

Fig. 2.3. Phylogenetic analysis of highly divergent astrovirus-like sequence reads. Maximum parsimony phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to the corresponding sequences from known astroviruses. 1,000 replicates were generated with bootstrap values over 700 shown. A) Representative sequence read mapping to astrovirus serine protease ORF (Accession number ET065575); B) Representative sequence read mapping to astrovirus RNA polymerase (Accession number ET065582).

confirmed the presence of the contigs in the original RNA extract as well as the contig assemblies (data not shown). Phylogenetic analysis of the two contigs yielded trees essentially identical to those generated from the individual sequence reads (data not shown).

In sample D09, we detected one sequence read which exhibited limited similarity to viruses in the family *Nodaviridae* (Table 2.2). RT-PCR of this sample using primers designed from the sequence read confirmed the presence of a 229 bp fragment in the original RNA extract (data not shown). Phylogenetic analysis of the sequence of the RT-PCR product demonstrated that the nodavirus in sample D09 was highly divergent from other known nodaviruses (Fig. 2.4).

Finally, one sample, D03, contained five sequence reads that, based on the top tBLASTX hits, contained 47% to 52% aa identity to endonuclease genes in the amoeba-

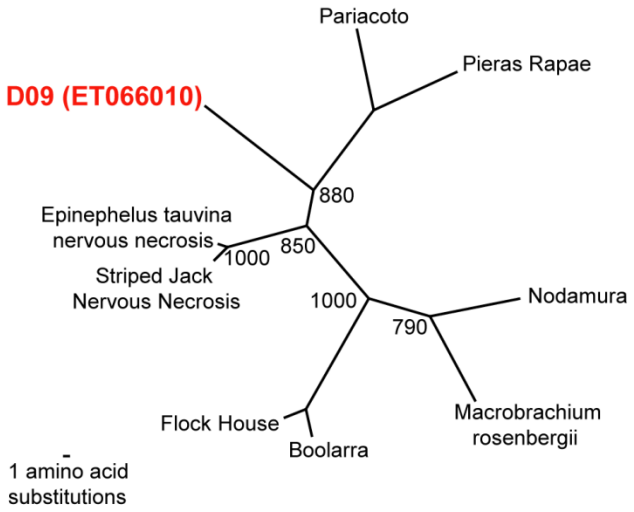

Fig. 2.4. Phylogenetic analysis of a highly divergent nodavirus-like sequence read. Maximum parsimony phylogenetic trees were generated by comparing the translated amino acid sequence of one sequence read (Accession number ET066010) to the corresponding RNA polymerase sequences of nodaviruses. 1,000 replicates were generated with bootstrap values over 700 shown.

infecting virus *Acanthamoeba polyphaga* mimivirus. These sequences also possessed approximately similar levels of sequence identity to a number of bacterial genomes and phage genomes containing putative endonuclease proteins. Phylogenetic analysis comparing the sequence reads to the top scoring BLAST hits (Fig. 2.S5) did not conclusively clarify the origin of these sequences. Further experimentation will be required to unambiguously determine if these sequences are derived from a mimi-like virus, phage, or a bacterial species.

Unassigned sequences

Some sequences in the libraries had no significant hits to any sequences in the Genbank nr database. Samples D03 and D07 had a large abundance of these ‘unassigned’ reads. Relaxing E-value thresholds for designating various sequence categories resulted in the ability to classify a greater number of these unassigned sequences; however, many of

these classifications likely represent artifactual alignments. Viral assignments remained largely unaffected, even when E-value thresholds as permissive as 10 were applied.

DISCUSSION

We examined the diversity of viral communities in stools from 12 children with diarrhea using a strategy we describe as 'micro-mass sequencing'. This strategy, which entails crude purification of fecal suspensions, nucleic acid purification, random PCR amplification, and cloning and sequencing of several hundred colonies, effectively detected known enteric viruses, viral co-infections, and novel viruses. In most traditional metagenomic studies, large sample volumes are subjected to multiple stages of filtration and purification before sequencing. For example, in previous metagenomic studies of the gut, 500 g of fecal samples were initially collected for the analyses. Because clinical pediatric diarrhea specimens are much more limited in volume, we chose to both minimally purify the samples and to employ a random PCR amplification strategy. These combined steps enabled us to rapidly generate sequencing libraries from small quantities of archived stools (30-100 mg). Furthermore, we wished to sample broadly from multiple patients because of the large number of viruses known or suspected to be associated with diarrhea. Therefore, rather than sequence few specimens in great depth as has been done previously (10,000 sequences per sample) [17], we focused on sequencing fewer clones (384 per sample) from more samples (12 specimens).

Our analysis detected viruses, bacteria, host, phage and other sequences (Fig. 2.1). The presence of non-viral sequences in the libraries was not surprising as only minimal efforts were made to enrich for viral sequences. In fact, the goal of this strategy was to

manipulate the specimens as little as possible in the interest of simplicity. Even so, in a few libraries, 100% of the sequence reads were of viral origin. Additional processing, such as treating the specimens with DNase, reduced the background signal and increased the percentage of viral reads in some instances (data not shown).

Viral sequences were detected in all but one sample. Interestingly, a number of DNA viruses (bacteriophages, adenoviruses, and adeno-associated viruses) were detected in our analysis, despite our use of a methodology focused on purification of RNA. While it is possible that RNA transcripts from these viruses were purified [60], it is more likely that viral DNA was co-purified with RNA, as is common in other RNA purification methods [61]. PCR analysis of samples D05 and D11 in the absence of reverse transcription, yielded positive results for adenovirus and adeno-associated virus, respectively, indicating that viral DNA was present in the RNA preparations (data not shown).

Analysis of this initial cohort of 12 specimens yielded a wealth of original findings. In contrast to previous metagenomic studies of stool [17], a number of known human viruses were detected in these clinical specimens. These included common enteric pathogens such as rotavirus, adenovirus, calicivirus, and astrovirus. In addition, putatively benign adeno-associated viruses (AAV) were also detected which are not generally associated with human diarrhea. Aside from one sample known to contain rotavirus, we intended to analyze the viral communities present in samples that were not infected by known enteric pathogens in order to identify viruses that might be responsible for the unexplained cases of diarrhea. The fact that micro-mass sequencing detected these

canonical viruses in some of the specimens, despite conventional diagnostic testing by EIA and PCR, underscores the sensitivity limits of conventional diagnostics.

Detection of novel viruses

Sequences were detected in this study from at least 9 putatively novel viruses. For 7 of these sequences, the degree of divergence observed based on phylogenetic analysis suggested that they might represent novel virus subtypes or genotypes of picobirnavirus, enterovirus, TT virus and norovirus (Fig. 2.S1-2.S4). Picobirnaviruses belong to an unclassified genus of double stranded RNA viruses and have been detected in fecal matter from human and other animals both with and without diarrhea [62]. Only a limited number of picobirnavirus sequences have been previously described in the literature and thus the identification of two novel picobirnaviruses significantly expands the known diversity of this taxonomic group, underscoring the unrecognized viral diversity inhabiting the human body.

Sequences representing a divergent norovirus were detected in sample D06 (Fig. 2.S4). Phylogenetic analysis of individual sequence reads that mapped to the RNA polymerase and the NS4 regions of human norovirus suggested that these sequences were derived from a novel or unsequenced member of norovirus genogroup 2. In the initial screening by conventional PCR, this sample tested negative for norovirus. Upon closer examination, four mutations were observed in one of the PCR primer binding sites, which plausibly hindered the PCR screening assay [7].

In two samples, much more highly divergent sequences were detected. In D08, phylogenetic analysis of 7 unique sequence reads strongly suggested that a novel

astrovirus species was present (Fig. 2.3). The observed sequence variation between these sequence reads and the known astrovirus genomes greatly exceeds the variation that exists between the 8 known serotypes of human astrovirus, suggesting that this virus is not simply another serotype of the known astroviruses. Astroviruses are non-enveloped, single stranded, positive sense RNA viruses that account for up to 10% of sporadic diarrhea cases [63]. Infections with astroviruses most frequently cause watery diarrhea lasting 2-4 days, and, less commonly vomiting, headache, fever, abdominal pain, and anorexia in children under the age of 2, the elderly, and immunocompromised individuals [64]. The detection of this genetically distinct astrovirus raises the question as to whether or not this is an authentic human virus, and if so, whether or not it is a causal agent of human diarrhea.

Another novel sequence detected appeared by phylogenetic analysis to belong to the family *Nodaviridae*. Nodaviruses are small single-stranded, positive sense, bipartite RNA viruses, divided into two genera, the alphanodaviruses (insect viruses) and the betanodaviruses (fish viruses). Currently, none of the established family members are known to naturally infect mammals although experimental manipulation of the viral genome has enabled viral replication in a wide array of organisms including mammals [65]. While it is tempting to speculate that this might represent the first instance of human infection with a nodavirus, further experimentation such as serological analysis is required to definitively answer this question. Another plausible explanation is that the virus may be present simply as a result of consumption of fish infected by the virus. A prior report describing the presence of plant virus RNAs in human stool has similarly been attributed to dietary exposure [17]. Incidentally, some fish genomic sequences were

detected in this particular sequence library (D09 “other” bin) supporting the possibility of dietary exposure. However, the potential piscine origin of this virus would not necessarily preclude its role as an etiologic agent of human disease.

The micro-mass sequencing approach, like any other experimental methodology capable of detecting novel viruses (such as culture or degenerate PCR), cannot of course by itself determine whether the newly detected agent is pathogenic. However, this strategy can generate novel, testable hypotheses such as “Are these novel viruses involved in the etiology of human diarrhea?” and “What is the true host of these viruses?” that could not be asked in the absence of the knowledge that these viruses existed.

Unassigned reads

556 out of the 2013 (28%) unique high quality sequences were binned as unassigned by the BLAST criteria. Of these, 23 were identified as containing repetitive elements or low-complexity sequence by RepeatMasker [66, 67] thus explaining the lack of meaningful BLAST alignments. The origin of the remaining 533 sequences that were unassigned is uncertain, but they could be derived from unannotated host genome, novel or unsequenced microbes, or dietary sources which have not been sequenced. However, it is also possible that some of these sequences could represent viruses that have no appreciable similarity to sequences of currently known viruses. Extracting more telling information from these sequences is a challenging problem that will require the development of new computational measures capable of detecting more distant evolutionary relationships than is possible with existing methods. In addition, as more

genome sequences from diverse organisms and other genomic/metagenomic projects become available, sequence similarity based methods may identify a greater fraction of these currently unassigned sequences.

Diagnostic Applications and Implications

Our data suggest that micro-mass sequencing might be of great diagnostic utility for a number of reasons. First, viruses escaping detection in conventional assays were detected by micro-mass sequencing. In theory, the sensitivity of this strategy is limited only by the depth of sequencing. As demonstrated here, even shallow sequencing performed better than conventional diagnostics in some instances. In addition, the unbiased nature of the method enabled detection of viruses not conventionally tested for. Moreover, co-infections were detected in multiple samples. Furthermore, for multi-segmented viruses such as rotaviruses, reassortment of segments between species is a major mechanism of viral evolution that can lead to the emergence of more virulent strains [68]. Complete genome sequencing of all segments simultaneously would yield completely unambiguous identification of the viral genotype. In contrast to typical PCR or antibody based assays that target a single segment or protein, micro mass sequencing detected all 11 genomic RNA segments of rotavirus. In terms of technical practicality, samples were only minimally manipulated relative to traditional metagenomic sequencing [17, 21, 52, 55], thereby avoiding the time, labor, and use of specialized equipment required to concentrate the specimens, rendering this methodology potentially amenable to use in diagnostic laboratories. As sequencing costs diminish and efficiencies improve, mass sequencing could become a powerful diagnostic tool.

CONCLUSIONS

We have shown that micro-mass sequencing can define the diversity of viral communities found in fecal samples from diarrhea patients. Both known viruses and novel viruses were detected by sequencing only a few hundred colonies from each sample library. These studies will serve as the springboard for further interrogations of the roles of these diverse viruses in the gastrointestinal tract. Finally, our detection of multiple novel viruses in this initial, limited exploration of a dozen samples suggests that broader sampling of patient specimens is likely to be highly fruitful in terms of identification of additional novel viruses.

MATERIALS AND METHODS

Clinical Archived Stool Specimens

Melbourne cohort: Stool samples were collected from children under the age of 5 who were admitted to the Royal Children's Hospital, Melbourne, Victoria, Australia with acute diarrhea between 1978 and 1999.

Seattle cohort: Stool samples were collected between 2003-2005 at the Emergency Department of the Children's Hospital and Regional Medical Center in Seattle, Washington, USA as part of a prospective study attempting to discern the cause of unexplained pediatric diarrhea.

Diagnostic testing of stool specimens for known microbial diarrheagenic agents

Melbourne cohort: Specimens were tested by routine enzyme immunoassays (EIA) and culture assays for rotaviruses, adenoviruses, and common bacterial and parasitic pathogens as previously described [7]. RT-PCR assays were used to screen specimens for the presence of caliciviruses and astroviruses [7, 69] .

Seattle cohort: Specimens were tested for the presence of a number of bacterial species (*Campylobacter jejuni*, *Escherichia coli* O157:H7 and non-O157:H7 Shiga toxin-producing *E. coli*, *Salmonella*, *Shigella*, and *Yersinia)* following standard culture assays, *Clostridium difficile* toxin by a cytotoxicity assay, parasites by microscopy and antigen testing [70]. Additionally, samples were tested by EIA for rotaviruses, adenoviruses, noroviruses 1 & 2, and astroviruses (Meridian Biosciences, DAKO). This study was approved by the institutional review boards of the CHRMC and of Washington University.

Library construction and mass sequencing

Chips of frozen archived fecal specimens (~30-150mg) were resuspended in 6 volumes of PBS. A subset of the archived specimens had been previously diluted and were further diluted 1:1in PBS. The stool suspensions were centrifuged (9,700 x g, 10 minutes) and supernatants were harvested and then passed through 0.45μm filters. RNA was extracted from 100μL of the filtrates using RNA-Bee (Tel Test, Inc., Friendswood, Texas) according to manufacturer's instructions. Approximately, 100-300 nanograms of RNA from each sample was randomly amplified following the Round AB protocol as previously described [29]. The amplified nucleic acid was cloned into pCR4 using the

TOPO cloning kit (Invitrogen, Carlsbad, CA), and transformed into Top10 bacteria. Positive colonies were subcloned into 384 well plates, DNA was purified using magnetic bead isolation, and followed by sequencing using standard Big Dye terminator (v3.1) sequencing chemistry and the universal primer M13 reverse. Reactions were ethanol precipitated and resuspended in 25uL of water prior to loading onto the ABI 3730xl sequencer.

Analysis of sequence reads

Sequence traces were subjected to quality assessment and base-calling using Phred [71, 72]. Lucy [73] was used to trim vector and low quality sequences. Default parameters were used except that high quality sequences identified by Lucy were allowed to be as short as 75 nucleotides. To define the set of reads with unique sequence content in each library, sequences that passed the quality filter were clustered using BLASTClust from the 2.2.15 version of NCBI BLAST to eliminate redundancy. Sequences were clustered based on 98% identity over 98% sequence length, and the longest sequence from each cluster was aligned to the NCBI nr database using the tBLASTx algorithm [74]. An E-value cutoff of 1e-5 was applied. Sequences were phylotyped as human, bacterial, phage, viral, or other based on the identity of the best BLAST hit. Sequences without any hits having an E-value of 1e-5 or better were placed in the "Unassigned" category. All eukaryotic viral sequences were further classified into viral families in similar fashion. Trimmed, high quality sequences that were not found by RepeatMasker to contain repetitive or low-complexity sequence have been deposited in Genbank (Accession numbers ET065304 through ET067293).

Phylogenetic analysis

ClustalX (1.83) was used to perform multiple sequence alignments of the protein sequences associated with select sequence reads. Available nucleotide or protein sequences from known viruses were obtained from Genbank for inclusion in the phylogenetic trees. Selected sequences from Genbank included those with the greatest similarity to the sequence read in question based on the BLAST alignments as well as representative sequences from all major taxa within the relevant virus family. The protein alignments created by ClustalX were input into PAUP [59], and maximum parsimony analysis was performed using the default settings with 1,000 replicates.

Astrovirus trees: Human astrovirus 1 (NC_001943); Human astrovirus 2 (L13745); Human astrovirus 3 (AAD17224); Human astrovirus 4 (DQ070852); Human astrovirus 5 (DQ028633); Human astrovirus 6 (CAA86616); Human astrovirus 7 (AAK31913); Human astrovirus 8 (AF260508); Turkey astrovirus 1 (Y15936); Turkey astrovirus 2 (NC_005790); Turkey astrovirus 3 (AY769616); Chicken astrovirus (NC_003790); Ovine astrovirus (NC_002469); and Mink astrovirus (NC_004579).

Nodavirus tree: Striped Jack Nervous Necrosis virus (Q9QAZ8); Macrobrachium rosenbergii nodavirus (Q6XNL5); Black Beetle virus (YP_053043.1); Flockhouse virus (NP_689444.1); Epinephelus tauvina nervous necrosis virus (NC_004136.1); Nodamura virus (NC_002691.1); Boolarra virus (NC_004145.1); Pariacoto virus (NC_003692.1); and Redspotted grouper nervous necrosis virus (NC_008041.1).

Picornavirus trees: Human coxsackievirus A1 (AAQ02675.1), Human coxsackievirus A18 (AAQ04836.1), Human coxsackievirus A19 (AAQ02681.1), Human coxsackievirus A21 (AAQ04838.1), Human coxsackievirus A24 (ABD97876.1), Human poliovirus 1 (CAD23059.1), Human coxsackievirus A2 (AAR38840.1), Human coxsackievirus A4 (AAR38842.1), Human coxsackievirus A5 (AAR38843.1), Human coxsackievirus A16 (AAV70120.1), Human enterovirus 89 (AAW30683.1), Human enterovirus 91 (AAW30700.1), Human enterovirus 90 (BAD95475.1), Human enterovirus 71 (CAL36654.1), Echovirus 1 strain Farouk (AAC63944.2), Human coxsackievirus B2 (AAD19874.1), Human enterovirus 86 (AAX47040.1), Human coxsackievirus B5 (AAF21971.1), Human echovirus 29 (AAQ73089.1), Human enterovirus 68 (AAR98503.1), Human enterovirus 70 (BAA18891.1), Bovine enterovirus

(NP_045756.1), Porcine enterovirus A (NP_653145.1), Porcine enterovirus B (NP_758520.1), Simian enterovirus A (NP_653149.1), Human rhinovirus A (ABF51203.1), Human rhinovirus B (NP_041009.1).

Picobirnavirus trees: Human picobirnavirus strain 1-CHN-97 (AF246939.1), Human picobirnavirus strain 4-GA-91 (AF246940.1), Human picobirnavirus strain Hy005102 (NC_007027.1), Human picobirnavirus strain 2-GA-91 (AF245701.1), Human picobirnavirus strain 1-GA-91 (AF246612.1), Porcine picobirnavirus 2 (EU104360.1).

Anellovirus trees: TGP96 Torque teno virus (AB041962), Pt-TTV8-II Torque teno virus (AB041963), CBD231 TTV-like mini virus (AB026930), Mf-TTV9 Torque teno virus (AB041959), Mf-TTV3 Torque teno virus (AB041958), KC009/G4 Torque teno virus (AB038621), TA278/G1 Torque teno virus (AB008394), Pt-TTV6 Torque teno virus (AB041957), TUS01/G3 Torque teno virus (AB017613), PMV/G2 Torque teno virus (AF261761), JT33F/G5 Torque teno virus (AB064606), MD1-073 Torque teno midi virus (AB290918), MD2-013 Torque teno midi virus (AB290919), Tbc-TTV14 Torque teno virus (AB057358), Sd-TTV31 Torque teno virus (AB076001), Fc-TTV4 Torque teno virus (AB076003), Cf-TTV10 Torque teno virus (AB076002), So-TTV2 Torque teno virus (AB041960), At-TTV3 Torque teno virus (AB041961).

Calicivirus trees: Camberwell (AAD33960.1), MD-2004 (ABG49508.1), Carlow(ABD73935.1), Snow Mountain virus (AAN08111.1), Mc37 (AAS47823.1), Hawaii(AAB97767.2), Norwalk(AAB50465.1), Southampton (AAA92983.1), Chiba(BAB18266.1), Hesse(AAC64602.1), BoJena-DEU-98 (CAA09480.1), Murine (AAO63098.2), SU17(BAC11827.1), Dumfries (AAM95184.2), SU25-JPN(BAC11830.1), SU1-JPN(BAC11815.1), Desert Shield (AAA16284.1), Melksham (CAA57461.1), Toronto-24 (AAA18929.1), Sw918 (BAB83515.1), OH-QW101 (AAX32876.1).

