

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Viral Detection and Discovery Using DNA Microarrays

Permalink

<https://escholarship.org/uc/item/4wx7t3k1>

Author

Urisman, Anatoly

Publication Date

2007-11-07

Peer reviewed|Thesis/dissertation

Viral Detection and Discovery Using DNA Microarrays

by

Anatoly Urisman

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

UMI Number: 3289306



UMI Microform 3289306

Copyright 2008 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Acknowledgments

I owe my success in graduate school to numerous people and several lucky coincidences. Joining the DeRisi lab was not a coincidence. Even before coming to UCSF, I knew I wanted to work with Joe – more about this coincidence later. Joe is a rare talent, and I am grateful for the opportunity and proud of being his student. I am profoundly grateful to Joe for his mentorship and support. I particularly value his infectious enthusiasm for science, exceptional experimental intuition, and unwavering focus.

I am also very grateful to my co-mentor and the committee chair Don Ganem. Don's clinical as well as scientific insights have been invaluable. Don's advice and encouragement related to scientific and clinical aspects of my training were also very important to me and helped me navigate through several hard decisions during my graduate work.

My deep gratitude also goes to my colleague and friend David Wang. I am grateful to Dave for being a patient and all-knowing mentor, an insightful and generous collaborator, and an encouraging, supportive, and fun lab mate.

I would like to thank my thesis committee members, Drs. Hao Li and Joel Palefsky, as well as my orals committee members, Drs. Eric Brown, Richard Locksley and Ken Dill, for their helpful comments and encouragement

I would like to thank my collaborators, both inside and outside the DeRisi lab. I am particularly grateful to Nicole Fischer, Ross Molinaro, Bob Silverman, Charles Chiu, Kael Fischer, Amy Kistler, Silvi Rouskin, Shoshannah Beck, Tara Greenhow, Patrick Tang, and YT Liu. I also would like to thank all the members of the DeRisi and Ganem labs for being so smart, funny, and a joy to work with.

A special thank you goes to Leslie Spector for proofreading this dissertation.

Now, about the coincidence responsible for my thesis lab choice... I thank heavens for my somewhat spontaneous decision to take a year off before medical school to work with David Botstein and Pat Brown at Stanford. I am grateful to David Botstein and Mike Fero for giving me the job, and to everyone in the two labs for being such an inspiration.

I thank my mother Luda Urisman, my father Victor Urisman, and my brother Igor Urisman for their love and support. I also thank my in-laws Yelena and Vladimir Yasnogorodsky for their encouragement and help. Finally, I would have never pulled it together without the love and devotion of my wife Tatiana Urisman and our two wonderful kids Jacob and Hannah.

Thesis Advisor’s Statement About Co-Author Contributions and Previously Published Work

Chapters 2, 3, and 4 of this dissertation are based on three previously published articles in PLoS Biology, Genome Biology, and PLoS Pathogens, respectively. All three journals allow free copying and redistribution of published materials.

Chapter 2 of this dissertation is based on material published in PLoS Biology titled “Viral Discovery and Sequence Recovery Using DNA Microarrays”. Dr. David Wang, who was a postdoctoral fellow in the laboratory at the time, coordinated this work. Under his supervision, Anatoly wrote and implemented the programs needed to upgrade the original Virochip to the second-generation of the microarray (MegaViro), which extended its scope to all sequenced viruses known at the time. Anatoly also participated in the processing of the SARS samples and in data analysis. In addition, Anatoly was the key contributor to the development and optimization of the “scratching” technique for recovering viral sequences that hybridize to microarray spots. Michael Springer helped with microscopy.

Chapter 3 of this dissertation is based on Anatoly’s work published in Genome Biology titled “E-Predict: a Computational Strategy for Species Identification Based on Observed DNA Microarray Hybridization Patterns”. Kael Fischer, Charles Chiu, and David Wang contributed helpful ideas. Charles Chiu wrote one of the modules used by the E-Predict software. In the course of this work, Anatoly also developed a web application, called

ViroView, for Virochip data retrieval and analysis, which is currently used by our group as well as two other labs.

Chapter 4 of this dissertation is based on material published in PLoS Pathogens entitled “Identification of a Novel Gammaretrovirus in Prostate Tumors of Patients Homozygous for R462Q *RNASEL* Variant”. This work was a collaboration between three labs with three primary contributors. Anatoly performed most of the experiments leading to the detection and molecular characterization of the virus described in the paper. Nicole Fischer also contributed to some of these experiments and in addition performed many important experiments, which did not become part of this publication. Ross Molinaro performed FISH and IHC experiments described in the report. Anatoly also contributed importantly to the writing of the paper and provided coordination and leadership for integrating and editing the original and revised manuscripts.

Chapter 5 of this dissertation is based on preliminary findings in an ongoing study of pediatric viral respiratory infections, which we are planning to publish in the next several months. This work is being carried out in collaboration with Charles Chiu and Silvi Rouskin in the lab and our clinical collaborator Dr. Tara Greenhow. Anatoly has been the project leader, carried out many of the experiments, contributed computational tools, and conducted most of the data analysis.

.....
Joseph DeRisi, PhD

Thesis Advisor

Viral Detection and Discovery Using DNA Microarrays

by

Anatoly Urisman

.....
Don Ganem, MD; Chair

Abstract

DNA microarray-based approaches to viral detection and discovery help overcome many limitations of the existing diagnostic methods. A panviral detection microarray has been developed in our laboratory and is comprised of oligonucleotides derived from ~1200 viral species representing all viral sequences present in the NCBI Nucleotide database as of Fall 2004. The array is capable of detecting viruses from all known viral families, including novel viruses through the use of oligonucleotides derived from the most conserved viral sequences. The platform was successfully used to discover a novel coronavirus in a viral culture sample derived from a patient with Severe Acute Respiratory Syndrome (SARS) during the 2003 outbreak. In addition, a novel retrovirus has been discovered in a subset of prostate tumors from patients with a mutation in the *RNASEL* gene. The platform has also been applied to clinical diagnostics for detecting viruses associated with acute respiratory infections in pediatric patients. The success of the platform has been linked to the development of new algorithms for microarray oligonucleotide selection, tools for microarray-based species identification, optimized techniques for viral nucleic acid extraction and amplification, and novel strategies for viral sequence recovery from analyzed samples.

Table of Contents

Acknowledgments.....	iii
Thesis Advisor’s Statement About Co-Author Contributions and Previously Published Work	v
Abstract.....	vii
Table of Contents.....	viii
List of Tables	xii
List of Figures.....	xiii
Chapter 1: Introduction.....	1
1.1 Motivation.....	1
1.2 The Virochip	2
1.3 Data Analysis.....	4
1.4 Viral Sequence Recovery.....	6
Chapter 2: Identification of the SARS Coronavirus	9
2.1 Abstract.....	9
2.2 Introduction.....	10
2.3 Results.....	11
2.4 Discussion	13
2.5 Materials and Methods.....	14
2.6 Supporting Data	17
2.7 Accession Numbers	17
2.8 Acknowledgements.....	18
Chapter 3: E-Predict.....	22

3.1 Abstract.....	22
3.2 Background.....	22
3.3 Results.....	25
3.3.1 The E-Predict algorithm.....	25
3.3.2 Normalization and similarity metric choice.....	26
3.3.3 Significance estimation.....	28
3.3.4 Examples.....	30
3.4 Discussion.....	34
3.5 Materials and Methods.....	40
3.5.1 Sample preparation and hybridization to microarrays.....	40
3.5.2 Training dataset.....	43
3.5.3 Theoretical energy profiles.....	43
3.5.4 Similarity scores.....	43
3.5.5 Probability estimation.....	44
3.5.6 Iterative E-Predict.....	45
3.5.7 Clustering of HRV serotypes.....	45
3.5.8 PCR.....	46
3.5.9 E-Predict software.....	46
3.6 Additional Files.....	46
3.7 Acknowledgements.....	47
Chapter 4: XMRV.....	57
4.1 Abstract.....	57
4.2 Introduction.....	58

4.3 Results.....	61
4.3.1 Detection of XMRV by microarray-based screening	61
4.3.4 XMRV sequence diversity in samples from different patients.....	67
4.3.5 Detection of XMRV in tumor-bearing prostatic tissues using FISH.....	69
4.3.6 Detection of XMRV in tumor-bearing prostatic tissues using IHC.....	71
4.4 Discussion	71
4.5 Materials and Methods.....	76
4.5.1 Genotyping of patients and prostate tissue processing	76
4.5.2 Microarray screening	77
4.5.3 Genome cloning and sequencing	78
4.5.4 PCR.....	79
4.5.5 Phylogenetic analyses	80
4.5.6 Antibodies	81
4.5.7 FISH.....	82
4.5.8 IHC.....	83
4.6 Supporting Information.....	85
4.7 Accession Numbers	85
4.8 Acknowledgements.....	86
Chapter 5: DNA Microarrays in Viral Diagnostics	104
5.1 Abstract.....	104
5.2 Introduction.....	105
5.3 Results.....	107
5.4 Discussion	110

5.5 Materials and Methods.....	112
5.6 Acknowledgements.....	115
Chapter 6: Concluding Remarks.....	118
6.1 Pros and Cons of Using DNA Microarrays for Viral Detection.....	118
6.2 Future Directions	119
References.....	122

List of Tables

Table 2.1: Oligonucleotides Hybridizing to Viral Sample	21
Table 3.1: Normalization Methods	54
Table 3.2: Similarity Metrics	54
Table 3.3: Example 1: Hepatitis Microarray – Predicted Virus Profiles	55
Table 3.4: Example 1: Hepatitis Microarray – Oligonucleotides Contributing to the Hepatitis B Virus Profile Prediction	55
Table 3.5: Example 2: FluA, RSV Double Infection.....	56
Table 3.6: Example 3: SARS Microarray	56
Table 4.1: XMRV Screening by <i>gag</i> Nested RT-PCR	101
Table 4.2: Frequency of XMRV Infected Prostatic Cells Determined by FISH	101
Table 4.3: Computational Viral Species Predictions Using E-Predict for the Virochip Microarrays Shown in Figure 4.1	102
Table 4.4: PCR Primers Used for Sequencing of XMRV Genomes	103
Table 4.5: Age, Clinical Parameters, and Geographical Locations of XMRV-Positive Prostate Cancer Cases	103
Table 5.1: Detection of Seven Common Respiratory Viruses by DFA and Microarray	117
Table 5.2: Detection of Double Infections by Microarray and DFA	117

List of Figures

Figure 2.1: Microarray Detection of SARS Coronavirus and Recovery of Viral Sequences	19
Figure 2.2: Viral DNA Recovery and Sequencing Scheme.....	20
Figure 3.1: E-Predict Algorithm	48
Figure 3.2: Evaluation of Normalization and Similarity Metric Parameters	50
Figure 3.3: Estimation of Significance of Individual Similarity Scores.....	51
Figure 3.4: HRV Serotype Discrimination Using E-Predict Similarity Scores	53
Figure 4.1: XMRV Detection by DNA Microarrays and RT-PCR	87
Figure 4.2: Complete Genome of XMRV.....	88
Figure 4.3: Phylogenetic Analysis of XMRV Based on Complete Genome Sequences..	90
Figure 4.4: Multiple-Sequence Alignment of Protein Sequences from XMRV and Related MuLVs Spanning SU Glycoprotein VRA, VRB, and VRC, Known to Determine Receptor Specificity.....	91
Figure 4.5: Multiple-Sequence Alignment of 5' <i>gag</i> Leader Nucleotide Sequences from XMRV and Related MuLVs	92
Figure 4.6: Comparison of XMRV Sequences Derived from Tumor Samples of Different Patients.....	93
Figure 4.7: Detection of XMRV Nucleic Acid in Prostatic Tissues Using FISH.....	94
Figure 4.8: Characterization of XMRV-Infected Prostatic Cells by FISH and FISH/Immunofluorescence	95
Figure 4.9: Detection of XMRV Protein in Prostatic Tissues Using Immunostaining.....	97

Figure 4.10: Phylogenetic Analysis of XMRV Based on Predicted Gag, Pro-Pol, and Env Polyproteins	98
Figure 4.11: Comparison of XMRV U3 Region to Representative Non-Ecotropic Sequences.....	99
Figure 5.1: Viruses Detected by DFA and Microarray.....	116

Chapter 1: Introduction

Copyright: © 2006 Anatoly Urisman et al.

1.1 Motivation

Numerous human diseases exist for which viral etiologies are suspected, yet specific causal agents are not known. Among these are up to 20% of cases of acute hepatic failure [1], up to 35% of cases of acute aseptic meningitis [2] and acute encephalitis [3], up to 50% of cases of acute respiratory infections [4-6], and numerous other conditions. In addition, infectious agents may be involved in the pathogenesis of a number of chronic conditions, most notably such disorders as chronic inflammation, autoimmune and degenerative conditions, as well as some forms of cancer. While it is unlikely that all of these diseases are caused by viruses, identifying causative agents in even a modest number of cases will have profound implications for understanding, diagnosis, and treatment of these conditions.

New approaches to viral discovery are needed to overcome the shortcomings of the existing methods. These methods include viral culture, electron microscopy, serology-based methods, PCR-based methods, and techniques based on subtractive hybridization. These methods have been critical for identifying many important human and non-human pathogens. However, each of these methods has at least one serious limitation. For example, many viruses are refractory to culture. Electron microscopy fails unless virus is present at a high titer. Serology- and PCR-based methods are targeted at specific viruses

and therefore are limited in scope. Finally, subtractive hybridization techniques are difficult to troubleshoot and essentially impossible to scale up for high throughput.

In addition to the need for novel approaches to viral discovery, new methodologies for comprehensive and unbiased detection of known viruses are also acutely needed.

Traditional methods, such as antibody-based tests and PCR, target only one or a few common agents, and may be too specific to detect emerging strains.

1.2 The Virochip

To address these needs, a novel viral detection and discovery method has been recently developed in our laboratory [7]. The method takes advantage of a DNA microarray (Virochip) consisting of 70-mer oligonucleotides derived from sequences of publicly available viral genomes, including human, animal, and plant viruses, as well as bacteriophages. The elements on the microarray are chosen from the most conserved segments of the viral genomes and therefore have the highest probability of being shared by as of yet undiscovered viruses.

In its initial version [7], the Virochip contained ~1600 oligonucleotides derived from ~140 viruses from families with known human pathogens and was heavily biased toward detection of respiratory viruses. This microarray was successfully used to detect several known respiratory pathogens from viral culture as well as clinical samples.

Based on the success of the original Virochip, we used the approach of choosing the most conserved viral sequences as the algorithm for picking microarray oligonucleotides and extended the design to include all fully sequenced NCBI Reference viral genomes available at the time (August 2002). This included ~950 viruses of humans, animals, plants and bacteria. In March 2003, we used this new version of the microarray (nicknamed “MegaViro”) to identify a novel coronavirus (SARS CoV) in a viral culture sample derived from a patient with SARS ([8]; Chapter 2). Detection of SARS CoV was an important validation of the overall strategy of using evolutionarily conserved probes, as this divergent Coronaviridae member was detected entirely via cross-hybridization to oligonucleotides derived from other viruses. In addition, MegaViro microarrays were used to discover a novel gammaretrovirus in a subset of prostate cancer samples derived from patients with a mutation in the *RNASEL* gene ([9]; Chapter 4).

More recently, the microarray has undergone another update (Kael Fischer et al. 2004, unpublished). In addition to the most conserved sequences, which tend to represent conservation on genus and family levels, new oligonucleotides were added to represent conserved elements at sub-genus and even species levels. This design was based on all partial as well as complete viral sequences in GenBank as of June 2004 (~277,000 sequences) and added ~9000 new oligonucleotides. This third generation microarray (nicknamed “Viro3”) contains ~22,000 oligonucleotides derived from ~1800 viral species. This microarray is now routinely used in our laboratory for viral detection and discovery projects targeting a broad range of diseases and a variety of patient samples. One example of such a project is the application of the microarray to clinical diagnostics

and detection of viral pathogens associated with acute respiratory tract infections in children (Chapter 5).