Endonuclease-like sequences for D03 tree (mimvirus-like sequences): *Bacteroides caccae* (ZP_01959575.1), *Acanthamoeba* mimivirus (YP_142599.1), *Eubacterium dolichum* (ZP_02077753.1), *Staphylococcus* phage K (YP_024462.1), *Lactobacillus* phage LP65 (YP_164778.1), *Lactococcus* phage bIL170 (NP_047162.1), *Lactococcus* phage r1t (NP_695069.1), *Burkholderia vietnamiensis* G4 (YP_001119011.1), *Streptococcus pyogenes* (NP_607538.1), *Tetrahymena thermophila* (XP_001029162.1), *Bacteroides vulgatus* (YP_001300673.1)

ACKNOWLEDGEMENTS

We would like to thank Henry Huang for helpful advice regarding the phylogenetic analysis. This work was funded in part by the Pilot Sequencing Program sponsored by the

Center for Genome Sciences at Washington University (DW), an NHMRC RD Wright Research Fellowship (CK), and by a USDA Grant NRI 2002-35212-12335 (PT).

Supplemental Figures

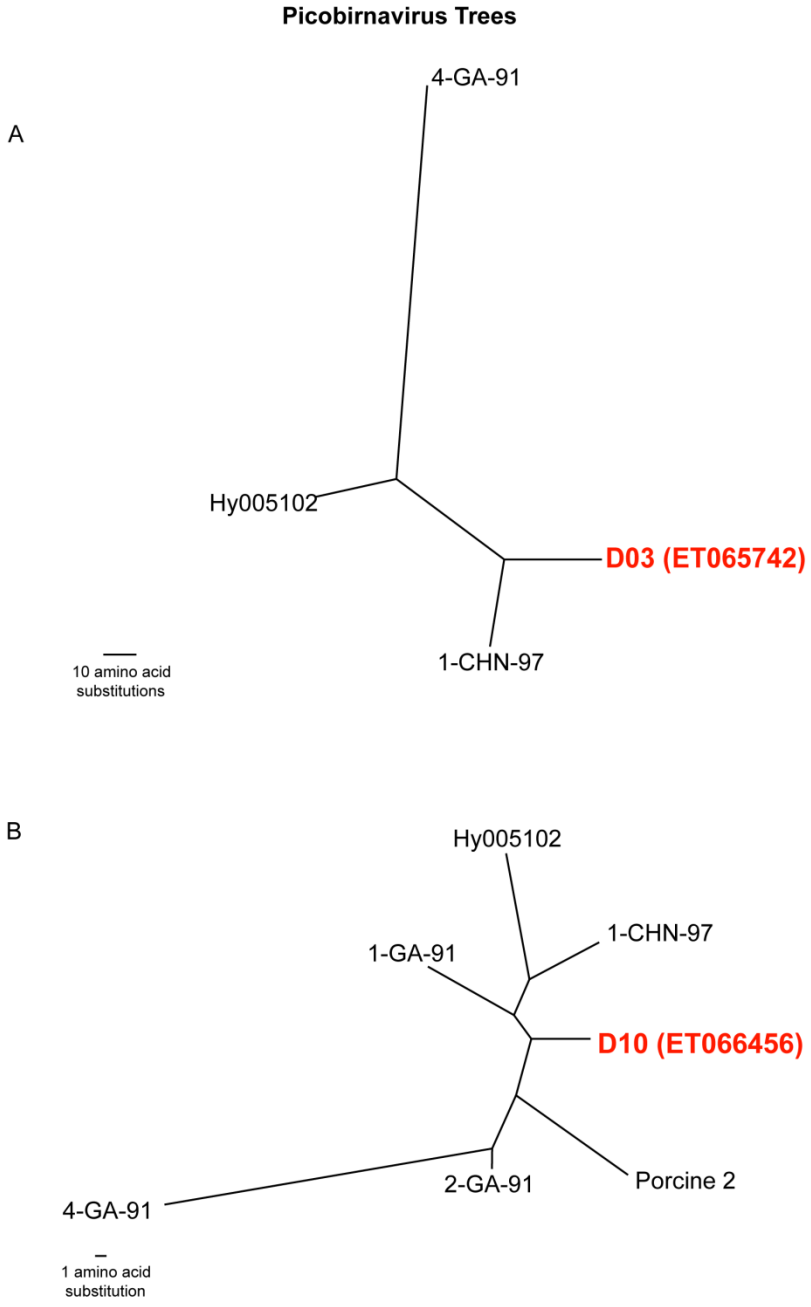

Fig. 2.S1. Phylogenetic analysis of picobirnavirus-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to members of the unclassified taxa picobirnavirus. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown.

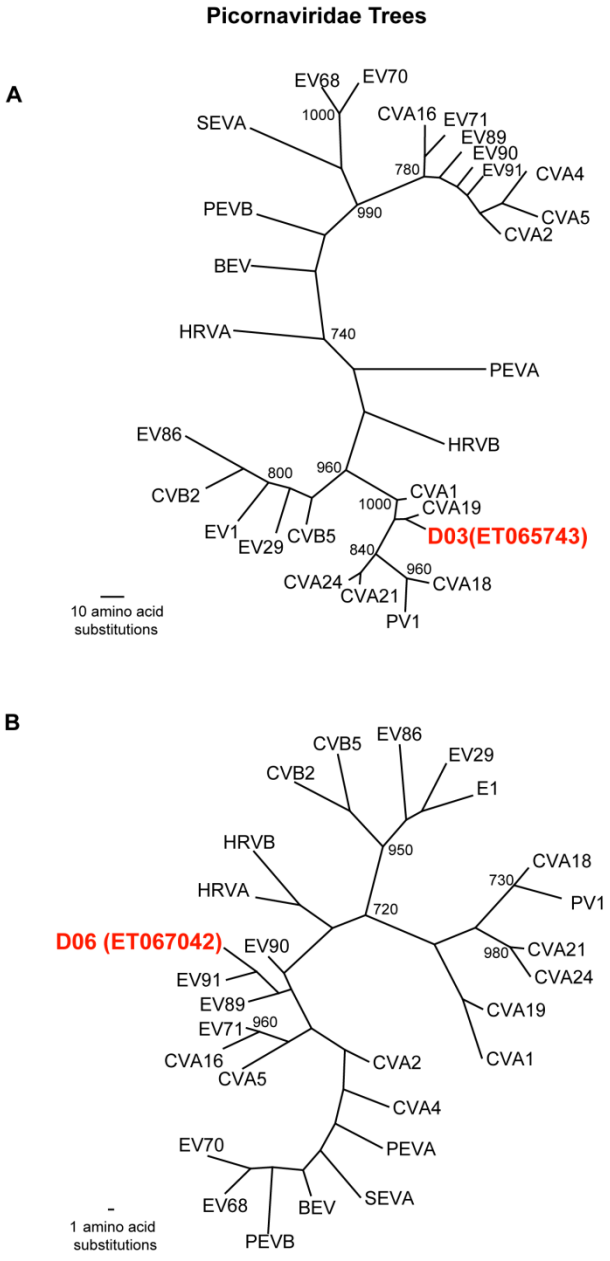

Fig. 2.S2. Phylogenetic analysis of *Picornaviridae*-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to members of the *Picornaviridae* family. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown. CVA=Coxsakievirus A, CVB=Coxsackievirus B, BEV=Bovine Enterovirus, EV=Enterovirus, HRVA=Human Rhinovirus A, HRVB=Human Rhinovirus B, PEV=Porcine Enterovirus, PV=Poliovirus, SEVA=Simian Enterovirus A.

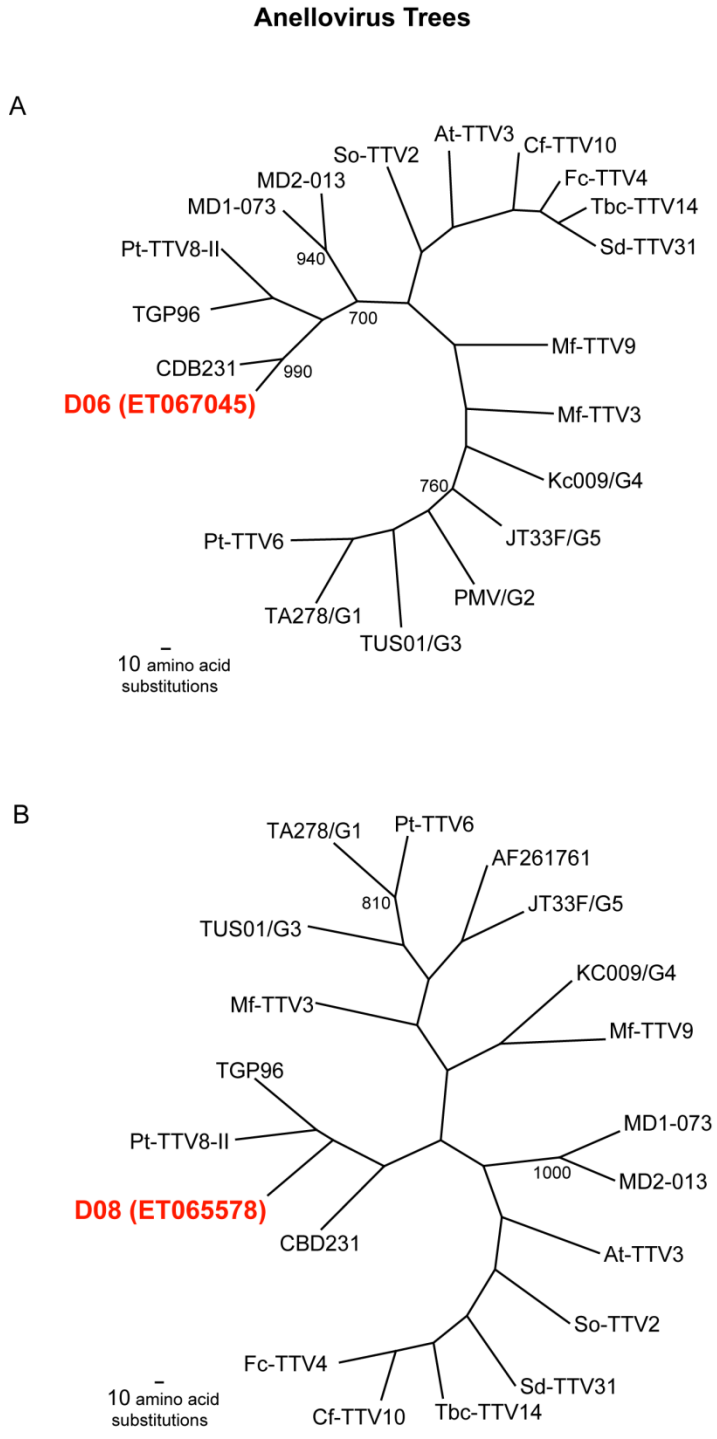

Fig. 2.S3. Phylogenetic analysis of anellovirus-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to anelloviruses. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown.

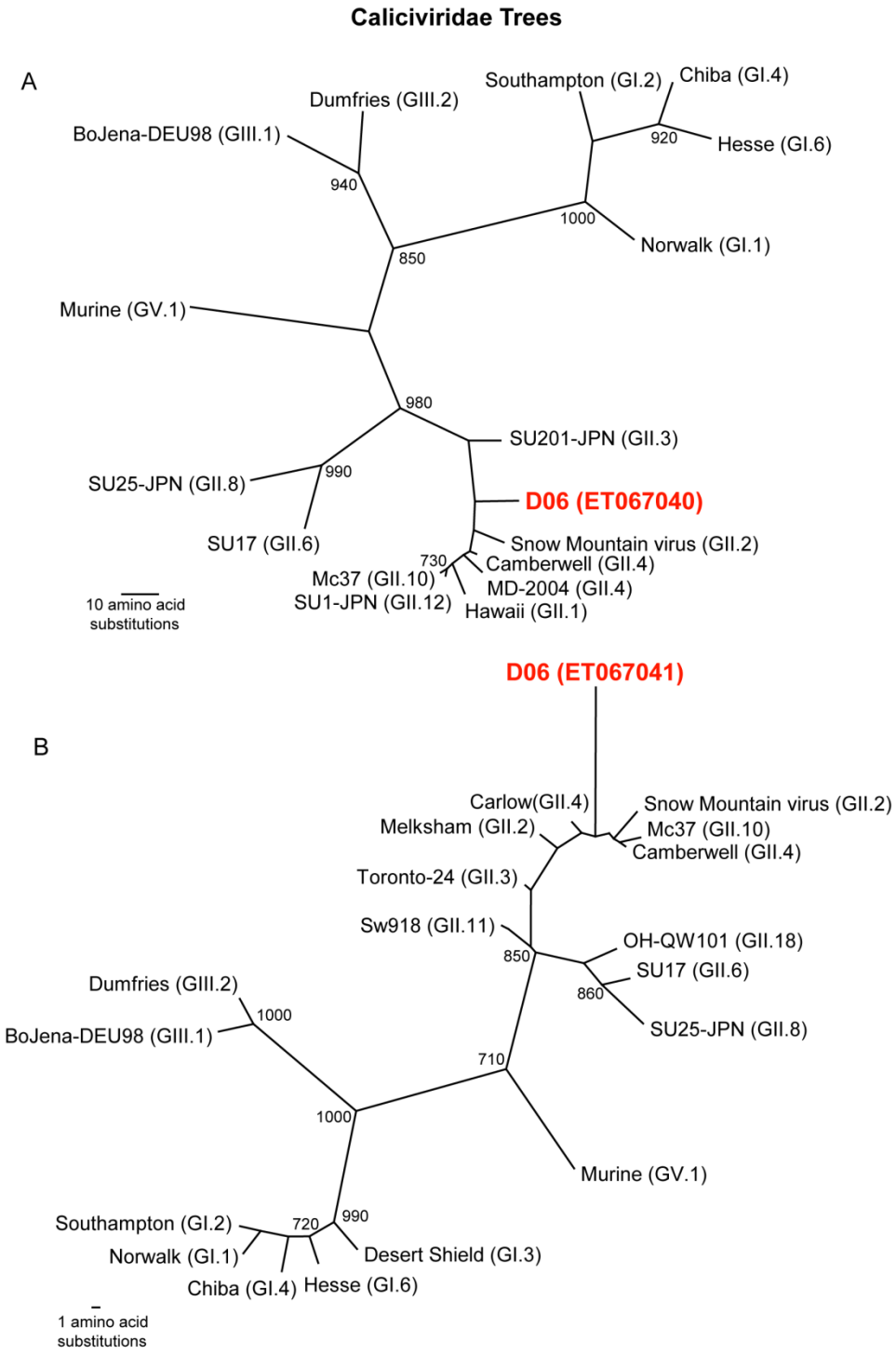

Fig. 2.S4. Phylogenetic analysis of *Caliciviridae*-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of individual sequence reads to the A) NS4 (3A-like) protein or B) NS7 (RNAP) protein of caliciviruses. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown.

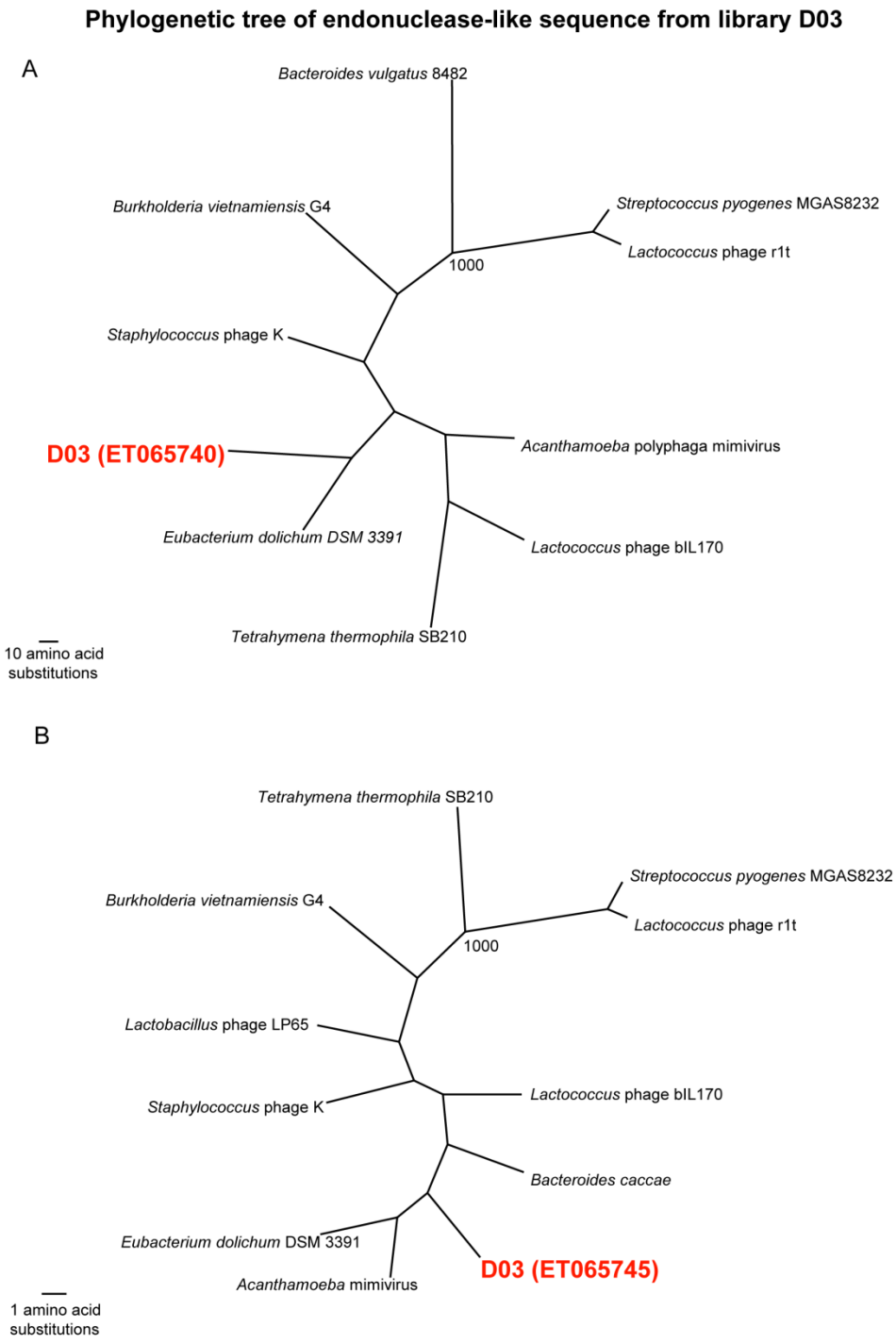

Fig. 2.S5. Phylogenetic analysis of endonuclease-like sequence reads. Phylogenetic trees were generated by comparing the translated amino acid sequence of two individual sequence reads to endonuclease sequences derived from mimivirus, phage, and bacterial species representing some of the top scoring BLAST hits. The trees were created using the maximum parsimony method with 1,000 replicates. Bootstrap values over 700 are shown.

CHAPTER 3:

VIPR: A Probabilistic Algorithm for Analysis of Microbial Detection Microarrays

This work is published in BMC Bioinformatics (2010) Jul 20; 11(1): 384.

Adam F. Allred[1], Guang Wu[1], Tuya Wulan[1], Kael F. Fischer[2], Michael R. Holbrook[3#], Robert B. Tesh[3], David Wang[1]

[1] Departments of Molecular Microbiology and Pathology & Immunology, Washington University School of Medicine, St. Louis, Missouri USA

[2]Department of Pathology, University of Utah School of Medicine, Salt Lake City, Utah USA

[3]Department of Pathology, University of Texas Medical Branch, Galveston, Texas USA

ABSTRACT

All infectious disease oriented clinical diagnostic assays in use today focus on detecting the presence of a single, well defined target agent or a set of agents. In recent years, microarray-based diagnostics have been developed that greatly facilitate the highly parallel detection of multiple microbes that may be present in a given clinical specimen. While several algorithms have been described for interpretation of diagnostic microarrays, none of the existing approaches is capable of incorporating training data generated from positive control samples to improve performance. To specifically address this issue we have developed a novel interpretive algorithm, VIPR (**V**iral **I**dentification using a **PR**obabilistic algorithm), which uses Bayesian inference to capitalize on empirical training data to optimize detection sensitivity. To illustrate this approach, we have focused on the detection of viruses that cause hemorrhagic fever (HF) using a custom HF-virus microarray. VIPR was used to analyze 110 empirical microarray hybridizations generated from 33 distinct virus species. An accuracy of 94% was achieved as measured by leave-one-out cross validation. VIPR outperformed previously described algorithms for this dataset. The VIPR algorithm has potential to be broadly applicable to clinical diagnostic settings, wherein positive controls are typically readily available for generation of training data.