1.3 Data Analysis

Very early in the development of the Virochip technology, it became apparent that the existing microarray data analysis methods were insufficient for interpreting the complex hybridization patterns typically observed on the Virochip. The main goal of the analysis is to identify the species of viruses, which are most likely to be present in the sample under investigation, given an observed hybridization pattern. This task is significantly complicated by the great nucleic acid complexity present in most clinical samples, the presence of multiple species in the same sample, and the infeasibility of obtaining positive controls for each virus targeted by the microarray.

Visual inspection of oligonucleotides with the highest hybridization intensities is occasionally sufficient to make an accurate conclusion as to what virus is present in the sample. However, most patterns are too complex to be interpreted by visual inspection alone. In addition, the technique is subjective and thus suffers from user-to-user variation.

Hierarchical clustering is a powerful technique (e.g. [10]), which groups oligonucleotides with similar intensity profiles in a set of microarray experiments, allowing, in some cases, visualization of groups of oligonucleotides belonging to the same viral species. However, in many cases clustering fails because observed viral signatures are frequently comprised of a small number of oligonucleotides with drastically different intensities. In addition, a

viral signature may represent only a minor fraction of the total mostly non-viral signal on a given microarray, in which case the clustering is typically “driven” by the predominating nonviral signal.

Data analysis can be greatly simplified through experimental efforts that enrich viral sequences and reduce overall nucleic acid complexity. Preliminary work in our laboratory (Patrick Tang et al. 2006, unpublished) shows that viral particle concentration via filtration or ultracentrifugation, enzymatic digestion of host nucleic acid, and even differential hybridization techniques prior to amplification can significantly reduce nucleic acid complexity in the analyzed samples. Therefore, we are constantly working on improving our nucleic acid extraction and amplification protocols. However, despite these efforts, hybridization pattern interpretation remains a challenging problem and is an area of active research in our laboratory.

One of the tools developed as part of this ongoing work is an algorithm called E-Predict ([11]; Chapter 3). E-Predict compares an observed microarray pattern to a set of theoretically derived hybridization energy profiles calculated from available viral genomic sequences. The result is a list of viruses whose profiles are most similar to the observed hybridization patterns. The algorithm also incorporates a method to estimate the statistical significance of each comparison based on empirical null distributions derived from previous cumulative microarray data. E-Predict is a robust and sensitive tool, which has become the gold standard of Virochip data interpretation in our laboratory.

Another data analysis technique we have developed examines the significance of each oligonucleotide on the microarray as an independent observation (Anatoly Urisman et al. 2005, unpublished; [12]). In its current implementation the significance of each oligonucleotide is calculated as the Z-score, i.e. the difference in standard deviations between the observed intensity and the median intensity of the oligonucleotide over a large number (50 or more) of predominantly negative control experiments. The result is a list of the most significant oligonucleotides, which is examined manually for possible enrichment of oligonucleotides from the same family. The technique, known in the lab as Single Oligo Analysis, in this simplest implementation requires human intelligence to decipher a meaningful pattern among the most significant oligonucleotides. As such, the method is relatively time consuming and suffers from user-to-user variation. Despite these limitations, the method has proven especially useful in detecting very weak signatures and signatures not detected by E-Predict. For example, Single Oligo Analysis was recently used to detect a very weak parainfluenza 4 signature in a bronchial aspirate sample from a patient with acute respiratory infection and respiratory failure [12]. E-Predict failed to detect parainfluenza 4 in this sample, because we did not have an energy profile from this only partially sequenced virus.

1.4 Viral Sequence Recovery

Detection of a microarray signature suggestive of a specific virus or virus family is not sufficient to unambiguously identify the virus in the sample, and frequently is the start of a follow-up process aimed at recovering viral nucleic acid from the sample. This is

particularly important for viral discovery projects and in cases of detecting unusual signatures consistent with divergent viruses.

In the simplest case, the virus in question can be confirmed using established PCR assays and subsequent sequencing of the amplified fragments. If such assays are not available or fail to yield a product, PCR primers can be designed based on oligonucleotides comprising the detected microarray signature. The signature must contain oligonucleotides from at least two genomic regions separated by a distance that can be spanned by a typical PCR reaction (<2 kb). In order to find the best primer, we typically align a group of viruses that share homology across the length of a given oligonucleotide. In addition, it can be advantageous to introduce degenerate bases to enable detection of diverse species.

Sequence recovery by *scratching* is a technique which captures nucleic acids hybridizing to the microarray oligonucleotides. This technique was developed during our work on SARS coronavirus ([8]; Chapter 2) and was used to recover a 1.1 kb clone from the 3' end of the virus, which was later used by a CDC team to obtain the complete genome of SARS CoV [13]. In that study, we were able to recover viral sequences directly from the Virochip. Fluorescence microscopy was used to visualize a target spot on the microarray, and nucleic acids were recovered by passing a tungsten micromanipulator needle across the spot surface. The material picked up by the needle was then PCR amplified, cloned, and sequenced. Since then, the protocol has been simplified, such that desired target oligonucleotides are hand-spotted on a glass slide at well-spaced and easily identifiable

locations, enabling easy access to the spots without microscopy or micromanipulation tools. This technique was successfully used to recover the first few clones from a novel gammaretrovirus, XMRV, found in a subset of prostate tumors from patients with R462Q *RNASEL* mutation ([9]; Chapter 4). Rough quantitation using colony hybridization shows that scratching carried out in this manner results in ~100 fold enrichment of target viral sequences (Anatoly Urisman et al. 2004, unpublished).

Another sequence recovery technique often used in our laboratory is screening of libraries made by cloning randomly amplified PCR fragments used for microarray hybridization. This can be accomplished either by colony hybridization using target microarray oligonucleotides as probes or by PCR using a plasmid backbone primer in combination with a primer derived from a target oligonucleotide.

Chapter 2: Identification of the SARS Coronavirus

Citation: Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J, Latreille JP, Wilson RK, Ganem D, DeRisi JL. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.* 2003 Nov;1(2):E2.

Copyright: © 2003 David Wang et al.

2.1 Abstract

Because of the constant threat posed by emerging infectious diseases and the limitations of existing approaches used to identify new pathogens, there is a great demand for new technological methods for viral discovery. We describe herein a DNA microarray-based platform for novel virus identification and characterization. Central to this approach was a DNA microarray designed to detect a wide range of known viruses as well as novel members of existing viral families; this microarray contained the most highly conserved 70-mer sequences from every fully sequenced reference viral genome in GenBank. During an outbreak of severe acute respiratory syndrome (SARS) in March 2003, hybridization to this microarray revealed the presence of a previously uncharacterized coronavirus in a viral isolate cultivated from a SARS patient. To further characterize this new virus, approximately 1 kb of the unknown virus genome was cloned by physically recovering viral sequences hybridized to individual array elements. Sequencing of these fragments confirmed that the virus was indeed a new member of the coronavirus family.

This combination of array hybridization followed by direct viral sequence recovery should prove to be a general strategy for the rapid identification and characterization of novel viruses and emerging infectious diseases.

2.2 Introduction

Over the past two decades, technological advances in molecular biology have fuelled progress in the discovery of new pathogens associated with human diseases. The identification of novel viruses such as hepatitis C virus [14], sin nombre virus [15], and Kaposi's sarcoma herpesvirus [16] has relied upon a diverse range of modern molecular methods such as immunoscreening of cDNA libraries, degenerate PCR, and representational difference analysis, respectively. In spite of these successes, there remain numerous syndromes with suspected infectious etiologies that continue to escape identification efforts, in part due to limitations of existing methodologies for viral discovery [17, 18]. These limitations, coupled with the constant threat posed by newly emerging infectious diseases of unknown origin, necessitate that new approaches be developed to augment the repertoire of available tools for pathogen discovery.

We have previously described a prototype DNA microarray designed for highly parallel viral detection with the potential to detect novel members of known viral families [7]. This microarray contained approximately 1600 oligonucleotides representing 140 viruses. Building upon this foundation, a more comprehensive second-generation DNA microarray consisting of 70-mer oligonucleotides derived from every fully sequenced reference viral genome in GenBank (as of August 15, 2002) was constructed. The most

highly conserved 70-mers from each virus were selected as described by Wang et al. [7] to maximize the probability of detecting unknown and unsequenced members of existing families by cross-hybridization to these array elements. On average, ten 70-mers were selected for each virus, totaling approximately 10,000 oligonucleotides from approximately 1,000 viruses. The objective was to create a microarray with the capability of detecting the widest possible range of both known and unknown viruses. This panviral microarray was used as part of the global effort to identify a novel virus associated with severe acute respiratory syndrome (SARS) in March 2003, as reported by Ksiazek et al. [19]. We describe here the experimental details of the microarray methodology for novel virus identification, using the SARS outbreak as an example.

2.3 Results

During the initial phase of research into the etiology of SARS, an unknown virus was cultured in Vero cells from a patient suffering from SARS [19]. Total nucleic acid purified from this viral culture, as well as a control culture, was obtained from the Centers for Disease Control and Prevention on March 22, 2003. These two samples, along with additional controls (HeLa cell RNA and water alone), were amplified and hybridized within 24 h to the virus DNA microarray. The strongest hybridizing array elements from the infected culture were derived from two families: astroviridae and coronaviridae. Table 2.1 lists the oligonucleotides from these families with the greatest hybridization intensity. By comparison, these oligonucleotides yielded essentially background levels of hybridization in the various control arrays performed in parallel. The initial suggestion from this hybridization pattern was that members of both of these

viral families might be present. However, alignment of the oligonucleotides using ClustalX revealed that all four hybridizing oligonucleotides from the astroviridae and one oligonucleotide from avian infectious bronchitis virus (IBV) (GenBank NC_001451), an avian coronavirus, shared a core consensus motif spanning 33 nucleotides (data not shown); thus, these five oligonucleotides behaved essentially as multiple redundant probes for the same sequence. This motif is known to be present in the 3' UTR of all astroviruses and the avian coronaviruses [20], but appears to be absent in the available sequenced mammalian coronaviruses (bovine coronavirus, murine hepatitis virus [MHV], human coronavirus 229E, porcine epidemic diarrhea virus, and transmissible gastroenteritis virus). The other three hybridizing oligonucleotides were derived from three conserved regions within the ORF1AB polyprotein common to all coronaviruses (Figure 2.1). Based on the aggregate hybridization pattern, the virus appeared to be a novel member of the coronavirus family.

To further characterize this virus, we sequenced fragments of the viral genome using two complementary approaches. First, BLAST alignment of two of the hybridizing viral oligonucleotides, one each from bovine coronavirus and human coronavirus 229E, to the IBV genome indicated that the oligonucleotides possessed homology to distinct conserved regions within the NSP11 gene (BLAST identity matches of 42/47 and 26/27, respectively). A pair of PCR primers was designed to amplify the intervening sequences between the two conserved regions, and a fragment that possessed 89% identity over 37 amino acids to MHV, a murine coronavirus, was obtained (Figure 2.1; sequence available as Data S1).

In a parallel approach, we directly recovered hybridized viral sequences from the surface of the microarray. This procedure took advantage of the physical separation achieved during microarray hybridization, which effectively purified the viral nucleic acid from other nucleic acid species present in the sample. Using a tungsten needle, the DNA microarray spot corresponding to the conserved 3' UTR motif was repeatedly scraped and the hybridized nucleic acid was recovered. This material was subsequently amplified, cloned, and sequenced (Figure 2.2). The largest clone spanned almost 1.1 kb; this fragment encompassed the 3' UTR conserved motif and extended into the most 3' coding region of the viral genome. BLAST analysis revealed 33% identity over 157 amino acids to MHV nucleocapsid, thus confirming the presence of a novel coronavirus (see Figure 2.1; see Data S1). We subsequently confirmed results obtained from both strategies described above by using a random-primed RT-PCR shotgun sequencing approach that generated contigs totaling ~25 kb of viral genome sequence (see Data S1).

2.4 Discussion

In this report, we have demonstrated the viability of detecting novel pathogens via cross-hybridization to highly conserved sequence motifs. With the recent sequencing of the complete SARS coronavirus genome (GenBank NC_004718) [21, 13], we were able to retrospectively determine the degree of nucleotide identity shared between the hybridizing oligonucleotides and the new coronavirus genome (see Table 2.1). Stretches of relatively uninterrupted nucleotide identity as short as 25 nucleotides yielded clearly

detectable hybridization signal, confirming that novel viruses with only limited homology to known viruses can be successfully detected by this strategy.

A key feature of this approach is that direct recovery of hybridized material from the microarray provides a rapid route for obtaining sequences of novel viruses. By contrast, conventional strategies for subsequent sequence identification would require time-consuming steps such as library screening or additional rounds of PCR primer design and synthesis. In the case of SARS, we were able to ascertain within 24 h that a novel coronavirus was present in the unknown sample, and partial genome sequences of this virus were obtained over the next few days without the need for specific primer design. To our knowledge, this is the first demonstration of the feasibility and utility of directly recovering nucleic acid sequences from a hybridized DNA microarray. In light of the continuous threat of emerging infectious diseases, this overall approach will greatly facilitate the rapid identification and characterization of novel viruses.

2.5 Materials and Methods

Nucleic acid isolation. Total nucleic acid was purified using the automated NucliSens extraction system (BioMerieux, Durham, NC, USA). Following the manufacturer's instructions, 100 μ l of each specimen was added to tubes containing 900 μ l of prewarmed NucliSens lysis buffer and incubated at 37°C for 30 min with intermittent mixing. Fifty microliters of silica suspension provided in the extraction kit was added to each tube and mixed. The mixtures were then transferred to a nucleic acid extraction cartridge and

loaded onto the extractor workstation for processing. Approximately 50 µl of total nucleic acid eluate was recovered.

Amplification. For the culture supernatants, 450 ng of nucleic acid was used as input for the amplification protocol. In parallel, 50 ng of HeLa cell RNA was used as a positive amplification control and water was used for a negative control. Samples were amplified using a random-primer protocol as described [7], with the following modifications: first- and second-strand synthesis were primed using primer-A (5'-GTTTCCCAGTCACGATANNNNNNNNN) followed by PCR amplification using primer-B (5'-GTTTCCCAGTCACGATA) for 40 cycles. Aminoallyl-dUTP was incorporated into the PCR product using an additional 20 cycles of thermocycling. A detailed protocol is available as Protocol S1.

Microarray hybridization and analysis. DNA microarrays were printed and hybridized essentially as described [7], with the following modifications: for array printing, a single-defined 70mer (spike-70) was mixed with each viral oligonucleotide in a 1:50 ratio. Array hybridizations used Cy5-labeled amplified probe from either virally infected cultures or controls (mock-infected culture, HeLa RNA, or water); a reference signal for every spot on each array was generated by using a Cy3-labeled version of the reverse complement of spike-70. Oligonucleotides were assessed by Cy5 intensity. Oligonucleotides from the astrovirus and coronavirus families that passed a conservative, arbitrarily set cutoff of $(\text{Cy5infection} - \text{Cy5mock}) > 1500$ intensity units are listed in Table 2.1. Additional oligonucleotides from these families and their homology to the SARS

coronavirus are listed in Table S1. Array data has been deposited in the Gene Expression Omnibus (GEO) database (accession number GSE546). A complete list of the viral oligonucleotide sequences on the microarray is also available as Table S2.

Conventional PCR using array element sequences. PCR primers were designed by aligning the hybridizing oligonucleotides (Oligo IDs 15081544_766 and 12175745_728) to the IBV genome (Fwd: 5'-TGTTTTGGAATTGTAATGTGGAT; Rev: 5'-TACAAACTACCTCCATTACAGCC) and selecting stretches of near-identity. Primer-B-amplified material was used as the template for 35 cycles of thermocycling using the following program: 94°C for 30 s, 56°C for 30 s, and 72°C for 60 s.