INTRODUCTION

The field of viral diagnostics, which has traditionally followed a "one virus-one assay" paradigm, has been revolutionized by the introduction of diagnostic microarrays [28-37]. It is now possible to test for the presence of thousands of viruses simultaneously in a

single assay. A microarray-based approach is particularly effective for viral diagnosis of diseases that have a common phenotype, but may be caused by any of a number of different viruses. For example, acute respiratory disease, encephalitis and hemorrhagic fever are all disease syndromes known to be caused by a spectrum of viral pathogens. Microarrays specifically focused on the diagnosis of respiratory disease [38, 75-77] and encephalitis [30-32] have been described, as have much broader pan-viral microarrays [28, 29, 35]. A wide range of probe design strategies and microarray platforms can be used for diagnostic microarrays. Independent of the probe design strategy or platform, a key component that is absolutely essential for all diagnostic microarrays is an objective method for interpreting the raw hybridization patterns. While many diagnostic microarrays have been described, there are only three published algorithms, E-Predict [43], DetectiV [44] and PhyloDetect [45], with downloadable or web-accessible software that are available for analyzing data from diagnostic microarrays.

The typical goal of diagnostic virology assays is to determine the presence or absence of one or more viruses from a finite, defined list of candidate viruses known to cause the disease in question. In clinical laboratories, samples of each candidate virus to be detected are typically readily available and can be used as positive controls. Our goal was to develop an interpretive algorithm for diagnostic microarrays that could take advantage of the existence of such positive controls to generate a training data set to guide subsequent analyses.

Toward this end, we developed a probabilistic algorithm for the purpose of analyzing diagnostic microarrays. This class of algorithms has been applied to numerous problems in biology. For example, hidden Markov models (HMMs) and the more

youthful conditional random fields (CRFs) have allowed researchers to make important inferences about sequence structure and function [47, 78]. Bayesian inference in a probabilistic framework offers a distinct advantage of capitalizing on empirical data to guide future predictions as compared with methods that are based solely on computational prediction of genome-to-probe binding. In addition, the power of utilizing probabilities as opposed to discretizing a host of parameters when considering possible solutions means that global calculations are less likely to be influenced by poor choices made locally. To date, no Bayesian algorithm for diagnostic microarrays has been described.

In this paper, we describe a novel probabilistic algorithm that relies on Bayesian inference for analysis of diagnostic microarrays. To validate this approach, we focused on analysis of the set of viruses known to cause hemorrhagic fever. HF symptoms include severe vascular damage, hemorrhage, high fever, and shock and can frequently lead to death [79, 80]. HF viruses belong to four virus families: *Arenaviridae*, *Bunyaviridae*, *Flaviviridae* and *Filoviridae*. A custom microarray was designed to detect all known HF viruses and many of their close relatives. Specimens representing virtually every virus species known to cause HF were procured and hybridized to microarrays for the purpose of validating our algorithm. Furthermore, we compared VIPR's performance to that of the existing interpretive algorithms that are not capable of utilizing training data in this fashion.

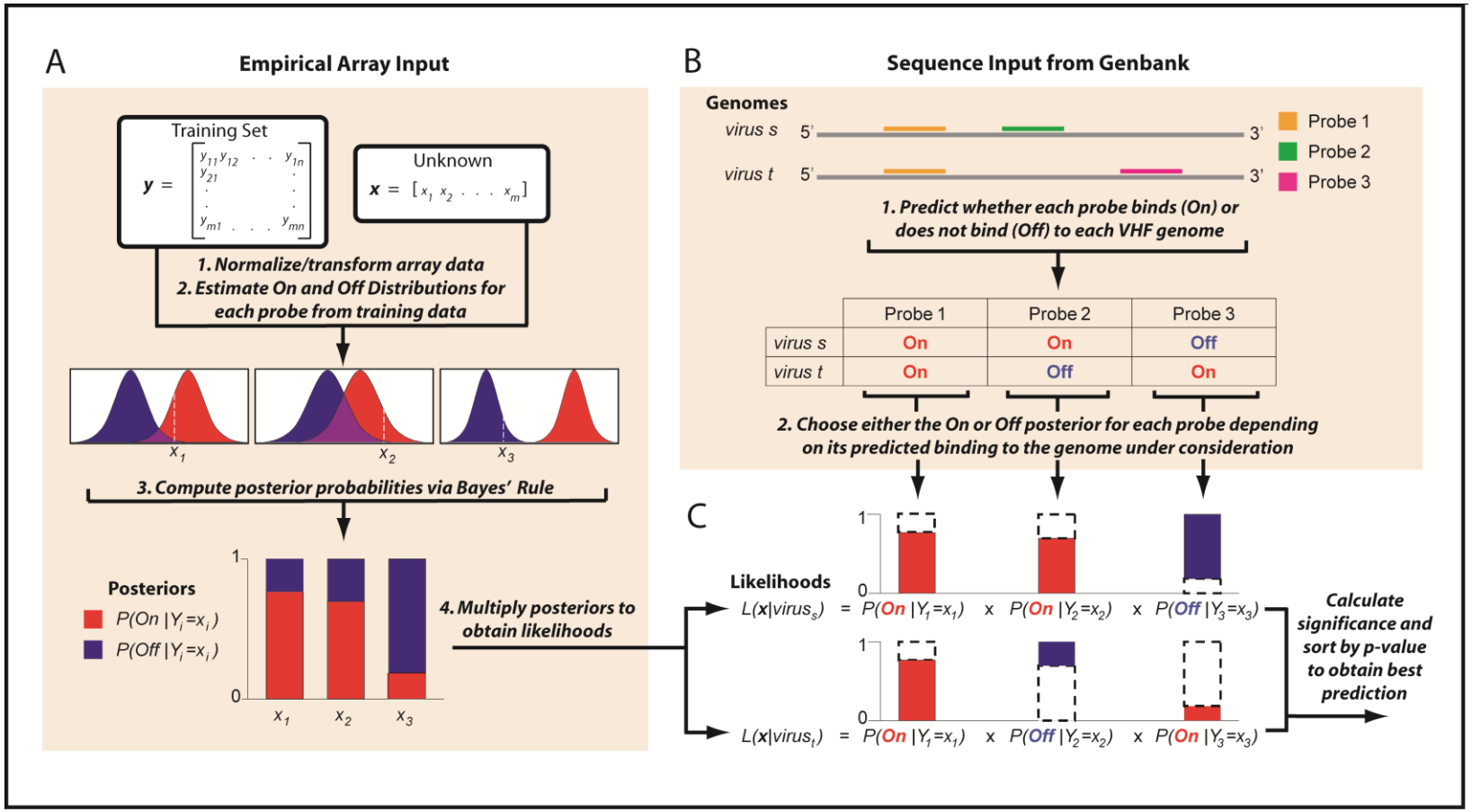

Fig. 3.1. Flow of VIPR. The VIPR probabilistic model incorporates both empirical array data as well as sequence data from GenBank to calculate likelihoods for each candidate virus. A) Posterior probabilities are calculated for each probe. B) The *On* or *Off* posterior is chosen for each probe based on predicted binding to candidate genomes. C) Posteriors are multiplied to obtain a likelihood for each candidate virus.

MATERIALS AND METHODS

Microarray design

14,864 oligonucleotide probes were designed using a taxonomy-based approach as described previously [81] except that the Agilent® 8 x 15K platform was used and probes were 35, 45 or 60 nucleotides in length. The probes were designed to bind to viral genomes from the four families that contain all viruses known to cause HF: *Arenaviridae*, *Bunyaviridae*, *Filoviridae*, and *Flaviviridae*. Probes of different lengths were designed to account for different levels of conservation between viral taxa. For example, longer probes were included to represent regions of strong conservation, while shorter probes were included to distinguish closely related virus species in order to increase specificity.

Hybridization of HF viruses to microarray

A total of 51 strains of 33 distinct virus species (see Table 3.1) obtained from the World Reference Center for Emerging Viruses and Arboviruses were grown in either Vero cells or C6/36 cells. RNA was extracted using standard Trizol® protocols and was randomly amplified as previously described [29]. The resulting amplified material was then coupled to a fluorescent dye and hybridized to the HF microarray. Raw data measurements were collected using GenePix Pro® software. In total, 110 hybridizations were performed (102 positive controls + 4 Vero negative controls + 4 C6/36 negative controls). All raw microarray data are available in NCBI GEO (accession GSM534862 through GSM534971). These 110 hybridizations constituted a set of positive and negative controls used for validation, a subset of which was used in training our algorithm.

Table 3.1. Viruses hybridized to the diagnostic microarray			
Virus	**Family**	**Causes HF**	**# of strains hybridized**
Amapari virus	*Arenaviridae*	No	1
Guanarito virus	*Arenaviridae*	Yes	4
Ippy virus	*Arenaviridae*	No	1
Junin virus	*Arenaviridae*	Yes	1
Lassa virus	*Arenaviridae*	Yes	2
Lymphocytic choriomeningitis virus	*Arenaviridae*	No	1
Machupo virus	*Arenaviridae*	Yes	1
Mobala virus	*Arenaviridae*	No	1
Mopeia virus	*Arenaviridae*	No	1
Sabia virus	*Arenaviridae*	Yes	1
Tacaribe virus	*Arenaviridae*	No	1
California encephalitis virus	*Bunyaviridae*	No	1
Crimean-Congo hemorrhagic fever virus	*Bunyaviridae*	Yes	4
Hantaan virus	*Bunyaviridae*	Yes	1
La Crosse virus	*Bunyaviridae*	No	1
Ngari virus	*Bunyaviridae*	Yes	1
Puumala virus	*Bunyaviridae*	Yes	1
Rift Valley fever virus	*Bunyaviridae*	Yes	3
Seoul virus	*Bunyaviridae*	Yes	1
Toscana virus	*Bunyaviridae*	No	1
Angola marburgvirus	*Filoviridae*	Yes	1
Reston ebolavirus	*Filoviridae*	No	1
Sudan ebolavirus	*Filoviridae*	Yes	1
Zaire ebolavirus	*Filoviridae*	Yes	1
Gabon ebolavirus	*Filoviridae*	Yes	1
Dengue virus 1	*Flaviviridae*	Yes	2
Dengue virus 2	*Flaviviridae*	Yes	2
Dengue virus 3	*Flaviviridae*	Yes	2
Dengue virus 4	*Flaviviridae*	Yes	2
Kyasanur Forest disease virus	*Flaviviridae*	Yes	2
Omsk hemorrhagic fever virus	*Flaviviridae*	Yes	4
Rocio virus	*Flaviviridae*	No	1
Yellow fever virus	*Flaviviridae*	Yes	2

VIPR normalization and transformation (Figure 3.1A)

For each sample in the training set, a unit-vector normalization was applied as shown, where x_i represents the i^{th} intensity for a given hybridization. Then, each normalized intensity was $\log_e$ transformed. As given in Equation (1), x_i^{NT} is the normalized, transformed value for that intensity. Normalization was performed to account for

variation in reagent concentrations or fluorescence across the microarray. Log transformation of the data was desirable for the estimation of normal distributions.

(1) $$x_i^{NT} = \log\left(\frac{x_i}{\sqrt{\sum x_i^2}}\right)$$

Note that in the following calculations, all intensities have been normalized and transformed although the superscript *NT* does not appear.

VIPR prediction of On and Off states (Figure 3.1B)

Candidate genomes to be scored in the VIPR algorithm were limited to all complete genomes in the NCBI virus RefSeq database as of 6/20/2008. The entire set of oligonucleotide probes on the microarray was aligned using BLASTN against each of the RefSeq viral genomes. Theoretical free energies of hybridization were then calculated from the aligned sequences using code included with OligoArraySelector [82]. If the free energy associated with binding of a given viral genome/oligonucleotide pair was computed to be less than -30 kcal/mol, the probe was assigned the *On* state for that genome; otherwise, the probe was assigned the *Off* state. The choice of -30 kcal/mol was based on the observation that this threshold represents the weakest binding reported for long-oligo broad specificity microarrays [82]. A given viral genome was included in the list of potentially detectable candidate viral genomes if at least three oligonucleotide probes were expected to bind to that genome (i.e. were expected to be *On*). A total of 101 candidate genomes met these criteria.

VIPR calculation of posteriors (Figure 3.1A)

Posterior probabilities were calculated for each probe *i* according to Bayes' rule, (2) and (3). Y_i and x_i represent a random variable and an observed intensity, respectively.

$$P(On \mid Y_i = x_i) = \frac{P(Y_i = x_i \mid On)P(On)_{marg}}{\sum_{On,Off} P(Y_i = x_i \mid state)P(state)_{marg}} \quad (2)$$

$$P(Off \mid Y_i = x_i) = 1 - P(On \mid Y_i = x_i) \quad (3)$$

Likelihoods for each probe were determined using normal distributions derived from two sets of normalized $\log_e$ transformed intensities: those corresponding to the *On* states for a given probe (4), and those corresponding to the *Off* states (5).

$$P(Y_i \mid On) \sim N(\mu_{i,on}, \sigma^2_{i,on}) \quad (4)$$

$$P(Y_i \mid Off) \sim N(\mu_{i,off}, \sigma^2_{i,off}) \quad (5)$$

The probe-specific *On* and *Off* distributions are derived from the training set where the probe *On*/*Off* states are defined by the identity of the virus in each hybridization.

VIPR priors

Priors were calculated (6,7) in a probe-specific manner and were designed to incorporate two calculations derived from the composition of the microarray as well as the composition of the set of candidate viruses under evaluation: (a) the percentage of probes predicted to be *On* for the candidate virus under consideration, represented as $P(On)_{pred}$; (b) the number of candidate viruses that share that probe's *On*/*Off* prediction (i.e. if four

candidate viruses, including $virus_s$, are predicted to be *On* for a given probe, then $P(virus_s|On) = ¼$ for that probe). Marginalizing over the possibility of an *On* or an *Off* prediction calls for a second invocation of Bayes' rule:

$$P(On)_{marg} = \frac{P(virus_s \mid On)P(On)_{pred}}{\sum_{On,Off} P(virus_s \mid state)P(state)_{pred}} \quad (6)$$

$$P(Off)_{marg} = 1 - P(On)_{marg} \quad (7)$$

VIPR calculation of hybridization likelihoods (Figure 3.1C)

Because of the possibility of underflow, all likelihood calculations were made in log space, though they are expressed here in probability space. The likelihood (8) of the observed hybridization vector, $\boldsymbol{x}$, was calculated for each of n viral genomes. The posteriors included in the product were chosen so as to reflect the expected state of a particular probe for $virus_s$. *On* states for $virus_s$ are indexed from i=1 to a while *Off* states are indexed from j=1 to b as shown in formulas (9) and (10), respectively.

$$L(\boldsymbol{x} \mid virus_s) = L_1 \times L_2 \quad (8)$$

$$L_1 = \prod_{i=1}^{a} P(On \mid Y_i = x_i) \quad (9)$$

$$L_2 = \prod_{j=1}^{b} P(Off \mid Y_j = x_j) \quad (10)$$

Calculating significance of VIPR results

To determine the significance of the results obtained, we computed a p-value for each candidate virus by permuting the set of priors for the candidate over the set of likelihoods $P(Off \mid Y_i=x_i)$ so as to estimate a null distribution of scores (n=100 permutations) against which the actual score for that candidate could be compared. From the 100 null scores for each candidate virus, a mean and standard deviation were calculated. The resulting p-value reflects the percentage of the null distribution that is greater than or equal to the actual score. When assessing the significance of a given candidate, a Bonferroni correction was applied so that 0.05, a generally accepted level of significance, was divided by the total number of candidate viruses (101) i.e. a candidate was considered significant if its p-value was less than $5x10^{-4}$.

Assessing the accuracy of VIPR

From the total 110 empirical hybridizations, 108 were chosen as suitable for training on the basis of percentage of well-behaved probes among those predicted to be *On*. Two hybridizations of Ippy virus to the array were excluded from training because the percentage of probes designed to bind to Ippy virus that evinced a sufficient separation ($p<0.001$ by student's t-test) between the *On* and *Off* distributions was less than ten percent. For the initial cross-validation, the subset of 108 arrays was divided into a training set consisting of 107 arrays and a validation set consisting of a single array. This was done 108 times, leaving out a different array each time. The two arrays that did not meet the criterion for inclusion in the training set were tested using all 108 selected arrays for training. For each array, the best prediction was determined by sorting significant candidate viruses ($p<5x10^{-4}$) by p-value and then by likelihood. In the case where no

virus was significant, the array was considered negative. Algorithm accuracy was computed using the formula, Accuracy = (TP + TN)/(TP + TN + FP + FN), where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the total number of false negatives. In the case where the fully sequenced genome of a viral subspecies was not available, an accurate prediction on the species level constituted a true positive. There was also one case where the genome of a subspecies (La Crosse virus) was used as a substitute for a hybridized species (California encephalitis virus) because the complete sequence of California encephalitis virus was not available. These designations of species and subspecies are according to NCBI taxonomy.

Exploring alternative priors

In a separate analysis, VIPR's accuracy was assessed over a space of arbitrary priors rather than deliberately specifying priors using Equations (6) and (7). Thus, the marginalized priors $P(On)_{marg}$ and $P(Off)_{marg}$ in Equation (2) were replaced with priors that ranged iteratively from 0.1 to 0.9. For each iteration, one prior pair i.e. *P(On)*, *P(Off)* where *P(Off)=1- P(On)* was chosen for all *On* probes, while a separate pair was chosen for all *Off* probes. Thus, while the prior pair between the *On* and the *Off* probes could differ, the prior pair between any two *On* probes or between any two *Off* probes was the same. Hence, the space explored represents successive iterations of independently varying the *On* prior pair and the *Off* prior pair with variations made at a step size of 0.1. As before, p-values were calculated to assess the significance of VIPR results, except that 20 permutations were run for each candidate virus instead of 100.

Exclusion of replicate hybridizations

Four independent strains of each of the following viruses, Crimean-Congo hemorrhagic fever, Guanarito virus and Omsk hemorrhagic fever virus were cultured in Vero cells. These viruses represent three of the four HF virus families. As with the other viruses in the positive control dataset, these viruses were hybridized in duplicate (3 viruses x 4 strains per virus x 2 hybridizations for each strain = 24 hybridizations). These 24 hybridizations were used to assess the effect that leaving out both replicates of a strain would have on cross-validation. VIPR predictions were made as described for the leave-one-out cross validation except that replicate hybridizations were excluded from training for the subset of 24 arrays. The number of accurate predictions made by VIPR out of the total 24 hybridizations was calculated.

Comparison to existing diagnosis algorithms

Three algorithms, E-Predict, DetectiV and PhyloDetect, were available for comparison to VIPR. E-Predict [43] was used to calculate Uncentered Pearson correlations. A custom E-matrix for the HF dataset was prepared as described by Urisman et al. A given viral genome was included in the list of potentially detectable candidate viral genomes if at least three oligonucleotide probes were expected to bind to that genome. Default normalizations ('Sum' for the intensity vector and 'Quadratic' for the E-matrix) were applied. 110 correlations were used to estimate each null distribution of correlations from the set of HF arrays. These distributions were fit using the Shapiro-Wilk normality test as described [43]. The same significance threshold that was applied to VIPR predictions

($p<5x10^{-4}$) was also applied to E-Predict. Background-corrected intensities were loaded into DetectiV [44] and normalized in two independent ways: first using the median option, and second, against a Vero or C6/36 array serving as a negative control. The filtered results (mean log ratio>1) for each array were then sorted by p-value to determine the top-scoring virus. The same significance threshold that was applied to VIPR predictions ($p<5x10^{-4}$) was also applied to DetectiV. Hybridization intensities were inputted to PhyloDetect [45] as binary vectors where a probe was considered 'present' if its intensity was greater than the median background signal plus twice the background standard deviation. The E-matrix constructed for E-Predict was converted to binary values ($x_i \leq$ -60 kcal/mol→1; otherwise→0). The *fnr* parameter was set to 0.10. Results were sorted first by likelihood and then by number of present probes to determine the top candidate. A likelihood above the threshold 0.05 constituted a positive prediction. The same formula to calculate accuracy for VIPR was used to calculate accuracies for E-Predict, DetectiV and PhyloDetect.