Direct sequence recovery from the microarray. Amplified viral sequences hybridized to individual microarray spots were recovered by scraping a 100 µm area of the microarray using a tungsten wire probe (Omega Engineering, Inc.) mounted on a micromanipulator while visualized by fluorescence microscopy (Nikon TE300). Recovered material was PCR amplified using primer-B, cloned into pCR2.1TOPO (Invitrogen), and sequenced. A detailed protocol is available as Protocol S2.

Shotgun sequencing. Primer-B-amplified nucleic acid (see above) was cloned in pCR2.1TOPO, plated on 2xYT/kan plates, and grown overnight at 37°C. White colonies were picked into 384-well plates containing 2xYT/kan plus 8% glycerol and incubated overnight at 37°C. DNA was purified by magnetic bead isolation. DNA sequencing involved adding 3 µl of water to each bead pellet, followed by 3 µl of Big Dye terminator

(v3.1) sequencing cocktail, and incubation for 35 cycles of 95°C for 5 s, 50°C for 5 s, and 60°C for 2 min. Reaction products were ethanol precipitated, resuspended in 25 µl of water, and loaded onto the ABI 3730xl sequencer. The resulting sequence reads were trimmed to remove primer sequences from the RT-PCR step and then assembled by Phrap (P. Green, unpublished data). Resulting contigs were screened by blast to remove any contigs with high human or monkey sequence similarity. The remaining contigs were edited to high quality, making any obvious joins. (Sequences are available as Data S1.)

2.6 Supporting Data

The following supporting data are available as a web supplement to the PLoS Biology article describing this work (<http://biology.plosjournals.org/>):

Data S1: Recovered SARS Coronavirus Sequences

Protocol S1: Round A/B/C Random Amplification Protocol

Protocol S2: Microarray Sequence Recovery Protocol

Table S1: List of Coronavirus and Astrovirus Microarray Oligonucleotides

Table S2: Complete List of Microarray Oligonucleotides

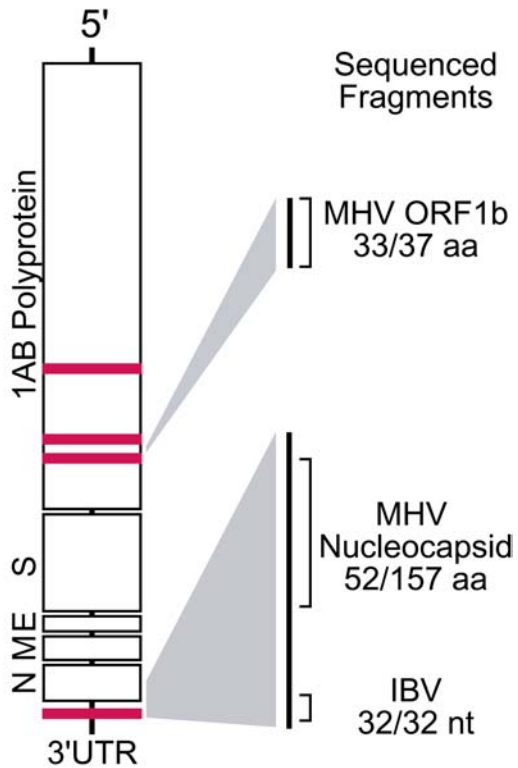
2.7 Accession Numbers

NCBI Gene Expression Omnibus accession number for the microarray series is GSE546.

2.8 Acknowledgements

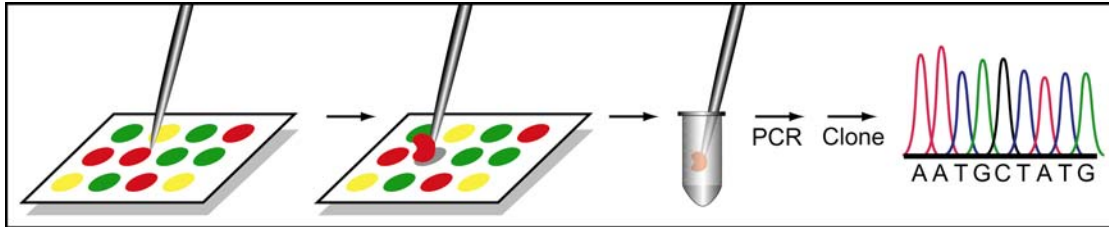
This work was supported by a grant from the Sandler Program for Asthma Research (to JLD).

Figure 2.1: Microarray Detection of SARS Coronavirus and Recovery of Viral Sequences



Red bars indicate physical location of virus microarray DNA elements mapped to a generic coronavirus genome. Portions of the coronavirus genome sequenced by physical recovery and PCR methods are highlighted with homologies to known coronaviruses. Abbreviations: aa, amino acid; nt, nucleotide.

Figure 2.2: Viral DNA Recovery and Sequencing Scheme



Hybridized viral sequences were physically scraped from a DNA microarray spot, amplified, cloned, and subsequently sequenced.

Table 2.1: Oligonucleotides Hybridizing to Viral Sample

Oligo ID	Family	Virus	Oligo Sequence	Identities to SARS Virus ^a
9626535_1099	Corona	IBV	TAGAGTAGGTATAAAGATGCCAGTGCCGG <u>GGCCACGCGGAGTAC-GATCGAGGGTACAG</u> CACTAGGACGCC	29/29
9626535_568	Corona	IBV	<u>TTATGGGTTGGGATTATCCTAAGTGTGATAGAGCAATGCCTAA-TTTGTTGCGTATAGCAGCATCCTTAGT</u>	39/43
15081544_766	Corona	Bovine corona	CTG <u>GGCTGTAATGGAGGTAGTTTGTATGTTAATAAACATGCATT-CACACT</u> AAACCCTTTTCTAGGGCAGC	44/48
12175745_728	Corona	Human 229E	AAATGGATGGCTTG <u>TGTTTGTGTTTGGAAATTGTAATGTGGAT</u> ATGT-ACCCTGAATTCTCAATTGTTGCAG	25/27
9635576_275	Astro	Turkey astro	GATGCTTGAGAAAATTGATGC <u>CGAGGCCACGCCGGGTAGGATC-GAGGGTACAG</u> CATCGGTGCACTAACT	29/32
9635572_255	Astro	Ovine astro	ATCACTTCAATCC <u>CGAGGCCACGCCGAGTAGGATCGAGGGTAC-AGGATTGTTTGATTTTTTAATCAATTA</u>	30/32
20514394_269	Astro	Avian nephritis	ACTTTCC <u>CGAGGCCACGCCGAGTAGCATCGAGGGTACAG</u> GAAA-GCTGGGACCATTGCATAGTCAACTAAT	28/32
9630726_269	Astro	Human astro	TGCACATCTGGAAGCCGC <u>GGCCACGCCGAGTAGGAACGAGGGT-ACAG</u> CTTCCTCTTTCTGTCTCTGT	26/29

Underlined nucleotides represent regions of identity to the SARS coronavirus. The table does not include reverse complement oligonucleotides. ^aBLAST identities to the SARS coronavirus genome (NC_004718).

Chapter 3: E-Predict

Citation: 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.

Copyright: © 2005 Anatoly Urisman et al.

3.1 Abstract

In metagenomic applications concerned with identifying particular microbial species present in environmental or clinical samples, DNA microarrays provide a time-efficient and cost-effective alternative to other methods. However, automated tools for reliable species identification based on observed microarray hybridization patterns are lacking. Here we present an algorithm, E-Predict, for microarray-based species identification. E-Predict compares an observed hybridization pattern with a set of theoretical energy profiles. Each profile represents a species that may be identified. We show the application of the algorithm to our recently described platform for viral detection and discovery, illustrate its versatility on a set of clinical examples, and discuss its relevance to other metagenomic applications.

3.2 Background

Metagenomics, an emerging field of biology, utilizes DNA sequence data to study unculturable microorganisms found in the natural environment. Metagenomic

applications include studies of diversity and ecology in microbial communities, detection and identification of representative species in environmental and clinically relevant samples, and discovery of genes or organisms with novel or useful functional properties (see recent reviews [22] [23] [24] [25]).

Common to all of these applications is the task of identifying (and often quantifying the abundance of) individual genes, species, or even groups of species from a large and often complex sequence space being explored. In the most general approach, shotgun sequencing is used both to identify and quantify the individual sequences in a sample of interest [26] [27] [28] [29]. In a more targeted approach, PCR is used to amplify a particular subset of sequences, which can then be cloned and analyzed. For example, 16S rRNA sequences are frequently used to identify bacterial and archaeal species [30] [31] [32] [33]. Another approach is based on functional screening of shotgun expression libraries to identify DNA fragments encoding proteins with desirable activities [34] [35] [36].

DNA microarrays are also emerging as an important tool in metagenomics [37] [23] [38] [39]. Particularly in applications concerned with real-time identification of known or related species, microarrays provide a practical high-throughput alternative to costly and time-consuming cloning and repetitive sequencing. For example, as previously reported, DNA microarrays have been successfully used to detect known viruses [7] [40] [41] [42] and to discover a novel human viral pathogen [8]. Other metagenomic applications where

microarrays have great potential include monitoring food and water quality [43], tracking of bioremediation progress [23] [44], and assessment of biological threat [45].

Use of DNA microarrays in metagenomics introduces a series of analytical challenges. First, the sequence space to explore may be very large, especially in the case of environmental samples. Given the technological constraints on the total number of probes that can be placed on a microarray, improved algorithms are required for optimal probe selection to maximize coverage. Second, microarray data generated in metagenomic studies can be very complex. In the case of viral diagnostics, nucleic acid extracted from clinical specimens usually contains host and bacterial contaminants in addition to viral RNA and DNA. As a result, hybridization patterns are complicated by substantial amounts of noise introduced by specific and non-specific cross-hybridization that cannot be anticipated or controlled. Third, multiple and potentially closely related species may be present in a single sample resulting in complex or even overlapping hybridization patterns. Finally, a species identification strategy based on the use of experimentally derived patterns alone is not feasible, because such empirical controls can be obtained only for a limited number of species available as pure cultures or genomic clones. New analytical tools capable of overcoming these challenges are acutely needed.

We have previously reported the development of a DNA microarray-based platform for viral detection and discovery [8] (NCBI GEO [46] accession GPL366). Briefly, the platform employs a spotted 70-mer oligonucleotide microarray containing approximately 11,000 oligonucleotides representing the most conserved sequences from 954 distinct

viruses corresponding to every NCBI reference viral genome available at the time of design. Nucleic acids are extracted from a sample of interest, typically a clinical specimen, amplified and labeled using random-primed reverse transcription, second strand synthesis, and PCR. The labeled DNA is then hybridized to the microarray, and hybridization patterns are analyzed to identify particular viruses present in the sample.

Here we report a computational strategy, called E-Predict, for species identification based on observed microarray hybridization patterns (Figure 3.1a). Using this strategy, an observed pattern of intensities is compared to a set of theoretical hybridization energy profiles, representing species with known genomic sequence. We illustrate the use of E-Predict on data obtained with our viral detection microarray and demonstrate its effectiveness in identifying viral species in a variety of clinical specimens. Based on these results, we argue that E-Predict is relevant for a broad range of microarray-based metagenomic applications.

3.3 Results

3.3.1 The E-Predict algorithm

Theoretical hybridization energy profiles were computed for every completely sequenced reference viral genome available in GenBank as of July 2004 (1,229 distinct viruses).

This set of profiles included all viruses represented on the microarray and many viruses whose genomes became available after the array design had been completed. All microarray oligonucleotides expected to hybridize to a given viral genome were identified using nucleotide BLAST alignment [47]. Free energy of hybridization (ΔG)

was then computed for each alignment using the nearest-neighbor method [48] [49]. Oligonucleotides that failed to produce a BLAST alignment were assumed to have hybridization energies equal to zero. Thus, a given theoretical energy profile consists of the non-zero hybridization energies calculated for the subset of oligonucleotides producing a BLAST alignment to the corresponding genome. Collectively, the energy profiles of all the viruses constitute a sparsely populated energy matrix, where each row corresponds to a viral species and each column corresponds to an oligonucleotide from the microarray (Figure 3.1b).

The general E-Predict algorithm for interpreting observed hybridization patterns is shown in Figure 3.1b. A vector of oligonucleotide intensities is normalized and compared to every normalized profile in the energy matrix using a simple similarity metric, resulting in a vector of raw similarity scores. Each element in this vector denotes the similarity between the observed pattern and one of the predicted profiles for a species represented in the energy matrix. The statistical significance of the raw similarity scores is estimated using a set of experimentally obtained null probability distributions. Profiles associated with statistically significant similarity scores suggest the presence of the corresponding viral species in the sample.

3.3.2 Normalization and similarity metric choice

In order to optimize the ability of E-Predict to discriminate between true positive and true negative predictions, we first evaluated the performance of several commonly used normalizations and similarity metrics. For this purpose we constructed a training dataset

of 32 microarrays obtained from samples known to be infected by specific viruses. Fifteen microarrays represented independent hybridizations of RNA extracted from HeLa cells, a human cell line permanently infected with human papillomavirus type 18 (HPV18). The remaining microarrays were obtained from seventeen independent clinical specimens from children with respiratory tract infections. Ten specimens contained respiratory syncytial virus (RSV) and seven contained influenza A virus (FluA) as determined by direct fluorescent antibody (DFA) test.

Intensity and energy vectors were independently normalized using sum, quadratic, unit-vector or no normalization (Table 3.1). Similarity scores between the vectors were computed using dot product, Pearson correlation, uncentered Pearson correlation, Spearman rank correlation, or similarity based on Euclidean distance (Table 3.2). All non-equivalent combinations of intensity vector normalization, energy vector normalization, and similarity metrics were evaluated. For each combination, similarity scores were obtained by comparing every microarray in the training dataset to every virus profile in the energy matrix. The performance of each combination was then evaluated by calculating the separation between the score obtained for the correct (match) virus profile and the best scoring non-match profile from either the same or a different virus family (Figures 3.2a and 3.2b, respectively). We defined separation as the difference between the similarity scores of a match and the appropriate non-match profiles, divided by the range of all similarity scores on a given microarray. Using this statistic, a value of one corresponds to the best possible separation, a value of zero corresponds to no separation,

and negative values represent cases where a match profile is assigned a score lower than a non-match profile.

With the exception of Spearman rank correlation, all considered metrics assigned the highest similarity scores to the match profiles on all 32 microarrays, independent of normalization choice. Not surprisingly, separation between interfamily profiles was greater than that between intrafamily profiles. In addition, changes of normalization and similarity metric had greater impact on intrafamily than interfamily separation. The best overall separation was determined by calculating the product of the means of the intra- and interfamily separations divided by the corresponding standard deviations. Sum normalization of the intensity vectors, quadratic normalization of the energy vectors and uncentered Pearson correlation as the similarity metric achieved the highest overall separation, producing a mean intrafamily separation of 0.69 (0.17 standard deviation) and a mean interfamily separation of 0.93 (0.08 standard deviation). Therefore, we settled on this combination of normalization and similarity metric parameters as our method of choice.

3.3.3 Significance estimation

Raw similarity scores, as described above, provide an effective means of ranking viral energy profiles based on similarity to an observed hybridization pattern. However, such ranking provides no explicit information regarding the likelihood that viruses corresponding to the best scoring profiles are actually present in a sample under investigation. For example, two profiles may have identical high scores, yet one of the

scores may reflect a true positive while the other may be the result of over-representation of cross-hybridizing oligonucleotides in a profile.

To facilitate the interpretation of individual raw similarity scores, we sought to develop a test of their statistical significance. For this purpose, we obtained empirical distributions of the scores for every virus profile in the energy matrix. The distributions were based on 1009 independent microarray experiments collected from a wide range of clinical and non-clinical samples representing different tissues, cell types, and nucleic acid complexities. Given such sample diversity, we assumed that any given virus was present in only a small fraction of all samples. Therefore, the empirical distributions are essentially distributions of true negative scores. The \log_e -transformed similarity scores were approximately normally distributed. Outliers on the right tails of the distributions, assumed to be true positives, were removed (see Materials and Methods), and parameters of the null distributions were estimated as the mean and standard deviation of the remaining observations. These parameters were used to calculate the probability associated with any observed similarity score. Probabilities obtained this way should be interpreted as one-tail p-values for the null hypothesis, that the virus represented by the profile is not present in the sample.