RESULTS

RNA was purified from cell cultures that were infected with each of the viruses shown in Table 3.1. These viruses were selected to include almost all of the viruses known to cause HF; only a few very recently identified HF viruses, such as Chapare virus [83] and Lujo virus [84], were not included. To assess whether these viruses could be distinguished from close relatives that are not associated with HF, additional viruses were also selected from the same families for testing. For each of the 51 virus cultures, following random amplification and fluorescent labeling, two microarrays were hybridized generating a

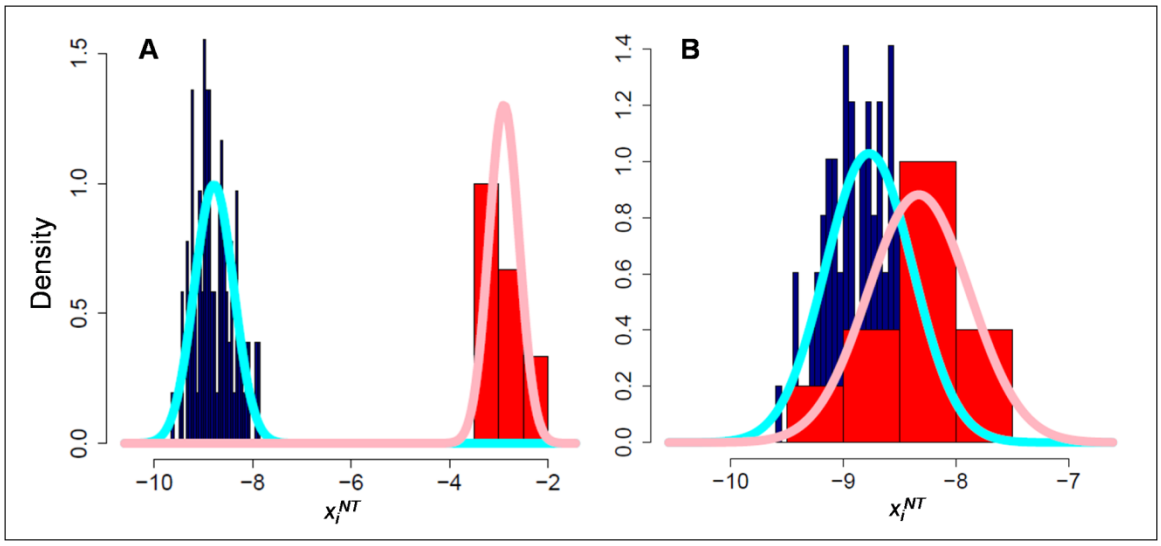

Fig. 3.2. Examples of *On* and *Off* distributions for two probes. A) One representative probe with highly resolved *On* and *Off* distributions based on the training set data. B) One representative probe where the *On* and *Off* distributions overlap. Empirical distributions (blue=*Off*, red=*On*) and estimated distributions (cyan=*Off*, pink=*On*) are shown.

total of 102 empirical hybridizations using virally infected samples. In addition, eight negative control hybridizations (four from uninfected Vero cells and four from uninfected C6/36 cells) were performed.

We developed VIPR as an objective approach for analyzing diagnostic microarray data (VIPR is available for download from http://ibridgenetwork.org/wustl/vipr). VIPR incorporates both sequence data from GenBank as well as empirical array data to classify microarray hybridizations of samples with unknown viral infections (Figure 3.1). From these data, normal distributions corresponding to *On* and *Off* states for each probe were estimated.

Empirical distributions and their normal approximations for two representative probes are shown in Figure 3.2. Figure 3.2A depicts a highly informative probe since there is effectively no overlap between the *On* and *Off* distributions for that probe. In contrast, the distributions in Figure 3.2B overlap substantially. Gradations between these

Table 3.2. Five highest scoring candidates for a Dengue 3 hybridization				
Rank	**Virus**	**Family**	**log(L)**	**p-value**
1	Dengue virus 3	*Flaviviridae*	-352	0
2	Dengue virus 4	*Flaviviridae*	-391	0
3	Dengue virus 2	*Flaviviridae*	-539	0
4	Dengue virus 1	*Flaviviridae*	-599	0
5	Psittacid herpesvirus 1	*Bunyaviridae*	-433	1.0

two extremes constitute probes of intermediate informative value. Posterior probabilities were calculated via Bayes' rule for each probe given the observed intensity from an unclassified array. These posterior probabilities were multiplied to obtain likelihoods for each candidate virus.

For identification of HF viruses, the algorithm was trained using a subset of the total 110 hybridizations. To select a suitable subset for training purposes, we identified 108 hybridizations for which at least 10% of the probes predicted to be *On* had intensities that differed significantly ($p<0.001$) from that probe's *Off* distribution. To assess VIPR's performance on the 108 selected arrays, we performed leave-one-out cross validation so that the selected arrays were divided into a training set (n=107) and a validation set (n=1). The remaining two arrays (not included in the training set) were tested using the entire set of selected arrays (n=108) for training.

An example of VIPR's output for a representative Dengue virus 3 hybridization is shown in Table 3.2. Likelihood scores for all candidate viruses for each microarray are available in the supporting material. We measured the accuracy with which we could make predictions for the virally infected and negative control arrays. VIPR made accurate predictions for 104 out the total 110 arrays. There were five false negatives and one false-positive, corresponding to an accuracy of 94%. The misclassified arrays are shown in Table 3.3.

Table 3.3. The six arrays that were misclassified by VIPR		
False positives		
Chip#	Hybridized virus	Top scoring virus (p<5e-4)
207	Dengue virus 3	Dengue virus 4
False negatives		
Chip#	Hybridized virus	Top scoring virus (p<5e-4)
462	Kyasanur Forest disease virus	none
463	Kyasanur Forest disease virus	none
464	Kyasanur Forest disease virus	none
221	Ippy virus	none
245	Ippy virus	none

For all Bayesian methods, one question that must be addressed is how to choose appropriate priors. From many possible choices, we selected priors in this study based upon the composition of the probes in the microarray as well as the makeup of the candidate genomes under evaluation. In order to define the dependency of our algorithm's accuracy on the choice of priors, VIPR's accuracy was assessed over a range of possible prior pairs, independently varying the pair used for probes expected to be *On* versus the pair used for those expected to be *Off*. Hence, the space explored represents different combinations of prior pairs whose values lie between 0.1 and 0.9 with variations made at step size of 0.1, and with the sum of *P(On)* and *P(Off)* defined as 1.0 for each prior pair. 20 permutations were run for each candidate virus to compute p-values. Results are shown in Figure 3.3. Accuracy varied depending of the choice of priors, but remained fairly stable (between 85% and 97%) over a wide range of prior pairs, suggesting that the method is relatively insensitive to the choice of priors.

For the 24 hybridizations representing four strains from each of three species (Crimean-Congo hemorrhagic fever, Guanarito virus, Omsk hemorrhagic fever virus), a second cross-validation was performed in which both replicates corresponding to a particular strain were excluded from training when making VIPR predictions for those

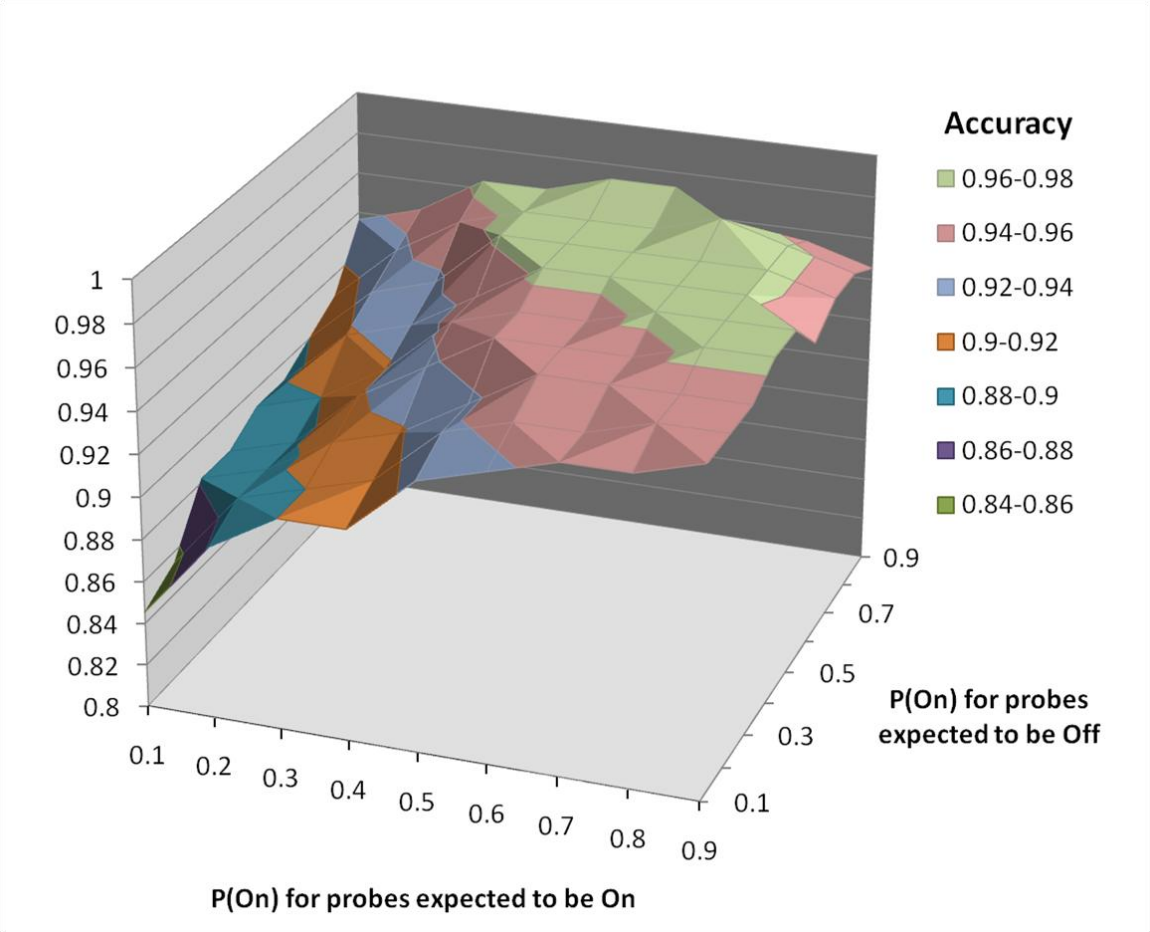

Fig. 3.3. Cross-validation results for different combinations of prior pairs.

arrays. Leaving out both replicates for these particular strains was possible because there remained three other positive control strains of the same species in the training set. This could not be done in the case where only one strain of a species was present among the positive controls because it would render the training set devoid of any representatives of that species. VIPR analysis of the subset of arrays that represent viruses where multiple strains are present in the training set demonstrated robust prediction (24/24 arrays accurately predicted).

We compared the performance of VIPR to that of existing algorithms for analyzing diagnostic microarrays. E-Predict [43], the first algorithm expressly designed

for interpretation of viral microarrays, uses a theoretical energy matrix to compute correlations between experimental hybridizations and genome-derived energy vectors [43]. As shown in Table 3.4, VIPR (94% accuracy) outperformed E-Predict (61% accuracy) for the same set of positive and negative control arrays. One possible explanation for E-Predict's low performance for this set of arrays is the lack of sufficient data to estimate accurate null distributions of scores by the Shapiro-Wilk criterion to be used to calculate p-values. For this dataset only 110 arrays were available for the estimation of null distributions for E-Predict, whereas over one thousand arrays were used by Urisman et al. [43] to calculate these distributions. This is supported by the fact that the virus with the highest raw score as determined by E-Predict is the true virus for 84 of the 102 positive control arrays.

DetectiV [44] is an R-based method for significance testing for microbial detection microarrays. Significance testing involves data normalization against one of the following: an array's median value for all probes, the mean value of a set of designated control probes, or a control array. No designated control probes, in the sense described by the DetectiV algorithm, were included in our design; therefore, the median and control array normalization options were used to analyze our data. After performing significance testing, the results were filtered to exclude groups whose mean log ratio was less than or equal to one. Sorting the filtered results by p-value then revealed a best prediction for each array. An accuracy of 69% was achieved using the median normalization method. Higher accuracies were achieved using the negative arrays with the control array normalization option. These accuracies ranged from 76% to 83% depending on which of the eight uninfected samples in our dataset was used as the control array.

Table 3.4. Accuracy of VIPR compared to other methods for this dataset	
Algorithm	**Accuracy (%)**
VIPR	94
DetectiV	76-83
E-Predict	61
PhyloDetect	49

PhyloDetect has previously been applied to viral diagnostic microarrays by increasing its 'false negative rate' parameter [45]. PhyloDetect, unlike VIPR, E-Predict and DetectiV, requires its hybridization inputs to be binary. To achieve this, we created a binary vector for each array where a probe was given a value of '1' if its intensity was greater than the median background signal plus twice the background standard deviation, and '0' otherwise. The theoretical microbial candidate profiles required for PhyloDetect are also binary. While the authors of PhyloDetect applied a stringent predicted binding energy threshold (-80 kcal/mol or less) to make binary present/absent predictions, our probe set, which included probes ranging in length from 35 to 60 nucleotides, could not tolerate such a stringent cutoff without resulting in some candidates having zero probes predicted as 'present.' Thus, we predicted a present probe when the corresponding binding energy was calculated to be -60 kcal/mol or less. After analysis of our data, we computed an accuracy of 49% for PhyloDetect.

DISCUSSION

The inherently parallel nature of DNA microarrays lends itself well to diagnostic applications seeking to simultaneously test for many microbial agents. While many diagnostic microarrays have been described [28-37] there is a relative lack of methods to objectively interpret these microarrays.

One key feature of a true diagnostic microarray is that the targets to be detected are typically well defined. Thus, specimens infected with these targets should be available for use as positive controls. In this study, we developed a novel interpretive algorithm for analysis of diagnostic microarrays that takes advantage of the existence of positive controls that can serve as a training set. VIPR performed with high accuracy (94%) as measured by leave-one-out cross validation. Since VIPR outperformed E-Predict, DetectiV and PhyloDetect for this dataset, this underscores the utility of using a set of known viruses together with a probabilistic algorithm to diagnose viral disease. Though we have not applied our algorithm to other diseases, we anticipate that this strategy would similarly be preferable to a non-Bayesian approach for diagnosis of other diseases of multiple etiologies whose microbial spectrum is well defined and for which positive and negative control specimens are available.

Only one false positive resulted from the cross-validation, which was a Dengue virus 3 sample being classified as Dengue virus 4. Dengue virus 3 was the second best prediction for this array, with both Dengue virus 4 and Dengue virus 3 achieving a p-value of 0.0. The other five microarrays that were misclassified by VIPR, all of which were called as virus negative, were derived from three virus cultures. None of these samples was accurately classified by E-Predict, DetectiV, or PhyloDetect. Given that these samples evaded accurate classification by all three algorithms, one possibility for the lack or detection of virus in these samples is the samples used as positive controls may have been present in abundance below the sensitivity limit of the microarrays. Another plausible explanation is that all or most of the probes designed to detect these viruses do not behave as predicted. In this case, redesigning the probes for these viruses

would be the best way to improve the accuracy of the platform. Comparing the *On* and *Off* distributions for probes designed to bind to these viruses reveals that among those viruses that were hybridized to the array, Ippy virus and Kyasanur Forest disease virus exhibited the highest percentage of *On* probes that displayed no significant difference ($p<0.001$) in intensity between the *On* and *Off* distributions (94% and 85% respectively). However, since VIPR's accuracy is inherently limited by the performance of the probe set, and the response of the probe set is determined by the identity and abundance of the target microbes, we are unable to distinguish between the possibilities of low-titer virus and misbehaving probes.

Other potential caveats related to our method include a limited ability to estimate the true intensity distribution of *On* states for a probe because of the small number of intensities in the training set that correspond to an *On* state. Hence, one way to improve the accuracy of estimation of these distributions would be to increase the number of positive control arrays in the training set. Depending on the degree of sequence divergence among the known strains of a given virus, it may also be important to represent the known diversity of related strains in the training set. However, we emphasize that even with the limited number of microarray hybridizations performed in this study, 94% accuracy was achieved.

The choice of prior probabilities could also be problematic in some circumstances. We found that prior estimation based on predicted binding of probes to viral genomes resulted in robust virus prediction. Moreover, accuracy remained fairly stable (between 85% and 97%) over a wide range of prior combinations. Another potential caveat with the VIPR algorithm is that the distribution of the $\log_e$ of the

intensities was assumed to be normally distributed. Gross violations of this assumption could have pejorative effects on prediction.

One limitation of a leave-one-out cross-validation in our case is that there is a possibility of overfitting due to the presence of replicate hybridizations in the training set. However, an analysis of a subset of arrays that represented several different strains of viruses (Crimean-Congo hemorrhagic fever virus, Guanarito virus and Omsk hemorrhagic fever virus) demonstrated that removing both replicate hybridizations for a given strain from the training set while retaining those from the other strains resulted in accurate prediction in every case. This subset of viruses represented three of the four families of HF viruses. While this analysis does not completely rule out the possibility of overfitting, it clearly demonstrates that VIPR can make accurate predictions even when replicate arrays are removed from training, as long as hybridizations representing strains from the same species are present. Additionally, VIPR outperformed the other three algorithms for this subset. E-Predict, DetectiV and PhyloDetect accurately classified 14, 16, and 8 of the 24 arrays, respectively.

While the results of our study represent a proof of principle using carefully controlled positive and negative controls for validation, it is anticipated that a probabilistic algorithm will be useful in clinical laboratory settings to analyze microarrays like the one described. Testing VIPR using clinical datasets will be the focus of future studies. In the case of diseases for which samples representing *in vivo* human infections are available, such would be the desired dataset for training. In the case of HF, however, clinical specimens from human infections are not generally available; therefore, it will be necessary to investigate the use of different kinds of specimens as training data

for the probabilistic algorithm. These datasets could include specimens from infected animals or viruses harvested from culture and spiked into human sera.

As currently implemented, VIPR only looks for single virus effects. Possible improvements to the software might include the addition of functionality to detect the presence of co-infections and reassortant viruses. This could be accomplished by including among the list of candidates for which likelihoods are computed theoretical combinations of sets of *On* posteriors from different viruses. Equations (11) through (14) extend the single-virus likelihood calculation implemented by VIPR to the case where two viruses, *s* and *t* are present.

(11) $$L(\boldsymbol{x} \mid virus_{s,t}) = L_1 \times L_2 \times L_3$$

(12) $$L_1 = \prod_{i=1}^{a} P_s(On \mid Y_i = x_i)$$

(13) $$L_2 = \prod_{j=1}^{b} P_t(On \mid Y_j = x_j)$$

(14) $$L_3 = \prod_{k=1}^{c} P_{s,t}(Off \mid Y_k = x_k)$$

CONCLUSIONS

We developed a probabilistic algorithm that relies on a training set of empirical hybridizations that accounts for probe-specific behaviors. Application of this algorithm to a dataset of cultured viruses that cause HF resulted in high accuracy virus identification. Though we report the application of VIPR only in the context of diagnosis of HF, our method of detection is theoretically applicable to any microbial detection problem in

which a set of positive and negative control hybridizations is available. Our implementation of a probabilistic algorithm demonstrates the power of a Bayesian approach for discerning important hybridization signals from a complex mixture of nucleic acids. This, in turn, should prove to be of great value as microarray-based diagnostics play more prominent roles in clinical and public health laboratories.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant U01 AI070374. We would like to thank Gary Stormo, Bill Shannon, and Rob Culverhouse for useful discussions.

CHAPTER 4:

VIPR HMM: A Hidden Markov Model for Detecting Recombination with Microbial Detection Microarrays

This work is currently in review with Bioinformatics.

Adam F. Allred[1], Hilary Renshaw[1], Scott Weaver[2], Robert B. Tesh[2] and David Wang[1]

[1]Departments of Molecular Microbiology and Pathology & Immunology, Washington University School of Medicine, St. Louis, Missouri USA

[2]Institute for Human Infections and Immunity, Center for Biodefense and Emerging Infectious Diseases and Department of Pathology, University of Texas Medical Branch, Galveston, Texas USA

ABSTRACT

Current methods in diagnostic microbiology typically focus on the detection of a single genomic locus or protein in a candidate agent. The presence of the entire microbe is then inferred from this isolated result. Problematically, the presence of recombination in microbial genomes would go undetected unless other genomic loci or protein components were specifically assayed. Microarrays lend themselves well to the detection of multiple loci from a given microbe; furthermore, the inherent nature of microarrays facilitates highly parallel interrogation of multiple microbes. However, none of the existing methods for analyzing diagnostic microarray data has the capacity to specifically identify recombinant microbes. In previous work, we developed a novel algorithm, VIPR, for analyzing diagnostic microarray data using a training set of empirical hybridizations of infected and uninfected samples. We have expanded upon our previous implementation of VIPR by incorporating a hidden Markov model (HMM) to detect recombinant genomes. We trained our HMM on a set of nonrecombinant parental viruses and applied our method to 11 recombinant alphaviruses and 4 recombinant flaviviruses hybridized to a diagnostic microarray in order to evaluate performance of the HMM. VIPR HMM correctly identified 95% of the 62 inter-species recombination breakpoints in the validation set and only two false positive breakpoints were predicted. This study represents the first description and validation of an algorithm capable of detecting recombinant viruses based on diagnostic microarray hybridization patterns. VIPR HMM could enhance our ability to rapidly identify novel recombinant viruses arising naturally or engineered as biological weapons.

INTRODUCTION

Recombination constitutes an important source of genetic variation among viruses. As an evolutionary mechanism, recombination leads to new viral genotypes with potentially novel biological properties and/or clinical manifestations. Vaccine-derived poliovirus is one example of a virus for which recombination may play an important role in the progression of disease. Recombination between vaccine-derived poliovirus and coxsackie virus has been shown to increase neurovirulence of recombinant progeny and may be responsible for the emergence of pathogenic vaccine-derived poliovirus [85]. In addition, H1N1 influenza and Ngari viruses provide examples in which novel genotypes consisting of genomic segments derived from multiple different parental viruses have led to disease outbreaks. H1N1, the influenza virus responsible for the 2009 outbreak of pandemic flu, is thought to have arisen from the successive reassortment of four different strains of influenza A [86]. Ngari virus, a hemorrhagic fever-causing bunyavirus, is thought to have resulted from the natural reassortment of two viruses, Bunyamwera and Batai viruses, neither of which is known to cause hemorrhagic fever [87, 88]. Given that recombination and reassortment can play important roles in producing novel variations that are implicated in pathological outcomes, the ability for clinicians to identify novel recombinant and reassortant viruses in diagnostic laboratories is highly desirable.

In addition to occurring naturally through evolution, recombinant and reassortant viruses can also be deliberately created in the laboratory. In vitro recombination has proven to be a useful tool for engineering novel viruses with properties desirable for the development of vaccines [89, 90]. However, this also means that recombination and reassortment have the potential to be used maliciously to develop novel agents of

bioterrorism. Such agents could be engineered as highly pathogenic new viral genotypes consisting of the components of previously described viruses including non-pathogenic viruses. Anticipating the possible use of recombinant/reassortant-based bioweapons should guide our efforts in preparing to respond to such attacks. In such cases, the ability to detect novel agents quickly and accurately would be critical. Thus, it is imperative that any assay used to detect agents of bioterrorism include novel recombinants and reassortants as possible outcomes.

Microarrays are well suited to detecting recombination and reassortment and have an important advantage over traditional diagnostic methods because they allow for the interrogation of multiple loci from multiple viruses in parallel. Traditional methods for microbial detection, such as PCR and antibody based methods, are generally limited to detecting only one genome segment or one protein per assay. The inference is then made that the entire genome is present given that a small part of the genome (or proteome) was detected. Unless other loci are specifically assayed, this diagnostic paradigm does not account for the possibility that a recombinant or reassortant virus is present. There have been many reports of the efficacy of microarrays as a tool for viral diagnosis and discovery [28-37, 91]. While many different probe design strategies and platforms have been proposed for diagnostic microarrays, all approaches require an objective method for interpreting the raw hybridization patterns.

The method must be able to make diagnostic calls in the presence of technical noise, biological noise (i.e. cross-hybridization to host) and probe saturation. Published examples of such methods with downloadable or web-accessible software include E-Predict [43], DetectiV [44], PhyloDetect [45], CLiMax [91] and VIPR [92]. While these

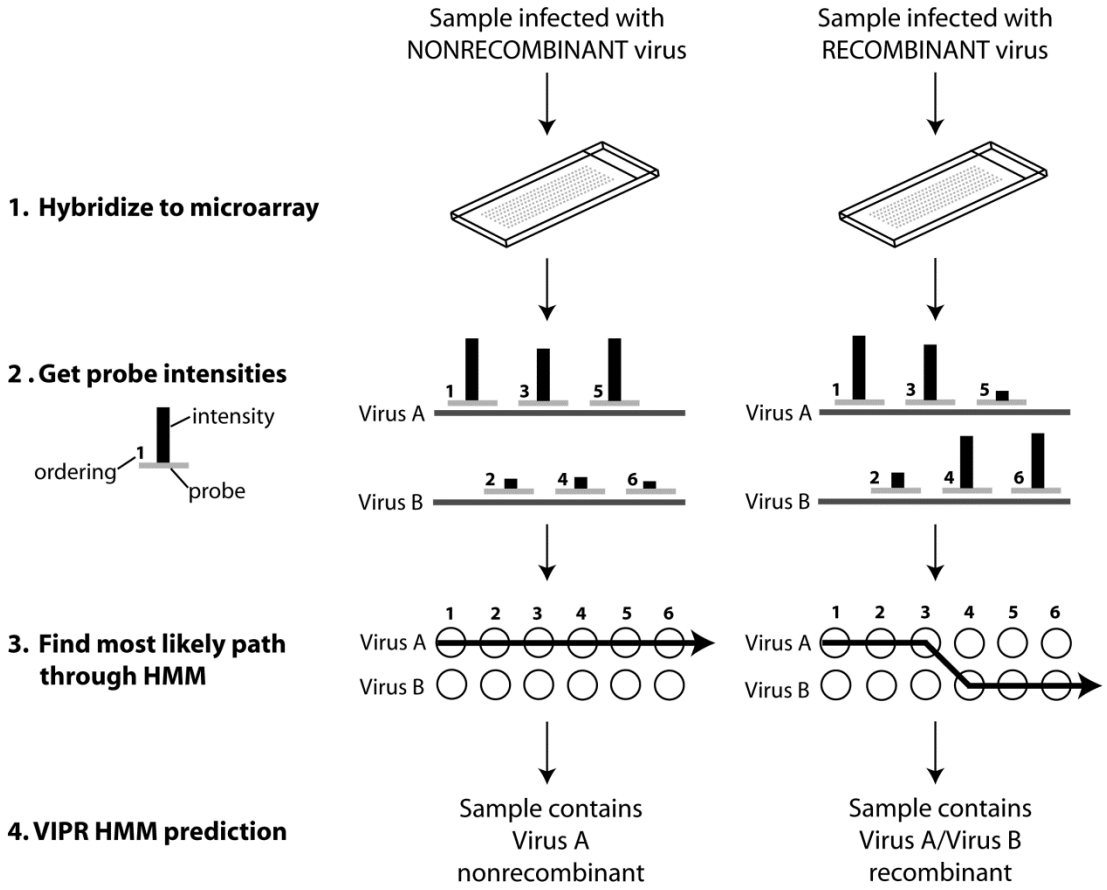

Fig. 4.1. Overall strategy for using an HMM to identify recombinant and nonrecombinant viruses hybridized to a microarray. Probe intensities indicative of binding can implicate the presence of a single virus (left) or the presence of different viruses for different loci (right). This pattern of intensities can be used to identify an optimal path through an HMM whose states represent binding or non-binding events between probes (columns) and virus genomes (rows). Nonrecombinant paths, such as the one on the left, involve transitions only between states in the same row, while paths that move from one row to another are indicative of recombination (as exemplified in the path on the right).

methods have been shown to perform with high accuracy, none of them was designed to be able to identify novel recombinant or reassortant viruses from a hybridization pattern.

One feature of VIPR, which stands for **V**iral **I**dentification with a **PR**obabilistic algorithm, is that it relies on an empirical training set of positive and negative control hybridizations to leverage diagnostic predictions. In this paper, we describe the expansion of VIPR to accommodate the possibility of recombination between candidate viruses

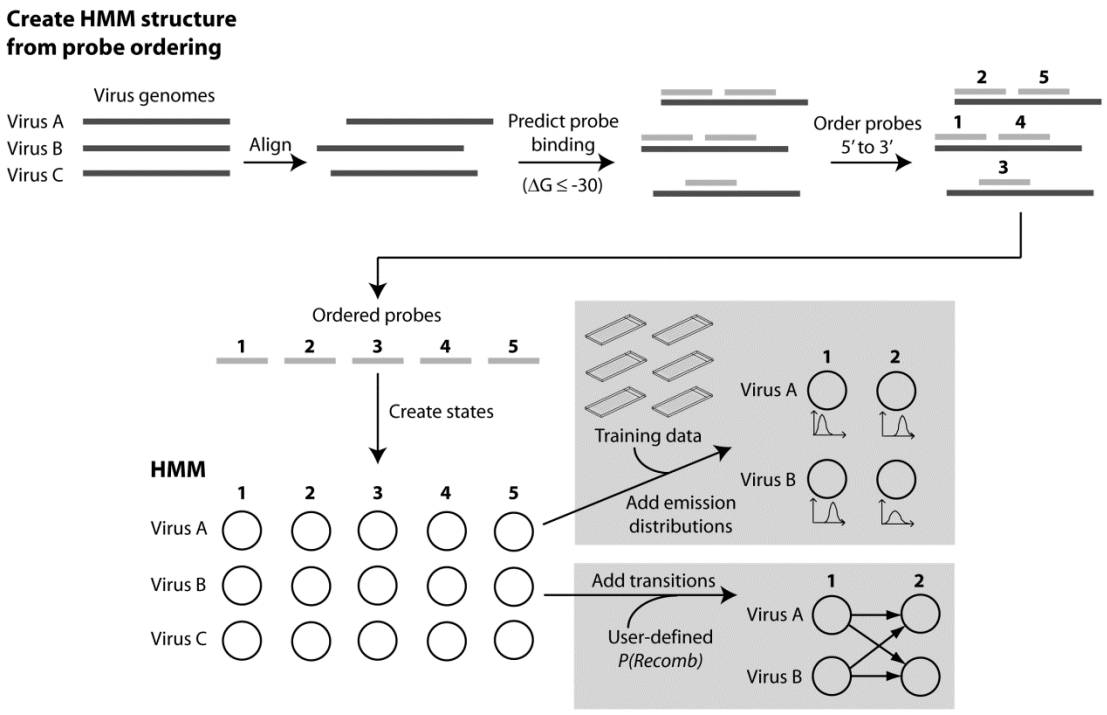

Fig. 4.2. Structure of the HMM used to detect recombinant and nonrecombinant viruses. First, candidate virus genomes are aligned. Probes are then mapped to their respective positions in the multiple alignment based on predicted free energy of binding in order to achieve a universal ordering of probes. A state is created for each probe:genome combination (representing either a predicted binding or non-binding event). The HMM is subsequently parameterized with emission distributions and transition probabilities based on probe intensity distributions from the training data and a user-defined probability of recombination parameter *P(Recomb)*, respectively.

included in the training set. We accomplished this by incorporating a hidden Markov model (HMM) into our method in order to define recombinant paths when calculating probabilities for candidate viruses (Figure 4.1). Figure 4.2 shows the details of constructing the HMM. The Viterbi algorithm was used to determine the optimal path from which recombination breakpoints could be inferred. As with VIPR, our HMM allows us to take advantage of training data consisting of hybridizations of known viruses to a microarray to make predictions for unknown infections. The incorporation of an HMM into VIPR now provides a probabilistic framework for assessing the presence of recombination between candidate parental viruses. To validate our approach, we applied

our HMM to a set of 15 recombinant viruses consisting of members of the *Alphavirus* and *Flavivirus* genera, each of which was hybridized in duplicate to a custom microarray. A set of microarrays to which nonrecombinant alphaviruses and flaviviruses were hybridized constituted the training data for the HMM. While our test focused on the validation of a set of recombinant alphaviruses and flaviviruses, the strategy should be generalizable to detecting recombination among members of a given viral family.

RESULTS

RNA was purified from cell cultures that were infected with each of the viruses shown in Table 4.1 and Table 4.2. Purified RNA was subsequently randomly amplified and hybridized to a custom diagnostic microarray. 65 hybridizations (60 representing nonrecombinant alphavirus and flavivirus parental viruses + 5 representing uninfected Vero cells) were performed in order to obtain a training set for the HMM. For validation of our algorithm, 49 hybridizations (30 representing alphavirus and flavivirus recombinants + 15 representing alphavirus and flavivirus nonrecombinants + 4 representing uninfected Vero cells) were performed.

In order to build the HMM, we first needed to establish a framework to define possible recombinant and nonrecombinant paths based on positional information inherent to each probe. The microarray probes were ordered by their position from 5' to 3' in the global alignment of candidate virus genomes (Figure 4.2). This was accomplished by mapping the set of oligonucleotide probes via local alignment (megablast) to each candidate virus genome, identifying probes for which the theoretical free energy associated with its probe:genome local alignment was $\leq$ -30 kcal/mol (indicative of

Table 4.1. Alphavirus and flavivirus parental viruses grown in culture and hybridized to the diagnostic microarray.

Genus	Species	Strain	Genbank	VIPR HMM strain designation
Alphavirus	CHIKV	LR	116047549	
Alphavirus	EEEV	BeAr436087	119633049	1
Alphavirus	EEEV	FL93-939	119633046	2
Alphavirus	SINV	AR339	9790313	
Alphavirus	VEEV	68U201	1144527	1
Alphavirus	VEEV	TC-83	323714	2
Alphavirus	VEEV	TRD	323714	2†
Alphavirus	VEEV	ZPC738	4689187	3
Alphavirus	WEEV	CO92-1356	254595918*	
Alphavirus	WEEV	McMillan	254595918	
Flavivirus	DENV-4	1228	12659201*	
Flavivirus	JEV	SA14-14-2	12964700	
Flavivirus	SLEV	CorAn9124	344221822*	
Flavivirus	WNV	NY99	158516887	
Flavivirus	YFV	17D	9627244	

*Genbank ID represents a closely related strain since the sequence of the exact strain was not available

†Since VEEV TRD and VEEV TC-83 genomes differ by only 11 nucleotides, they were considered to be the same strain (VEEV strain 2)

binding using previously explained criteria [92]), and converting the midpoint of the probe:genome local alignment for each of those probes to its corresponding position in the global alignment [93] of candidate virus genomes. Probes that mapped to multiple genomes at similar positions but were offset relative to each other by 30 nucleotides or fewer were consolidated to a single position in the global alignment. Probes were then sorted by their positions in the global alignment of candidate virus genomes.

Once the probes were ordered, they were assigned *On* and *Off* states for each genome. These assignments were based on the same theoretical free energy of binding calculated in the mapping step. *On* and *Off* states emit normalized and $\log_e$ transformed intensities according to normal distributions estimated from training data as previously

Table 4.2. Recombinant alphaviruses and flaviviruses hybridized to the diagnostic microarray for validation of the HMM.

Virus	Type of recombinant	Parents	Coordinates in parental genomes
R01	Double	EEEV BeAr436087 CHIKV LR	1-7499;11291-11638 7504-11313
R02	Double	SINV AR339 VEEV TC-83	1-7601;11394-11703 7536-11382
R03	Double	SINV AR339 CHIKV LR	1-7601;11383-11703 7502-11313
R04	Double	SINV AR339 WEEV CO92-1356	1-7602;11385-11703 7466-11210†
R05	Double	SINV AR339 EEEV BeAr436087	1-7601;11312-11703 7498-11291
R06	Double	SINV AR339 VEEV TRD	1-7601;11394-11703 7536-11382†
R07	Double	VEEV TC-83 CHIKV LR	1-7533;11328-11446 7500-11313
R08	Double	YFV 17D DENV-4 1228	1-481;2453-10862 441-2423†
R09	Double	YFV 17D JEV SA14-14-2	1-481;2453-10862 477-2477
R10	Double	YFV 17D SLEV CorAn9124	1-481;2453-10862 456-2465†
R11	Double	YFV 17D WNV NY99	1-481;2453-10862 466-2469
R12	Double*	SINV AR339 VEEV TC-83 VEEV 68U201	1-7601;11394-11703 7536-8286 8298-11398
R13	Double*	SINV AR339 EEEV BeAr436087 EEEV FL93-939	1-7601;11312-11703 7498-7640(7641-7675)‡ (7673-7707)7708-11323
R14	Double*	SINV AR339 VEEV TC-83 VEEV ZPC738	1-7601;11394-11703 7536-8353(8354-8406) (8331-8383)8384-11359
R15	Triple*	SINV AR339 EEEV BeAr436087 EEEV FL93-939 WEEV McMillan	1-7601;11385-11703 7498-7640(7641-7675) (7673-7707)7708-7902 7802-11210

Coordinates corresponding to the parental genomes listed in Table 4.1 are given. For the recombinant alphaviruses, a short cloning sequence (between three and ten nucleotides) is present at the 3'-most recombination breakpoint.

*additional intra-species breakpoints present

†coordinates derived from closely related strain listed in Table 4.1

‡parentheses represent regions of overlap between two parents sharing identical sequence

described [92]. Thus, all emission probabilities *e(state, intensity)* were derived from distributions estimated in a manner identical to the estimation of probe-specific *On* and *Off* distributions in VIPR except in the case where there were fewer than 8 intensities available in the training set for a given probe. In that case, the mean of the distribution was calculated from the available intensities, but the standard deviation was derived from the average standard deviation over all probes with a similar *On* or *Off* prediction. In addition to the candidate virus genomes, a null genome was included which represented a none-of-the-above genome prediction and was assigned an *Off* state for each probe.

Finally, the states in the HMM were connected via transitions *t(state, state)* as depicted in Figure 4.2. As with HMMs that have been developed to detect recombination in sequence, probabilities representing recombination transitions could not be estimated directly from the training data as could the other HMM parameters [94]. Thus, a user-specified probability of recombination parameter *P(Recomb)* was introduced to compute transition probabilities. Transitions connecting states within the same genome i.e. $t(state_{Virus\_A}, state_{Virus\_A})$ represented non-recombination events and had the associated probability *1-P(Recomb)*. Transitions between genomes i.e. $t(state_{Virus\_A}, state_{Virus\_B})$ represented recombination events and had the associated probability *P(Recomb)/(n-1)* where *n* is the number of candidate virus genomes (including the 'null' genome). In some cases, multiple probes mapped to the same position in the global alignment of candidate virus genomes. Transitions between states whose probes mapped to the same position were only allowed if those states correspond to the same genome and were assigned a probability of 1.0, such that recombination events were not permitted between such states. Because the next state in the model is dependent only on the current state, and

because states in the model emit from continuous intensity distributions, the model is a first-order continuous HMM.

Two models were built and were used to analyze the alphaviruses and the flaviviruses, respectively. In order to experimentally define a suitable *P(Recomb)* for computing transition probabilities, we evaluated the performance of VIPR HMM on a subset of the parental viruses, varying *P(Recomb)* over a range of values. We selected the maximum value of *P(Recomb)* that resulted in zero false positive recombination breakpoints when Viterbi was applied to the parental alphaviruses. This value, $P(Recomb) = 10^{-25}$, was subsequently used when applying VIPR HMM to the parental flaviviruses as well as to the alphavirus and flavivirus recombinants. The Viterbi results were compared to expected results based on the known sequences of the recombinant constructs. When applied to the 5 flavivirus nonrecombinants, VIPR HMM classified each as the correct species. Additionally, the four uninfected Vero samples were accurately classified as null. VIPR HMM detected no recombination breakpoints for these samples except for one false positive breakpoint at the 3' end of the dengue virus 4 genome, which bypassed the final 254 nucleotides of the genome in favor of null states. VIPR HMM results for the nonrecombinant alphaviruses and flaviviruses are shown in Figure 4.S1.