As shown in Figure 3.3, the most significant similarity scores for all 32 microarrays in the training dataset were correctly matched to the virus known to be present in the input sample: HPV18 for HeLa samples, RSV for RSV-positive samples, and FluA for FluA-positive samples. Corresponding p-values ranged between 8.7×10^{-3} and 7.7×10^{-7} (median

= 2.1×10^{-5}), 4.0×10^{-4} and 1.4×10^{-8} (median = 5.1×10^{-8}), and 1.8×10^{-6} and 1.4×10^{-7} (median = 4.7×10^{-7}) respectively (red circles in Figure 3.3). Energy profiles of unrelated viruses from six representative families (black circles) as well as profiles of divergent members belonging to the same families as the match viruses (blue circles) had similarity scores of essentially background significance (p-values > 0.14). Even p-values of the most closely related intrafamily virus profiles (purple circles) were separated from those of the match viruses by more than 1.1 (HPV45), 2.1 (HMPV), and 3.4 (FluB) logs. Although the p-values obtained for these profiles are more significant than background, their similarity scores are entirely based on oligonucleotides that also belong to the match virus profiles. P-values resulting from such profile overlaps can be easily recognized and masked if desired (see *Example 3* below).

3.3.4 Examples

Our laboratory is conducting a series of studies focused on human diseases suspected to have viral etiologies. The E-Predict algorithm was developed to assist in the analysis of samples obtained as part of these investigations. As an illustration of its versatility, we present four examples of E-Predict as used in our laboratory.

Example 1. In this example, E-predict was used to interpret a hybridization pattern complicated by a low signal-to-noise ratio (Tables 3.3 and 3.4). The microarray result was obtained as part of our ongoing study of viral agents associated with acute hepatitis. Total nucleic acid from a serum sample was amplified, labeled, and hybridized to the microarray using our standard protocol (see Materials and Methods). Despite the fact that

very few oligonucleotides had intensity higher than background (Table 3.4), E-Predict assigned highly significant scores to Hepatitis B virus ($p = 0.002$) and several closely related hepadnaviruses (Table 3.3). Specifically, no hepadnavirus oligonucleotide had intensity greater than 500 (for reference, background intensities are around 100, and the possible range is between 0 and 65,536). PCR with Hepatitis B specific primers confirmed the presence of the virus in the sample. Complete E-Predict output for this example is available as Additional File 1. The microarray data have been submitted to the NCBI GEO database [46] (accession GSE2228).

Example 2. In this example, E-Predict was used to identify the presence of two distinct viral species in the same sample (Table 3.5). The microarray result was obtained from a nasopharyngeal aspirate sample, which was collected as part of our ongoing investigation of childhood respiratory tract infections. On this microarray, E-Predict assigned highest significance to two unrelated viruses, Influenza A virus ($p < 10^{-6}$) and RSV ($p = 0.008$), suggesting a double infection. The sample was independently confirmed to contain Influenza A and RSV, by DFA and specific PCR respectively. Complete E-Predict output for this example is available as Additional File 2. The microarray data have been submitted to the NCBI GEO database [46] (accession GSE2228).

Example 3. This example illustrates the ability of E-Predict to identify a virus that was not included in the microarray design. Table 3.6 shows E-Predict results for a microarray used to identify a novel coronavirus (SARS CoV) during the 2003 outbreak of Severe Acute Respiratory Syndrome as reported previously [8] [19]. Since our microarray was

designed prior to 2003, it did not contain oligonucleotides derived from the SARS CoV genome. However, after the entire genome sequence of the virus became available [13], its theoretical energy profile was added to the E-Predict energy matrix. Reanalysis of the original SARS microarray data (NCBI GEO [46] accession GSM8528) using E-Predict revealed that the SARS CoV energy profile attained the highest similarity score and a highly significant p-value ($p = 1 \times 10^{-6}$), despite the fact that the microarray, and therefore the profile, did not contain any oligonucleotides derived from the SARS CoV genome.

In addition to the SARS CoV prediction mentioned above, several astrovirus and picornavirus profiles had similarity scores with significant p-values. However, these predictions were based on oligonucleotides corresponding to a conserved 3' UTR region shared by these viruses with the SARS CoV [8] [20]. To identify incorrect predictions, such as these, resulting from partial profile overlaps with a match virus, we implemented an iterative version of E-Predict, where oligonucleotide intensities corresponding to the top scoring profile from one iteration are set to zero before running the next iteration. As a consequence, misleading predictions resulting from oligonucleotides shared with the top scoring profile fail to attain significant similarity scores in subsequent iterations. Conversely, only those predictions that are based on alternative oligonucleotides, i.e. predictions representing distinct species, remain. When iterative E-Predict was used on the SARS microarray, no astrovirus or picornavirus profile attained a statistically significant score ($p > 0.04$) in the second iteration, effectively removing these profiles from consideration. Complete E-Predict output for this example is available as Additional File 3.

Example 4. This example illustrates the use of E-Predict to discriminate between closely related viral species such as human rhinovirus (HRV) serotypes (Figure 3.4).

Rhinoviruses are a genus in the picornavirus family, which also includes enterovirus, aphthovirus, cardiovirus, hepatovirus, and parechovirus genera. Partial sequence analysis [50] [51] [52] indicates that HRV serotypes can be divided into two major groups (A and B), with the exception of HRV87, which is more closely related to enteroviruses. Only 2 complete rhinovirus reference genomes are available, one for each group: HRV89 (group A) and HRV14 (group B). Energy profiles of both viruses are included in our energy profile matrix as well as profiles of several enteroviruses and other more distant members of the picornavirus family. RNA samples from cultures of 22 representative serotypes were individually hybridized to the microarray, and the results were analyzed by E-Predict. In the absence of complete genome sequence data and corresponding energy profiles for each of the 22 serotypes, the E-Predict results revealed whether a particular serotype was most similar to HRV89, HRV14 or one of the enterovirus genomes in the energy matrix. To further refine our analysis, we clustered the E-Predict similarity scores from all 22 microarrays across all picornavirus profiles (Figure 3.4a). The resulting cluster dendrogram of the serotypes had striking similarity to a phylogenetic tree based on nucleotide sequences of VP1 capsid protein (Figure 3.4b and [50]). Serotypes 4, 26, 27, 70, and 83 were correctly grouped together on the basis of their similarity to the profile of HRV14 (group B); HRV87 formed a separate node, and the remaining serotypes were grouped together on the basis of their similarity to the profile of HRV89 (group A). Complete E-Predict output for this example is available as Additional File 4.

The microarray data have been submitted to the NCBI GEO database [46] (accession GSE2228).

3.4 Discussion

Identifying individual species present in a complex environmental or clinical sample is an essential component of many current and proposed metagenomic applications. Given a foundation of genomic sequence information, DNA microarrays are a high-throughput and cost-effective methodology for detecting species in an unbiased and highly parallel manner. Metagenomic applications employing DNA microarrays include characterization of microbial communities from environmental samples such as soil and water [23] [38], pathogen detection in clinical specimens and field isolates [37], monitoring of bacterial contamination of food and water [43], and detection of agents involved in potential cases of bioterrorism [45].

Despite the increasing use of DNA microarrays for species detection and identification, bioinformatics tools for interpreting hybridization patterns associated with complex clinical and environmental samples are lacking. Existing methods have utilized direct visual inspection of hybridizing oligonucleotides [8] [53] or inspection following clustering [7] [54]. Such methods are intractable for interpreting complex hybridization patterns, are time-consuming, and suffer from user bias. Improved data interpretation tools must address several challenges. First, hybridization patterns may represent signal from dozens or even hundreds of species. Also, several closely related species may be present in a sample, giving rise to overlapping hybridization signals. A likely additional

source of noise is unanticipated cross-hybridization, since many of the genomes present in a complex sample may be uncharacterized. Finally, obtaining pure samples of each possible species for the purpose of generating reference hybridization patterns is impractical or impossible in most cases.