A total of 30 hybridizations of recombinant viruses was analyzed by VIPR HMM. Figures 4.S2-4.S5 shows results for all recombinant alphaviruses and flaviviruses analyzed by VIPR HMM. Of the 30 hybridizations, 28 represented double recombinants between two parent viruses of distinct species and two represented triple recombinants composed of three distinct parental species. Thus, the total number of expected inter-

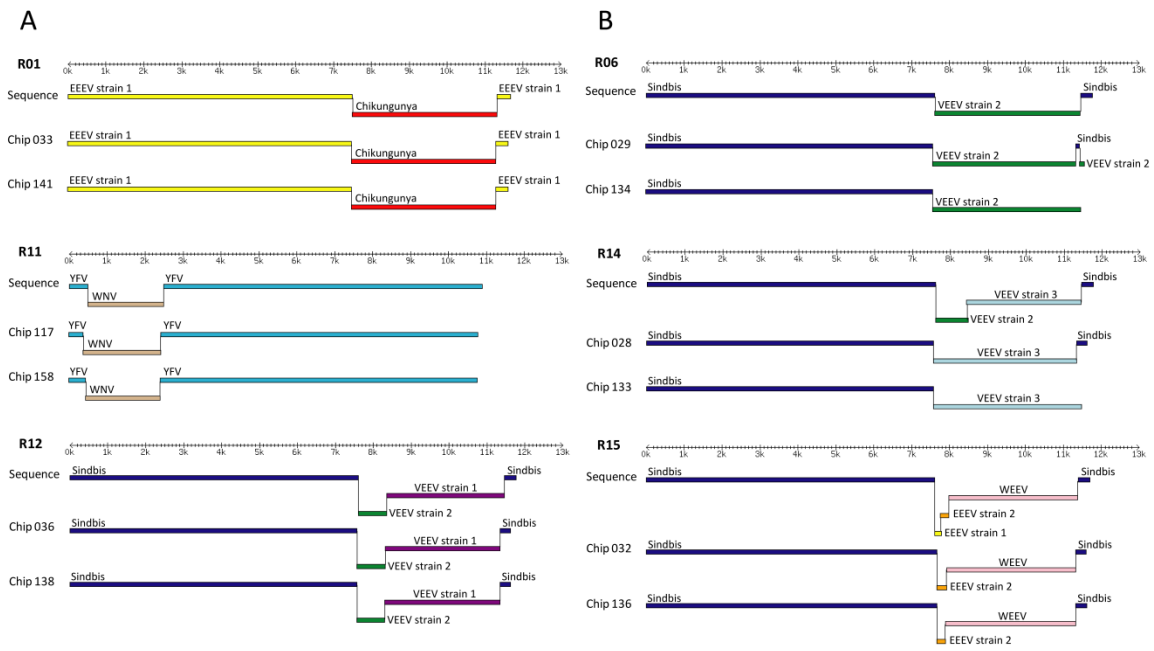

Fig. 4.3. VIPR HMM results for a subset of recombinants tested. A**)** VIPR HMM results for three recombinants that gave expected results. For each recombinant, the expected output based on sequence is shown, followed by the VIPR HMM output for the two hybridizations performed. B**)** VIPR HMM results for three recombinants that gave unexpected results. R06 is a double recombinant for which an additional false positive recombination breakpoint was identified at the 3' end in one hybridization, and for which a 3' inter-species recombination breakpoint was not identified in the other hybridization. R14 is a double recombinant for which a 3' inter-species recombination breakpoint was identified in one of the hybridizations, but not the other. Additionally, an intra-species recombination breakpoint was not identified in either hybridization. R15 is a triple recombinant for which all three inter-species recombination breakpoints were identified in both hybridizations, but for which an intra-species recombination breakpoint was not identified in either.

species recombination breakpoints was (28 x 2) + (2 x 3) = 62. VIPR HMM correctly identified breakpoints and the identity of the parental species for 59 of the 62 total breakpoints. VIPR HMM results for a subset of the recombinant viruses that were identified unambiguously are shown in Figure 4.3A. In the remaining three instances, VIPR HMM yielded false negatives. Of all the recombinant and nonrecombinant samples analyzed by VIPR HMM, only two false positive breakpoints were predicted (one in a nonrecombinant virus and one in a double recombinant virus).

In some cases, the recombinant viruses we used included intra-species recombination breakpoints. Of the 8 intra-species breakpoints, 2 were identified by VIPR

HMM. For those 2 breakpoints, the correct viruses 5' and 3' of the breakpoint were identified (both species and strain). VIPR HMM results for a subset of the recombinant viruses that gave unexpected results are shown in Figure 4.3B.

VIPR HMM was used to estimate the nucleotide positions of each breakpoint in each parental genome. The nucleotide positions associated with recombination breakpoints were estimated based on the position in the alignment of the probes associated with the recombinant transition in the Viterbi path. For each such probe, its position in the alignment was correlated with a position in the Viterbi-specified parental virus genome to estimate the nucleotide position of the recombination breakpoint in that genome. The differences between the nucleotide positions estimated by VIPR and the actual sequence positions ranged from 0 to 90 nucleotides.

DISCUSSION

The ability of DNA microarrays to simultaneously assess the presence of multiple loci in microbial genomes is highly advantageous for detecting recombination between virus species in a diagnostic setting. Despite this, none of the existing methods for analyzing diagnostic microarrays is designed to accommodate the detection of recombinant viruses. In previous work, we developed VIPR, a method for objectively interpreting diagnostic microarrays. One of the advantages of VIPR relative to other methods is that it relies on a training set of empirical hybridizations of virally infected and uninfected samples to leverage diagnostic predictions. We anticipated that relying on a training set of hybridizations from known viral infections would also help us predict recombination between virus species. In this study, we developed a hidden Markov Model (HMM)

parameterized with VIPR probability distributions to detect recombination in unknown infections.

VIPR HMM performed with high accuracy when identifying recombination breakpoints between viral species (59/62 such breakpoints were identified and the correct virus species 5' and 3' to the breakpoint were identified in each case). Of the 8 intra-species breakpoints in our data set, two were identified by VIPR HMM. Given that a much higher percentage of inter-species breakpoints were detected than were intra-species breakpoints (95% versus 25%), these results demonstrate that VIPR HMM is more effective at detecting recombination between species than between strains belonging to the same species. The ability of VIPR HMM to distinguish between strains of the same species involved in recombination is likely influenced by the degree of sequence divergence between the two strains. VIPR HMM correctly identified the intra-species breakpoint in both hybridizations of R12 (Figure 4.3). The two strains comprising the intra-species breakpoint for R12 are 23% divergent on the nucleotide level. However, VIPR HMM was not able to identify the intra-species breakpoint in either hybridization of R14 (Figure 4.3), whose recombinant regions 5' and 3' to the intra-species breakpoint were similar in size to those of R12, but whose strains comprising the intra-species breakpoint are only 4% divergent on the nucleotide level. The ability of VIPR HMM to distinguish between strains of the same species may also be influenced by the size of the recombinant segment. The four other intra-species breakpoints that VIPR HMM failed to detect had greater dissimilarity between flanking strains (25%) but were proximal to other breakpoints (within 200 nt). Given the cost in probability associated with following

a recombinant transition in the HMM, our results suggest that Viterbi may opt to bypass small recombinant regions.

Since microarray probes are mapped to their position in an alignment of candidate genomes, VIPR HMM can use the probes located at the boundary of a predicted recombination event to estimate nucleotide positions of recombination breakpoints. Although it was not expected that using a microarray tiling scheme wherein probes were non-overlapping and spaced 63 nucleotides apart would give the precise nucleotide positions of recombination breakpoints, we compared the estimates given by VIPR HMM to the nucleotide positions known from sequence. For the 61 correctly identified breakpoints (59 inter-species, 2 intra-species), the differences between microarray estimates and actual positions ranged from 0 to 90 nucleotides. Therefore, the maximum distance observed falls within the span of about a two probe tiling (i.e. 90 < 60mer + 3 nt spacing + 60mer). We expect that using higher density tiling strategies would result in higher resolution mapping of the breakpoints.

Only two false positive recombination breakpoints were predicted by VIPR HMM, both near the 3' ends of their respective genomes. One bypassed the final 254 nucleotides of dengue virus 4 in favor of null states. The other bypassed the final 191 nucleotides of Sindbis virus in favor of VEEV states. From analysis of the training data, it was observed that the mean of the *On* distributions approach the mean of the *Off* distributions for probes near the 3' end of each genome, due to lower intensities for *On* probes in the training set for that region. This trend was observed in the training data universally for all genomes. The tendency for *On* probes to give lower intensities when approaching the 3' end may be attributable to the fact that random PCR amplification,

which was used in the preparation of each sample for hybridization, is less efficient at the ends of a linear genome. This could also explain why VIPR failed to detect three inter-species recombination breakpoints, all of which are localized near the 3' end of a genome. A similar pattern of lower intensities was also observed for *On* probes approaching the 5' end, although there appeared to be more probes in those regions that behaved as expected based on ΔG compared to the 3' end. Despite the observed decrease in hybridization intensity proximal to the 3' and 5' termini, VIPR HMM was still able to make accurate predictions in those regions in most cases.

Although we did not specifically validate VIPR HMM for reassortant viruses, we anticipate that viral reassortants would be readily detected. Reassortment can occur during co-infection when virus progeny inherit genome segments from two or more parental viruses with multi-segmented genomes. The resulting chimeric genotypes associated with reassortment are similar to those generated though recombination except that the exchange of genetic material occurs at discrete, predictable points in the genome i.e. at the boundary between genome segments.

VIPR HMM relies on a multiple alignment of candidate viral genomes to order microarray probes. One limitation of this approach is that only recombination between members of the same family will be considered as candidates since it is not generally feasible to globally align members of different families. In addition, because paths through the HMM follow a specific 5' to 3' ordering, only recombination at homologous sites is detectible by VIPR HMM as currently implemented. In future versions of VIPR, recombinants composed of viruses from different families could be detected by running multiple iterations of Viterbi, one for each HMM representing a different virus family.

For a hypothetical recombinant between members of two different virus families, we anticipate that the HMM for each family would predict the presence of only a portion of the viral genome from its family (with the rest of the prediction being the null genome).

One challenge in building an HMM for detecting recombination is finding an appropriate value for *P(Recomb)*, a user-inputted probability of recombination parameter used to calculate different transition probabilities in the model. Our choice of *P(Recomb)* was based on minimizing false positive recombinations in nonrecombinant samples. However, in some cases, it may be advantageous to increase *P(Recomb)* in order to increase detection sensitivity.

CONCLUSIONS

We developed a hidden Markov model (HMM) to identify recombination in viruses that have been hybridized to a microbial detection microarray. This model builds on previous work in which empirical hybridizations of cultured viruses were used as training to classify unknown infections (VIPR). Applying the HMM in conjunction with VIPR enabled the detection of inter-species recombination breakpoints with high accuracy in two different families of viruses. This is the first report of a method for analyzing diagnostic microarrays that includes recombination as a possible diagnostic outcome. Our method is theoretically applicable to detecting homologous recombination or reassortment between members of any family of viruses for which a set of nonrecombinant parental viruses is available for training and for which genome sequences are available. The inherently parallel nature of diagnostic microarrays coupled with powerful methods for analysis enhance our ability to rapidly and accurately identify

novel recombinant viruses responsible for disease outbreaks, either due to emergence by natural means or by engineered recombinant viruses.

MATERIALS AND METHODS

Design of the diagnostic microarray

60mer oligonucleotide probes were designed from sequences representing three virus families (*Bunyaviridae, Flaviridae and Togaviridae*) using a tiling strategy. 145 RefSeq genomes and genome segments from the aforementioned virus families were obtained from Genbank. To the RefSeq set we added from Genbank as many complete genome sequences as were available of the parental viruses of the 11 recombinant alphaviruses. Partial genome sequences for the parental alphaviruses were added if complete genomes were not available. Additionally, complete genome sequences of alphaviruses that did not represent parents of the recombinant viruses were added until there were in the set at least three complete genomes of each of EEEV, VEEV, WEEV, Chikungunya and Sindbis viruses. The final set of Genbank records totaled 193, of which 175 were complete or nearly complete genomes or genome segments. Probes were selected as 60 nucleotide windows tiled over all 193 sequences with a spacing of three nucleotides between the 3' end of one probe and the 5' end of the following probe. The reverse complement of each 60mer was also included in the microarray. The resulting set of probes including reverse complements totaled 43414 and the Agilent® 4 x 44 K platform was used (GEO accession GSE34490).

Hybridization of alphavirus and flavivirus parental and recombinant viruses to the diagnostic microarray

21 alphaviruses (11 recombinants + 10 parental viruses) and 9 flaviviruses (4 recombinants + 5 parental viruses) which have been previously described [89, 95-99] were obtained from the World Reference Center for Emerging Viruses and Arboviruses and were grown in Vero cells. RNA was extracted using standard Trizol® protocols and was reverse transcribed and randomly amplified as previously described [29]. For each recombinant, two independent amplifications were performed, while five independent amplifications were performed for each parental virus. The resulting amplified material was then coupled to a fluorescent dye and hybridized to the tiling microarray. Raw data measurements were collected using GenePix Pro® software. In total, 114 hybridizations were performed (30 recombinant + 75 parental + 9 uninfected Vero cells). All raw microarray data are available in NCBI GEO (accession GSE34490). The training set for our HMM consisted of 60 parental hybridizations + 5 Vero negative control hybridizations, while the test set for validating the algorithm consisted of the 30 recombinant hybridizations + 15 parental hybridizations + 4 Vero negative control hybridizations.

Viterbi algorithm for finding the optimal path

By multiplying emission probabilities *e(state, intensity)* and transition probabilities *t(state, state)* across a series of states, it is possible to obtain a probability for an entire path through an HMM. For our HMM, the set of emission and transition parameters is

abbreviated as θ. The probability of a particular path (π) and a given hybridization ($\boldsymbol{x}$) of length L can be expressed as a joint probability:

$$P(\boldsymbol{x}, \pi \mid HMM, \theta) = t(0, \pi_1) \prod_{i=1}^{L} t(\pi_i, \pi_{i+1}) e(\pi_i, x_i)$$

The Viterbi algorithm falls into a class of algorithms called dynamic programming algorithms that are commonly used in conjunction with HMMs. Using the Viterbi algorithm allows us to identify the most probable series of states (π') through our HMM where

$$\pi' = \arg\max_{\pi} P(\boldsymbol{x}, \pi \mid HMM, \theta)$$

Points of recombination can be inferred from places in the path where a transition between states of different genomes has occurred. As with other dynamic programming algorithms, the Viterbi algorithm consists of an initialization step, an iteration step and a termination step. Once the dynamic programming matrix (V) is populated, the optimal path is traced back through a shadow matrix (τ) of stored pointers. Except for the begin and end states s_{begin} and s_{end} and states in a given path (π_i), all other states ($s_{g,i}$) are indexed by genome (g) and probe-column (i). The V matrix and τ matrix are similarly indexed. Calculations are performed in log space although they are shown here in probability space. The Viterbi algorithm adapted from [47] is as follows:

Initialization (*g = 1 to n*)

$$V_{g,1} = t(s_{begin}, s_{g,1}) e(s_{g,1}, x_1)$$

Iteration (*i = 2 to L; g = 1 to n*)

$$V_{g,i} = e(s_{g,i}, x_i) \max_{j=1}^{n} \left[V_{j,i-1} t(s_{j,i-1}, s_{g,i}) \right]$$

$$\tau_{g,i} = \arg\max_{j=1}^{n} \left[V_{j,i-1} t(s_{j,i-1}, s_{g,i}) \right]$$

Termination

$$P(\boldsymbol{x}, \pi' \mid HMM, \theta) = \max_{j=1}^{n} \left[V_{j,L} t(s_{j,L}, s_{end}) \right]$$

$$\pi'_L = \arg\max_{j=1}^{n} \left[V_{j,L} t(s_{j,L}, s_{end}) \right]$$

Traceback (*i = L to 2*)

$$\pi'_{i-1} = \tau(\pi'_i, i)$$

Traceback reveals the optimal path through the HMM. If the path includes states representing only one genome, the optimal path is a nonrecombinant path. If the optimal path includes transitions between states of different genomes, the path is recombinant, and the global alignment positions corresponding to the probes associated with the states involved in each transition are referenced. These global alignment positions are then back-converted to genomic positions in the predicted virus parents in order to define the recombinant breakpoints between virus genomes on the nucleotide level.

Availability

VIPR HMM is freely available for academic use and can be downloaded from http://ibridgenetwork.org/wustl/vipr

ACKNOWLEDGMENTS

We thank Michael Brent for useful discussions.

Funding: This work was supported by NIH grant U01 AI070374 and by a grant from the National Institute of Allergy and Infectious Disease (NIAID) through the Western Regional Center of Excellence for Biodefense and Emerging Infectious Disease Research, National Institutes of Health (NIH) grant U54 AIO57156.

Supplemental Figures

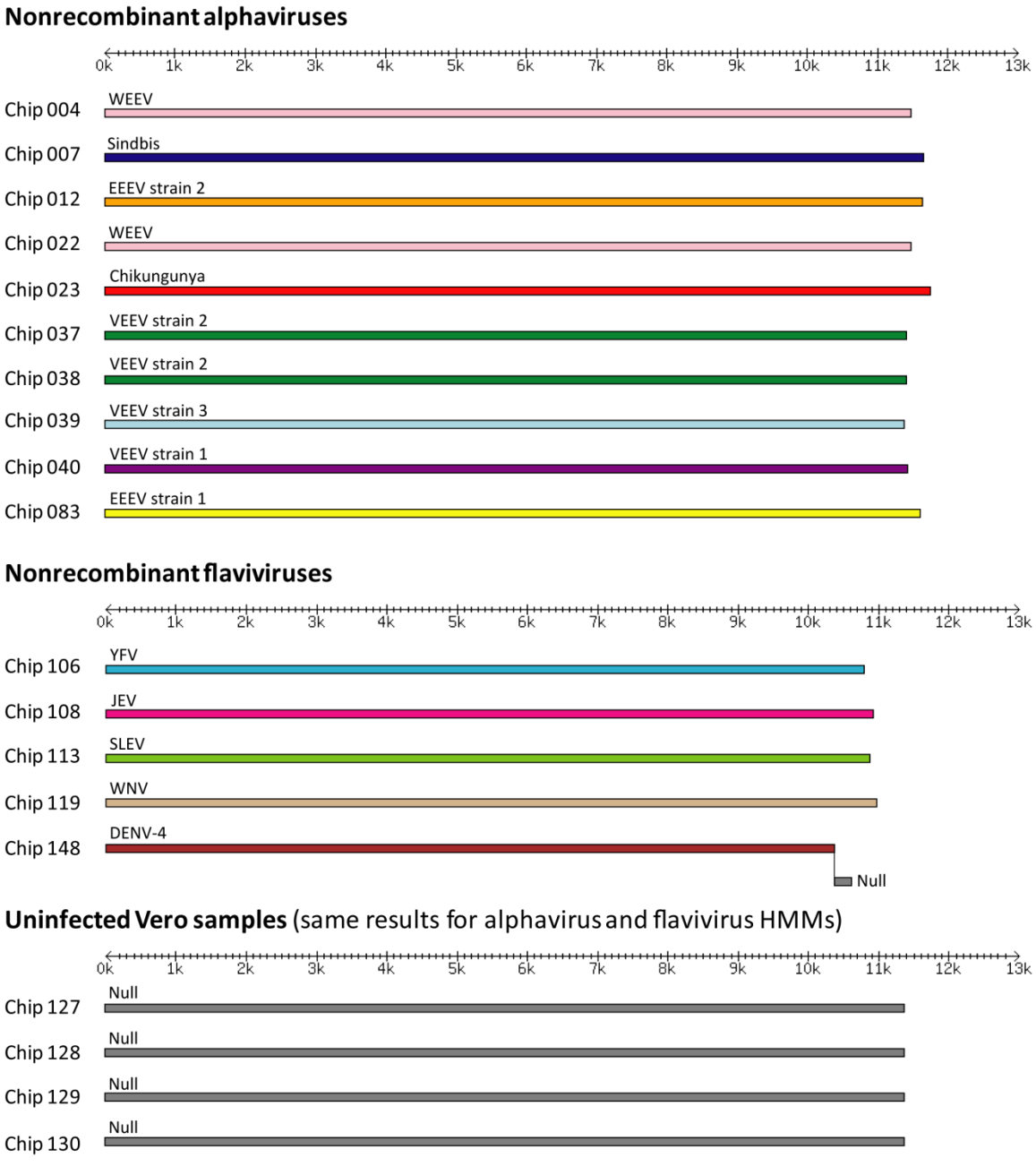

Fig. 4.S1. VIPR HMM output for nonrecombinants (alphaviruses and flaviviruses) and uninfected Vero samples.

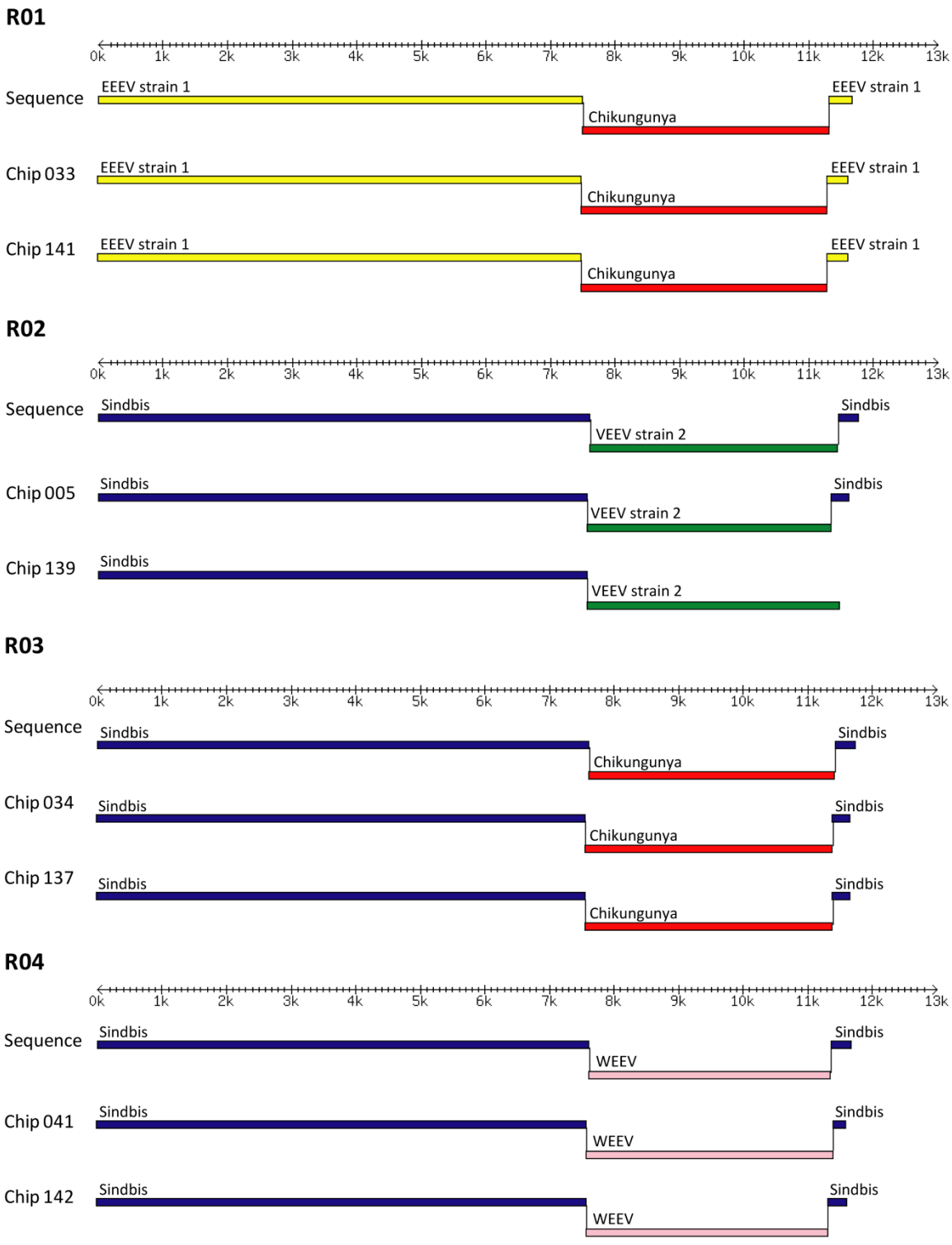

Fig. 4.S2. VIPR HMM output for recombinant samples R01-R04

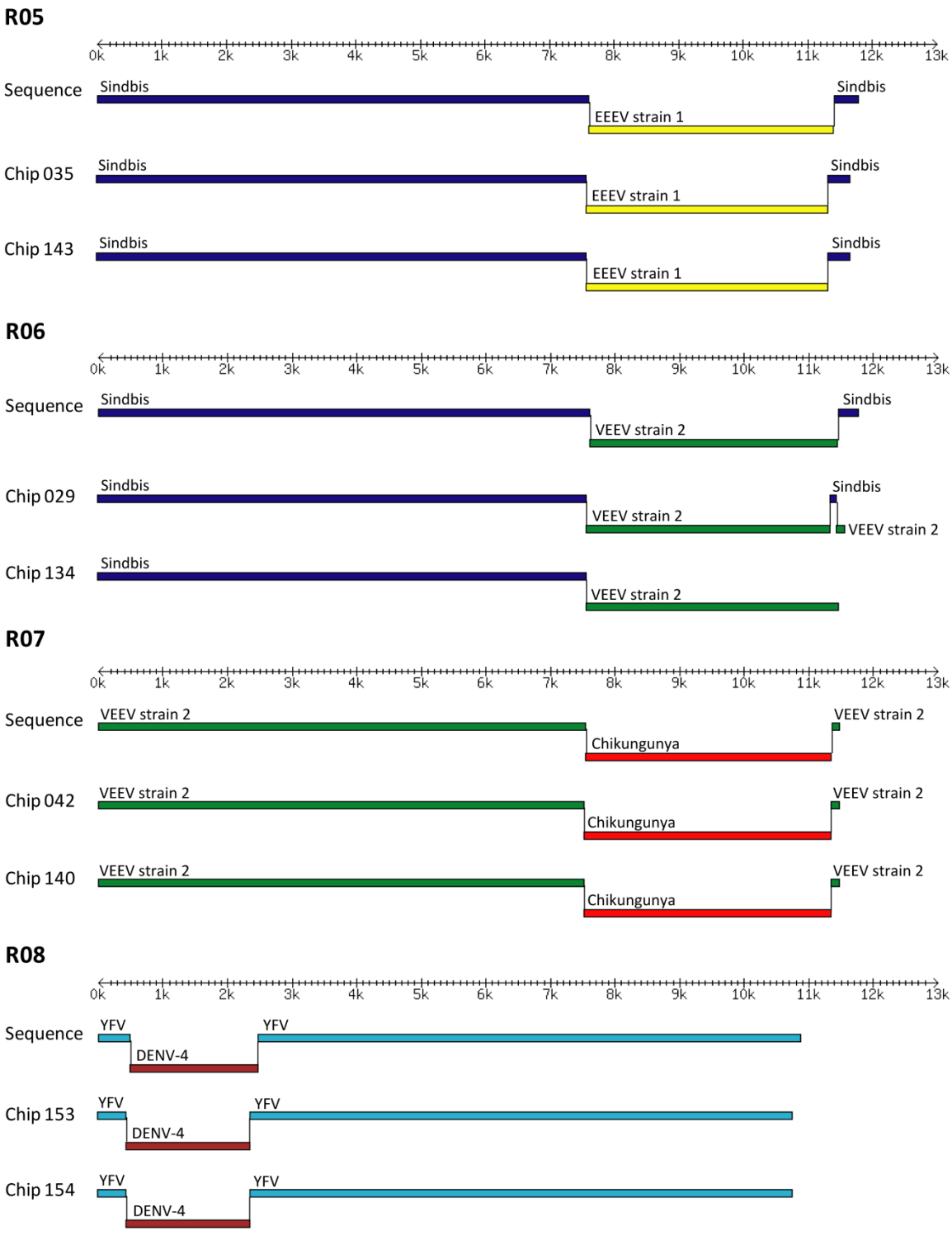

Fig. 4.S3. VIPR HMM output for recombinant samples R05-R08

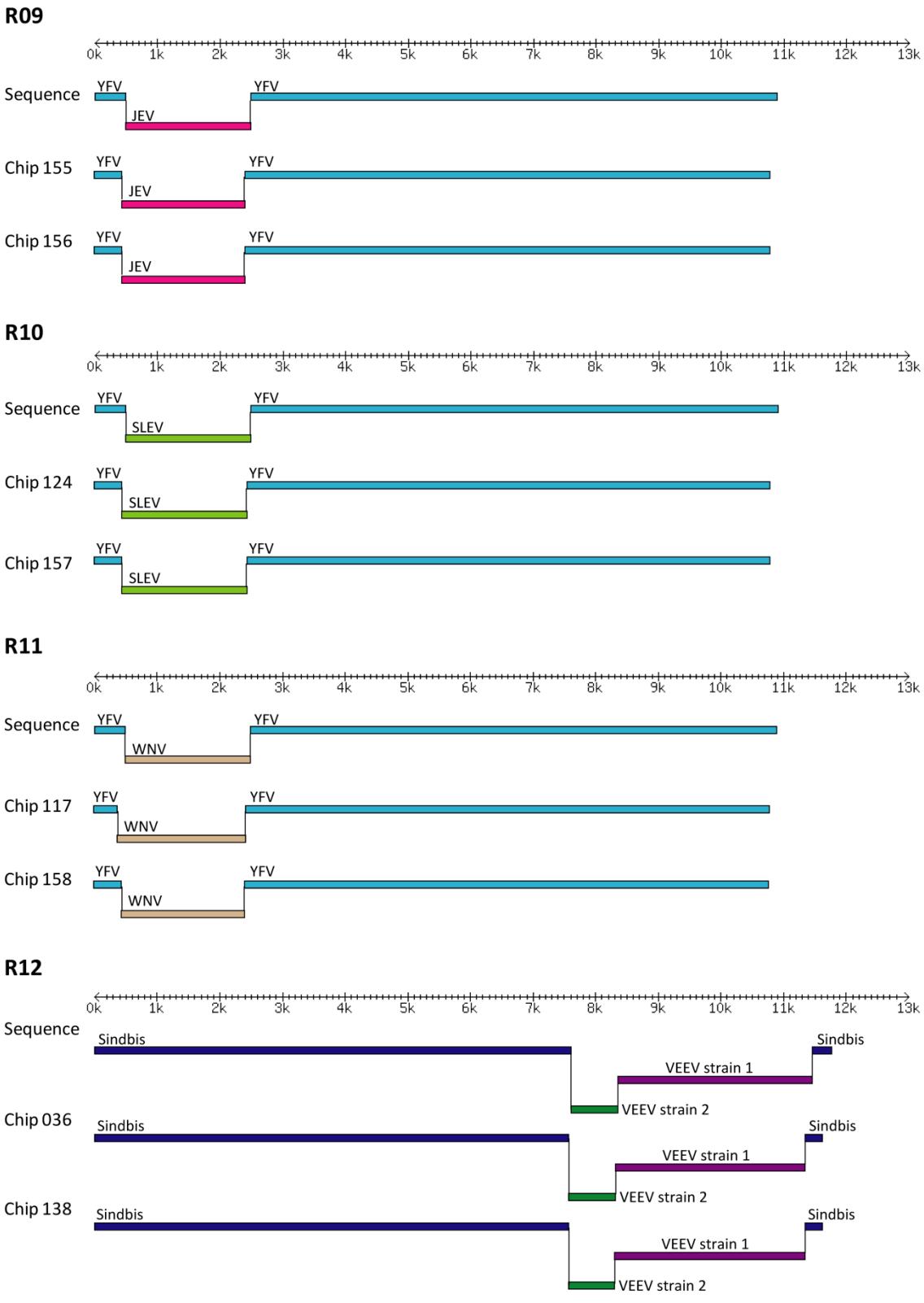

Fig. 4.S4. VIPR HMM output for recombinant samples R09-R12

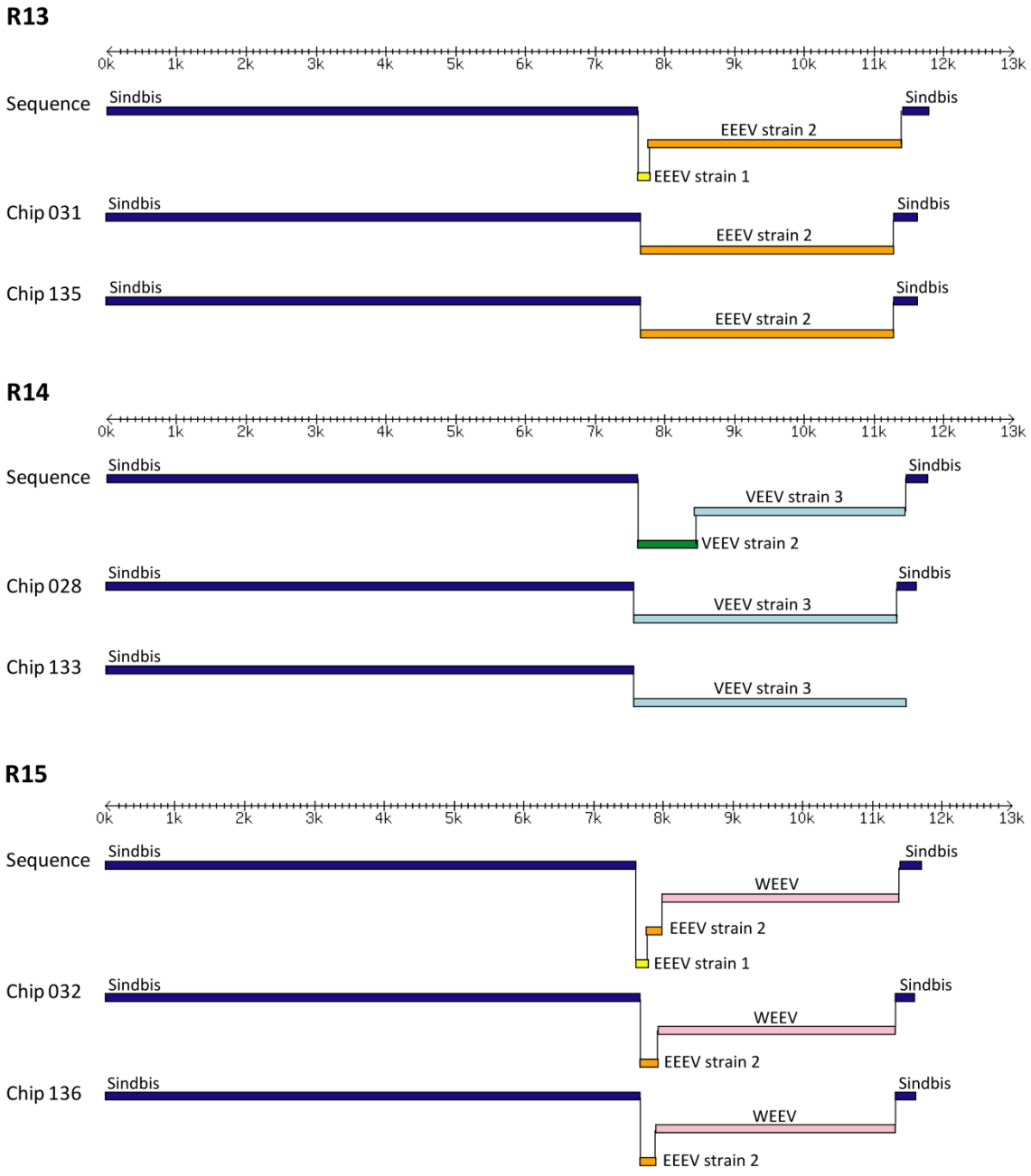

Fig. 4.S5. VIPR HMM output for recombinant samples R13-R15

CHAPTER 5:

Conclusions

Given the limitations of traditional assays for pathogen detection, there is a great need in diagnostic microbiology for high-throughput assays capable of detecting many agents in parallel and in an unbiased fashion. Metagenomic sequencing and microarray technology are powerful tools which can help circumvent the challenges associated with the detection of uncultured microbes. Such data-intensive approaches require robust computational tools to process raw measurements, make predictions in the presence of technical and other sources of noise, capitalize upon experimental design, and facilitate interpretation.

Metagenomic sequencing provides an unprecedented opportunity to explore the makeup of microbial communities in environmental as well as human-derived specimens. Moreover, metagenomic surveys of clinical specimens can provide a basis for further investigations regarding the role of microbial communities in disease. As demonstrated by this work, such surveys can lead to the identification of both known and novel microbial species. Prior to this work, nothing was known about the diarrhea virome since only stool specimens from asymptomatic patients had been sequenced previously. Using a unique sequencing strategy designed specifically for pediatric diarrhea specimens coupled with a robust sequence analysis pipeline, viruses were detected in all but one specimen. Additionally, many sequences were identified which had only limited similarity to known viruses. A particularly exciting result that came from this study was evidence for the presence of multiple novel viruses. Sequences were detected from at least nine putatively novel viruses in these specimens. For two of these putative viruses, the sequence divergence from known viruses was especially pronounced, suggesting the possibility that these sequences represent novel virus species. Following analysis and

taxonomic binning of sequences in one specimen, seven unique sequences were revealed which had limited similarity to known astroviruses. Astroviruses are implicated in up to 10% of cases of sporadic diarrhea [63]. Phylogenetic analysis confirmed marked divergence of this virus relative to other known human astroviruses. Following up on this result, Finkbeiner et al. later sequenced the entire genome of this virus which is now known as astrovirus MLB1 [100]. This demonstrates the efficacy of metagenomic approaches which can be used as a springboard for further characterization of novel viruses such as astrovirus MLB1, including investigation of the possible role of such viruses in diarrheal disease. Future applications of similar metagenomic methodologies to other diseases or specimen types for the purpose of microbial detection and discovery also appear bright.

While the metagenomic sequence pipeline I developed was applied exclusively to Sanger sequencing reads, it later served as a prototype for the development a pipeline designed to process 454 pyrosequencing reads (Zhao et al., unpublished). While platforms featuring longer reads (i.e. Sanger, 454) are preferable for the taxonomic classification of sequences with only limited similarity to known microbes, the development of assembly tools which can join short reads in a metagenomic context for downstream taxonomic assignment may help make these platforms more amenable to discovery of divergent microbes in the future.

One challenge associated with metagenomic sequencing analysis is the preponderance of sequence reads which cannot be reliably assigned to any taxonomic category. For the diarrhea study, 26% of unique sequences were unable to be classified. Since all of these sequences were determined to be high-quality, it is not likely that

altering the experimental procedure used for processing and sequencing of specimens would have had any effect on the ability to reliably classify such reads. Rather, future studies devoted to the development of computational tools capable of detecting more distant evolutionary relationships could help to identify the species of origin for currently unclassifiable reads. The identification of the origin of such reads will also be facilitated as genetic databases accumulate more sequences to be used in similarity searches.

Microarrays have also garnered attention in recent years as a rapid way to carry out detection of many microbes in parallel. Considering the data-rich nature of a diagnostic microarray experiment wherein thousands of probe intensities can factor into a prediction and where there are many potential virus candidates, each of which is likely to share some degree of sequence similarity with others of the candidates, it is imperative that an objective software tool be available for the interpretation of such data.

While several tools had been developed previously for the interpretation of diagnostic microarrays, none of them capitalized on training data as part of a machine learning approach to virus prediction. I developed a computational tool, VIPR, which relies on hybridizations of known viruses to diagnostic microarrays as a training set in order to gauge probe-specific behaviors and improve future predictions. VIPR accomplishes this using a probabilistic approach wherein probe probabilities are multiplied together under the assumption of independence.

VIPR performed with high accuracy (94%) when applied to a set of hemorrhagic fever viruses and their relatives. VIPR outperformed previously published methods for this data set. While the choice of prior can potentially be problematic for Bayesian approaches, VIPR predictions were found to be robust to changes in this parameter. VIPR

is theoretically applicable to any diagnostic scenario where positive control specimens are available for use as training data.

Traditional diagnostics fail to detect recombination since they are focused on the detection of a single locus in candidate microbes. In order to accommodate the detection of recombinant viruses, I developed a hidden Markov model which expands upon VIPR's original probabilistic implementation. VIPR HMM was parameterized with emissions derived from probe intensity distributions and transition probabilities derived from a user-selected probability of recombination parameter. Applying VIPR HMM to a recombinant set of viral encephalitis vaccines resulted in accurate detection of 95% of inter-species breakpoints. Additionally, VIPR HMM was able to identify intra-species breakpoints in some cases. The ability of VIPR HMM to detect intra-species breakpoints may be dependent upon the degree of sequence divergence between recombining strains. Using the probes located at the boundary of a predicted recombination breakpoint, VIPR HMM can estimate nucleotide positions of breakpoints. All predicted breakpoints fell within 90 nucleotides of their actual positions (i.e. within a two-probe tiling since probes were tiled 63 nucleotides apart).

While VIPR HMM was designed to detect homologous recombination between members of the same family, future studies could be devoted to adding functionality to detect recombination between members of different families or recombination at non-homologous sites. In fact, the incorporation of a "null" genome makes it possible to identify portions of a virus that are present while allowing for a null prediction for portions of the genome which are absent. Running multiple iterations of Viterbi using models for different families could allow for identification of these portions in other

families. Next generation sequencing could also prove a valuable tool for detection of recombination. Reads whose full sequence could be mapped to one particular genome would serve to identify parental species, while the remaining reads could be screened for potential breakpoint-scanning sequences for fine resolution of breakpoints.

High-throughput genomic approaches such as metagenomic sequencing and microarray technology offer highly parallel and unbiased detection of viruses in clinical specimens. Due to the data-intensive nature of these technologies, robust bioinformatics tools are required for objective analysis. As the cost associated with high-throughput approaches decreases, their use in clinical and public health laboratory settings will become more salient, enhancing our ability to rapidly and accurately detect pathogens and to respond to infectious disease.

References

1. Knipe DM (ed.): **Fields Virology**, 4th edn. Philadelphia: Lippincott Williams &Wilkins; 2001.
2. Storch GA: **Essentials of Diagnostic Virology**. New York: Churchill Livingstone; 2000.
3. Bon F, Fascia P, Dauvergne M, Tenenbaum D, Planson H, Petion AM, Pothier P, Kohli E: **Prevalence of group A rotavirus, human calicivirus, astrovirus, and adenovirus type 40 and 41 infections among children with acute gastroenteritis in Dijon, France**. *J Clin Microbiol* 1999, **37**(9):3055-3058.
4. Chikhi-Brachet R, Bon F, Toubiana L, Pothier P, Nicolas JC, Flahault A, Kohli E: **Virus diversity in a winter epidemic of acute diarrhea in France**. *J Clin Microbiol* 2002, **40**(11):4266-4272.
5. Denno DM, Klein EJ, Young VB, Fox JG, Wang D, Tarr PI: **Explaining unexplained diarrhea and associating risks and infections**. *Animal health research reviews / Conference of Research Workers in Animal Diseases* 2007, **8**(1):69-80.
6. Kapikan A: **Viral Gastroenteritis**. *The Journal of the American Medical Association* 1993, **269**(5):627-630.
7. Kirkwood CD, Clark R, Bogdanovic-Sakran N, Bishop RF: **A 5-year study of the prevalence and genetic diversity of human caliciviruses associated with sporadic cases of acute gastroenteritis in young children admitted to hospital in Melbourne, Australia (1998-2002)**. *J Med Virol* 2005, **77**(1):96-101.
8. Edwards RA, Rohwer F: **Viral metagenomics**. *Nat Rev Microbiol* 2005, **3**(6):504-510.
9. Handelsman J: **Metagenomics: application of genomics to uncultured microorganisms**. *Microbiol Mol Biol Rev* 2004, **68**(4):669-685.
10. Stahl DA, Lane DJ, Olsen GJ, Pace NR: **Characterization of a Yellowstone hot spring microbial community by 5S rRNA sequences**. *Appl Environ Microbiol* 1985, **49**(6):1379-1384.
11. Healy FG, Ray RM, Aldrich HC, Wilkie AC, Ingram LO, Shanmugam KT: **Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose**. *Appl Microbiol Biotechnol* 1995, **43**(4):667-674.
12. Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF: **Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon**. *J Bacteriol* 1996, **178**(3):591-599.
13. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C *et al*: **Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms**. *Appl Environ Microbiol* 2000, **66**(6):2541-2547.
14. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF: **Community-wide analysis of microbial genome sequence signatures**. *Genome Biol* 2009, **10**(8):R85.

15. Flores GE, Bates ST, Knights D, Lauber CL, Stombaugh J, Knight R, Fierer N: **Microbial biogeography of public restroom surfaces**. *PLoS One* 2011, **6**(11):e28132.
16. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F: **Metagenomic analyses of an uncultured viral community from human feces**. *J Bacteriol* 2003, **185**(20):6220-6223.
17. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, Hibberd ML, Liu ET, Rohwer F, Ruan Y: **RNA viral community in human feces: prevalence of plant pathogenic viruses**. *PLoS Biol* 2006, **4**(1):e3.
18. Delwart EL: **Viral metagenomics**. *Rev Med Virol* 2007, **17**(2):115-131.
19. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F: **PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information**. *BMC Bioinformatics* 2005, **6**:41.
20. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, Rohwer F: **Diversity and population structure of a near-shore marine-sediment viral community**. *Proc Biol Sci* 2004, **271**(1539):565-574.
21. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F: **Genomic analysis of uncultured marine viral communities**. *Proc Natl Acad Sci USA* 2002, **99**(22):14250-14255.
22. Persing DH: **Molecular microbiology: diagnostic principles and practice**, 2nd edn. Washington, DC: ASM Press; 2011.
23. Allander T, Andreasson K, Gupta S, Bjerkner A, Bogdanovic G, Persson MA, Dalianis T, Ramqvist T, Andersson B: **Identification of a third human polyomavirus**. *J Virol* 2007.
24. Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, Andersson B: **Cloning of a human parvovirus by molecular screening of respiratory tract samples**. *Proc Natl Acad Sci U S A* 2005, **102**(36):12891-12896.
25. Feng H, Shuda M, Chang Y, Moore PS: **Clonal integration of a polyomavirus in human Merkel cell carcinoma**. *Science* 2008, **319**(5866):1096-1100.
26. Gaynor AM, Nissen MD, Whiley DM, Mackay IM, Lambert SB, Wu G, Brennan DC, Storch GA, Sloots TP, Wang D: **Identification of a Novel Polyomavirus from Patients with Acute Respiratory Tract Infections**. *PLoS Pathog* 2007, **3**(5):e64.
27. Kapoor A, Victoria J, Simmonds P, Wang C, Shafer RW, Nims R, Nielsen O, Delwart E: **A highly divergent picornavirus in a marine mammal**. *J Virol* 2008, **82**(1):311-320.
28. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL: **Microarray-based detection and genotyping of viral pathogens**. *Proc Natl Acad Sci U S A* 2002, **99**(24):15687-15692.
29. Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J *et al*: **Viral discovery and sequence recovery using DNA microarrays**. *PLoS Biol* 2003, **1**(2):E2.
30. Boriskin YS, Rice PS, Stabler RA, Hinds J, Al-Ghusein H, Vass K, Butcher PD: **DNA microarrays for virus detection in cases of central nervous system infection**. *J Clin Microbiol* 2004, **42**(12):5811-5818.

31. Nordstrom H, Falk KI, Lindegren G, Mouzavi-Jazi M, Walden A, Elgh F, Nilsson P, Lundkvist A: **DNA microarray technique for detection and identification of seven flaviviruses pathogenic for man**. *J Med Virol* 2005, **77**(4):528-540.
32. Korimbocus J, Scaramozzino N, Lacroix B, Crance JM, Garin D, Vernet G: **DNA probe array for the simultaneous identification of herpesviruses, enteroviruses, and flaviviruses**. *J Clin Microbiol* 2005, **43**(8):3779-3787.
33. Malanoski AP, Lin B, Wang Z, Schnur JM, Stenger DA: **Automated identification of multiple micro-organisms from resequencing DNA microarrays**. *Nucleic Acids Res* 2006, **34**(18):5300-5311.
34. Wong CW, Heng CL, Wan Yee L, Soh SW, Kartasasmita CB, Simoes EA, Hibberd ML, Sung WK, Miller LD: **Optimization and clinical validation of a pathogen detection microarray**. *Genome Biol* 2007, **8**(5):R93.
35. Palacios G, Quan PL, Jabado OJ, Conlan S, Hirschberg DL, Liu Y, Zhai J, Renwick N, Hui J, Hegyi H *et al*: **Panmicrobial oligonucleotide array for diagnosis of infectious diseases**. *Emerg Infect Dis* 2007, **13**(1):73-81.
36. Phillippy AM, Mason JA, Ayanbule K, Sommer DD, Taviani E, Huq A, Colwell RR, Knight IT, Salzberg SL: **Comprehensive DNA signature discovery and validation**. *PLoS Comput Biol* 2007, **3**(5):e98.
37. Wang Z, Malanoski AP, Lin B, Kidd C, Long NC, Blaney KM, Thach DC, Tibbetts C, Stenger DA: **Resequencing microarray probe design for typing genetically diverse viruses: human rhinoviruses and enteroviruses**. *BMC Genomics* 2008, **9**:577.
38. Chiu CY, Urisman A, Greenhow TL, Rouskin S, Yagi S, Schnurr D, Wright C, Drew WL, Wang D, Weintrub PS *et al*: **Utility of DNA microarrays for detection of viruses in acute respiratory tract infections in children**. *J Pediatr* 2008, **153**(1):76-83.
39. Chiu CY, Greninger AL, Kanada K, Kwok T, Fischer KF, Runckel C, Louie JK, Glaser CA, Yagi S, Schnurr DP *et al*: **Identification of cardioviruses related to Theiler's murine encephalomyelitis virus in human infections**. *Proc Natl Acad Sci U S A* 2008.
40. Mihindukulasuriya KA, Wu G, St Leger J, Nordhausen RW, Wang D: **Identification of a novel coronavirus from a beluga whale by using a panviral microarray**. *J Virol* 2008, **82**(10):5084-5088.
41. Kistler AL, Gancz A, Clubb S, Skewes-Cox P, Fischer K, Sorber K, Chiu CY, Lublin A, Mechani S, Farnoushi Y *et al*: **Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: identification of a candidate etiologic agent**. *Virology journal* 2008, **5**(1):88.
42. Urisman A, Molinaro RJ, Fischer N, Plummer SJ, Casey G, Klein EA, Malathi K, Magi-Galluzzi C, Tubbs RR, Ganem D *et al*: **Identification of a novel Gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant**. *PLoS Pathog* 2006, **2**(3):e25.
43. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, Wang D, DeRisi JL: **E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns**. *Genome Biol* 2005, **6**(9):R78.

44. Watson M, Dukes J, Abu-Median AB, King DP, Britton P: **DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data**. *Genome Biol* 2007, **8**(9):R190.
45. Rehrauer H, Schonmann S, Eberl L, Schlapbach R: **PhyloDetect: a likelihood-based strategy for detecting microorganisms with diagnostic microarrays**. *Bioinformatics* 2008, **24**(16):i83-89.
46. McLoughlin KS: **Microarrays for pathogen detection and analysis**. *Brief Funct Genomics* 2011, **10**(6):342-353.
47. Durbin R: **Biological sequence analysis: probalistic models of proteins and nucleic acids**. Cambridge, UK New York: Cambridge University Press; 1998.
48. Cahan P, Godfrey LE, Eis PS, Richmond TA, Selzer RR, Brent M, McLeod HL, Ley TJ, Graubert TA: **wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data**. *Nucleic Acids Res* 2008, **36**(7):e41.
49. Gelfond JA, Gupta M, Ibrahim JG: **A Bayesian hidden Markov model for motif discovery through joint modeling of genomic sequence and ChIP-chip data**. *Biometrics* 2009, **65**(4):1087-1095.
50. Zhang D, Wells MT, Smart CD, Fry WE: **Bayesian normalization and identification for differential gene expression data**. *J Comput Biol* 2005, **12**(4):391-406.
51. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**(6978):37-43.
52. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al*: **Environmental genome shotgun sequencing of the Sargasso Sea**. *Science* 2004, **304**(5667):66-74.
53. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC *et al*: **Comparative metagenomics of microbial communities**. *Science* 2005, **308**(5721):554-557.
54. Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM, DeLong EF: **Reverse methanogenesis: testing the hypothesis with environmental genomics**. *Science* 2004, **305**(5689):1457-1462.
55. Culley AI, Lang AS, Suttle CA: **Metagenomic analysis of coastal RNA virus communities**. *Science* 2006, **312**(5781):1795-1798.
56. **World Health Report**. In.: World Health Organization; 2004.
57. Dennehy PH: **Acute diarrheal disease in children: epidemiology, prevention, and treatment**. *Infect Dis Clin North Am* 2005, **19**(3):585-602.
58. Berns KP, CR: **Parvoviridae**. In: *Fields Virology*. Edited by Knipe DH, PM, vol. 2, 5th edn: Lippincott Williams & Wilkins; 2007: 2437-2477.
59. Swofford DL: **PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).** Version 4 edn. Sunderland, Massachusettes: Sinauer Associates; 1998.
60. Xing L, Tikoo SK: **Viral RNAs detected in virions of porcine adenovirus type 3**. *Virology* 2004, **321**(2):372-382.

61. Mannhalter C, Koizar D, Mitterbauer G: **Evaluation of RNA isolation methods and reference genes for RT-PCR analyses of rare target RNA**. *Clin Chem Lab Med* 2000, **38**(2):171-177.
62. Gallimore CI, Appleton H, Lewis D, Green J, Brown DW: **Detection and characterisation of bisegmented double-stranded RNA viruses (picobirnaviruses) in human faecal specimens**. *J Med Virol* 1995, **45**(2):135-140.
63. Glass RI, Noel J, Mitchell D, Herrmann JE, Blacklow NR, Pickering LK, Dennehy P, Ruiz-Palacios G, de Guerrero ML, Monroe SS: **The changing epidemiology of astrovirus-associated gastroenteritis: a review**. *Arch Virol Suppl* 1996, **12**:287-300.
64. Moser LA, Schultz-Cherry S: **Pathogenesis of astrovirus infection**. *Viral Immunol* 2005, **18**(1):4-10.
65. Friesen P: **Insect Viruses**. In: *Fields Virology*. Edited by Knipe DH, PM, vol. 1, 5th edn: Lippincott Williams & Wilkins; 2007: 725-727.
66. Smit A, Hubley, R & Green, P.: **RepeatMasker Open-3.0.** 1996-2004.
67. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements**. *Cytogenet Genome Res* 2005, **110**(1-4):462-467.
68. Iturriza-Gomara M, Isherwood B, Desselberger U, Gray J: **Reassortment in vivo: driving force for diversity of human rotavirus strains isolated in the United Kingdom between 1995 and 1999**. *J Virol* 2001, **75**(8):3696-3705.
69. Mustafa H, Palombo EA, Bishop RF: **Improved sensitivity of astrovirus-specific RT-PCR following culture of stool samples in CaCo-2 cells**. *J Clin Virol* 1998, **11**(2):103-107.
70. Klein EJ, Boster DR, Stapp JR, Wells JG, Qin X, Clausen CR, Swerdlow DL, Braden CR, Tarr PI: **Diarrhea Etiology in a Children's Hospital Emergency Department: A Prospective Cohort Study**. *Clin Infect Dis* 2006, **43**(7):807-813.
71. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Res* 1998, **8**(3):175-185.
72. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**(3):186-194.
73. Chou HH, Holmes MH: **DNA sequence quality trimming and vector removal**. *Bioinformatics* 2001, **17**(12):1093-1104.
74. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
75. Kumar S, Wang L, Fan J, Kraft A, Bose ME, Tiwari S, Van Dyke M, Haigis R, Luo T, Ghosh M *et al*: **Detection of 11 common viral and bacterial pathogens causing community-acquired pneumonia or sepsis in asymptomatic patients by using a multiplex reverse transcription-PCR assay with manual (enzyme hybridization) or automated (electronic microarray) detection**. *J Clin Microbiol* 2008, **46**(9):3063-3072.
76. Quan PL, Palacios G, Jabado OJ, Conlan S, Hirschberg DL, Pozo F, Jack PJ, Cisterna D, Renwick N, Hui J *et al*: **Detection of respiratory viruses and**

subtype identification of influenza A viruses by GreeneChipResp oligonucleotide microarray. *J Clin Microbiol* 2007, **45**(8):2359-2364.
77. Lin B, Malanoski AP, Wang Z, Blaney KM, Ligler AG, Rowley RK, Hanson EH, von Rosenvinge E, Ligler FS, Kusterbeck AW *et al*: **Application of broad-spectrum, sequence-based pathogen identification in an urban population**. *PLoS One* 2007, **2**(5):e419.
78. Liu Y, Carbonell J, Klein-Seetharaman J, Gopalakrishnan V: **Comparison of probabilistic combination methods for protein secondary structure prediction**. *Bioinformatics* 2004, **20**(17):3099-3107.
79. Marty AM, Jahrling PB, Geisbert TW: **Viral hemorrhagic fevers**. *Clin Lab Med* 2006, **26**(2):345-386, viii.
80. Pigott DC: **Hemorrhagic fever viruses**. *Crit Care Clin* 2005, **21**(4):765-783, vii.
81. Chiu CY, Rouskin S, Koshy A, Urisman A, Fischer K, Yagi S, Schnurr D, Eckburg PB, Tompkins LS, Blackburn BG *et al*: **Microarray detection of human parainfluenzavirus 4 infection associated with respiratory failure in an immunocompetent adult**. *Clin Infect Dis* 2006, **43**(8):e71-76.
82. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL: **Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray**. *Genome Biol* 2003, **4**(2):R9.
83. Delgado S, Erickson BR, Agudo R, Blair PJ, Vallejo E, Albarino CG, Vargas J, Comer JA, Rollin PE, Ksiazek TG *et al*: **Chapare virus, a newly discovered arenavirus isolated from a fatal hemorrhagic fever case in Bolivia**. *PLoS Pathog* 2008, **4**(4):e1000047.
84. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, Khristova ML, Weyer J, Swanepoel R, Egholm M *et al*: **Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa**. *PLoS Pathog* 2009, **5**(5):e1000455.
85. Jegouic S, Joffret ML, Blanchard C, Riquet FB, Perret C, Pelletier I, Colbere-Garapin F, Rakoto-Andrianarivelo M, Delpeyroux F: **Recombination between polioviruses and co-circulating Coxsackie A viruses: role in the emergence of pathogenic vaccine-derived polioviruses**. *PLoS Pathog* 2009, **5**(5):e1000412.
86. Neumann G, Noda T, Kawaoka Y: **Emergence and pandemic potential of swine-origin H1N1 influenza virus**. *Nature* 2009, **459**(7249):931-939.
87. Gerrard SR, Li L, Barrett AD, Nichol ST: **Ngari virus is a Bunyamwera virus reassortant that can be associated with large outbreaks of hemorrhagic fever in Africa**. *J Virol* 2004, **78**(16):8922-8926.
88. Briese T, Bird B, Kapoor V, Nichol ST, Lipkin WI: **Batai and Ngari viruses: M segment reassortment and association with severe febrile disease outbreaks in East Africa**. *J Virol* 2006, **80**(11):5627-5630.
89. Atasheva S, Wang E, Adams AP, Plante KS, Ni S, Taylor K, Miller ME, Frolov I, Weaver SC: **Chimeric alphavirus vaccine candidates protect mice from intranasal challenge with western equine encephalitis virus**. *Vaccine* 2009, **27**(32):4309-4319.
90. Brandler S, Brown N, Ermak TH, Mitchell F, Parsons M, Zhang Z, Lang J, Monath TP, Guirakhoo F: **Replication of chimeric yellow fever virus-dengue**

serotype 1-4 virus vaccine strains in dendritic and hepatic cells. *Am J Trop Med Hyg* 2005, **72**(1):74-81.
91. Gardner SN, Jaing CJ, McLoughlin KS, Slezak TR: **A microbial detection array (MDA) for viral and bacterial detection**. *BMC Genomics* 2010, **11**:668.
92. Allred AF, Wu G, Wulan T, Fischer KF, Holbrook MR, Tesh RB, Wang D: **VIPR: A probabilistic algorithm for analysis of microbial detection microarrays**. *BMC Bioinformatics* 2010, **11**:384.
93. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.
94. Schultz AK, Zhang M, Leitner T, Kuiken C, Korber B, Morgenstern B, Stanke M: **A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes**. *BMC Bioinformatics* 2006, **7**:265.
95. Ni H, Yun NE, Zacks MA, Weaver SC, Tesh RB, da Rosa AP, Powers AM, Frolov I, Paessler S: **Recombinant alphaviruses are safe and useful serological diagnostic tools**. *Am J Trop Med Hyg* 2007, **76**(4):774-781.
96. Paessler S, Fayzulin RZ, Anishchenko M, Greene IP, Weaver SC, Frolov I: **Recombinant sindbis/Venezuelan equine encephalitis virus is highly attenuated and immunogenic**. *J Virol* 2003, **77**(17):9278-9286.
97. Paessler S, Ni H, Petrakova O, Fayzulin RZ, Yun N, Anishchenko M, Weaver SC, Frolov I: **Replication and clearance of Venezuelan equine encephalitis virus from the brains of animals vaccinated with chimeric SIN/VEE viruses**. *J Virol* 2006, **80**(6):2784-2796.
98. Wang E, Petrakova O, Adams AP, Aguilar PV, Kang W, Paessler S, Volk SM, Frolov I, Weaver SC: **Chimeric Sindbis/eastern equine encephalitis vaccine candidates are highly attenuated and immunogenic in mice**. *Vaccine* 2007, **25**(43):7573-7581.
99. Wang E, Volkova E, Adams AP, Forrester N, Xiao SY, Frolov I, Weaver SC: **Chimeric alphavirus vaccine candidates for chikungunya**. *Vaccine* 2008, **26**(39):5030-5039.
100. Finkbeiner SR, Kirkwood CD, Wang D: **Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea**. *Virology journal* 2008, **5**:117.