When challenged with each of these problems, E-Predict proved to be a useful tool for interpreting hybridization patterns, correctly identifying viruses from diverse viral families present in a variety of clinical samples. In particular, E-Predict does not rely on the use of empirically generated reference hybridization patterns, since species identification is based instead on theoretical hybridization energy profiles. The energy profile matrix currently represents over 1200 distinct viruses whose complete genomic sequences are known. As new viral genomes are sequenced, profiles are added to the matrix to broaden the range of species detection. For example, addition of the SARS CoV profile enabled accurate identification of the virus, even though no oligonucleotides derived from its genome were present on the microarray. Conversely, even when a perfectly matching profile is not available due to limited sequence coverage, E-Predict will identify the closest related species, as long as such species are represented on the microarray. This feature is particularly useful for detection of novel viruses as well as for discrimination between closely related viruses such as HRV serotypes. Naturally, maximum range and precision of detection is achieved through addition of new profiles and periodic microarray updates to include specific oligonucleotides from newly sequenced species.

E-Predict is also useful in overcoming problems related to nucleic acid complexity frequently encountered in clinical samples. For example, E-Predict correctly identified Hepatitis B virus in a serum sample, despite the fact that the hybridization pattern was complicated by a low signal-to-noise ratio. In another example, E-Predict deconvoluted a complex hybridization pattern, correctly suggesting the presence of two viruses (FluA and RSV) in a nasopharyngeal aspirate sample. In yet another example, iterative application of E-Predict (see Materials and Methods) to a hybridization pattern involving oligonucleotides derived from seemingly unrelated families (coronaviridae and astroviridae) allowed an objective recognition that the pattern represented the presence of only one virus (SARS CoV).

Using a training dataset of 32 microarrays derived from samples known to contain specific viral species, we identified a set of normalization and similarity metric parameters, which yielded the best discrimination between true positive and true negative species predictions. The combination of sum normalization of the intensity vectors, quadratic normalization of the energy vectors, and uncentered Pearson correlation as the similarity metric was the optimal choice for our data. However, a different set of parameters may be required for applications that use a different nucleic acid amplification or detection strategy. An independent evaluation of potentially useful normalization and similarity metric parameters is therefore recommended for each specific application of the algorithm.

Using our best combination of normalization and similarity metric parameters, we obtained a set of null distributions representing true negative scores. These distributions were based on over 1000 independent hybridizations and the assumption that the majority of samples were negative for the presence of any given virus. Although valid for our data, this assumption will not hold for all cases. For example, in applications concerned with bacterial species detection, some species may be present in most or even all samples and others encountered only rarely. In this case, a more complicated model will be required to assess whether a specific distribution represents negative, positive, or both negative and positive scores. For example, in cases where distributions appear bimodal, one mode may represent true negatives and the other true positives. In some cases targeted experimental verification of a subset of representative scores may be necessary. If both positive and negative score distributions are available, p-values can be calculated for each distribution.

Several modifications to the algorithm may potentially result in improved prediction accuracy. First, in the current implementation, oligonucleotides exhibiting non-specific cross-hybridization are filtered, and the remaining oligonucleotides are weighted equally. Since oligonucleotides exhibit a continuous range of non-specific hybridization [40] [49], a more sophisticated system of oligonucleotide weights may result in a better performance. For example, using a procedure similar to that used to generate null distributions for the virus profile scores, empirical distributions can be obtained for individual oligonucleotide intensities, and individual oligonucleotide contributions may be weighted by the probabilities associated with the corresponding observed intensities. Such weighting may allow a more accurate assessment of significance.

Second, no attempt was made to normalize nucleic acid abundances of individual species, which may vary widely in different samples depending on factors such target-to-background ratio, number of species present, and efficiency of nucleic acid extraction and amplification. While individual nucleic acid abundances are difficult or impossible to estimate in most metagenomic applications, particularly before the corresponding species have been identified, in applications where such estimates can be made, either experimentally or on theoretical grounds, use of correction factors for calculating similarity scores or stratification of p-value estimation may be needed. In addition, for highly abundant species, care should be taken to avoid saturation of individual oligonucleotides, as E-predict performance drops sharply after 20-25% of oligonucleotides in a given profile are saturated (data not shown).

Third, even though viral genomes were used as the basis for calculating energy profiles, the concept can be easily extended to other taxonomy nodes such as genera or families of viruses. This requires every sequence element to be classified at the appropriate node in the taxonomy hierarchy.

Finally, iterative use of E-Predict was intended for identification of multiple species that may be present in a sample. In this setting, it is important to distinguish between true predictions representing unique species present in the sample and misleading predictions arising from partially overlapping profiles. In each iteration, it is assumed that the profile attaining the highest score corresponds to the species most likely to be among those

present in the sample. When a novel species is present, this assumption may not hold due to limited oligonucleotide coverage. For instance, in the SARS CoV example, although SARS CoV attained a higher similarity score than mink astrovirus (MAV), the corresponding p-values were comparable. However, even if MAV were the top prediction in the first iteration, SARS CoV would be the top prediction in the second iteration ($p = 2 \times 10^{-6}$, data not shown) and therefore would not be missed as a true positive. In our current studies, p-values in all iterations are estimated using the same set of null probability distributions. In addition, we use two iterations as our default, and essentially never need to run more than three iterations, as detection of more than two or three viruses is rare. However, iterative resolution of hundreds or thousands of species present in a sample may necessitate other normalization methods or adjustments to the null distributions for p-value estimation. As an alternative, non-iterative algorithms for analyzing overlapping profile signatures are also being explored.

In conclusion, E-Predict is a novel computational approach for species identification, which is generally applicable to a wide range of metagenomic applications using DNA microarrays. In particular, as more sequencing efforts are being directed at natural microbial communities, DNA microarrays are bound to become a central tool for various downstream applications such as identification of microbial species or detection of genes and biochemical pathways in such communities. E-Predict addresses an acute need for computational tools capable of interpreting the highly complex microarray data obtained through such studies. E-Predict was developed for viral species identification and therefore has immediate implications for medical diagnostics and viral discovery. In

addition, the concept of theoretical energy profiles can be extended to represent other microorganisms, particular genes, or biochemical pathways.

3.5 Materials and Methods

3.5.1 Sample preparation and hybridization to microarrays

All patient samples were collected according to protocols approved by the UCSF Committee on Human Research.

HeLa samples. HeLa cells were grown to confluence in a T150 tissue culture flask in DMEM supplemented with 10% FBS and antibiotics. The cells were harvested by adding 10 mL of Trizol reagent (Invitrogen), and total RNA was isolated according to the manufacturer's protocol. 50 ng of HeLa total RNA were used for each amplification and hybridization.

Pediatric respiratory samples. Frozen nasopharyngeal aspirate samples were thawed, and 200 μ L aliquots were used to extract RNA using RNeasy Mini Kit (Qiagen) as follows. 750 μ L of RLT buffer containing 1% 2-Mercaptoethanol were added to each sample and mixed. 1 mL of 100% ethanol was added next, and the resulting mixture was applied to the columns in three 650 μ L aliquots. The remaining steps were carried out according to the manufacturer's protocol including on-column DNase digest. RNA was eluted from the columns with 30 μ L of nuclease-free water, and 9 μ L were used for amplification and hybridization.

Hepatitis sample. Frozen serum sample was thawed, and 150 μ L aliquot was used to extract total nucleic acid using MagNA Pure LC Total Nucleic Acid Isolation Kit (Roche) according to the manufacturer's protocol. RNA was eluted in 50 μ L of nuclease-free water, and 9 μ L were used for amplification and hybridization.

HRV serotypes. Frozen samples of low passage viral culture supernatants were thawed on ice and pre-filtered with a 0.2 μ m syringe filter. 200 μ l aliquots of the pre-filtered supernatants were treated with 600 U of micrococcal nuclease (Fermentas) in the presence of 10 mM CaCl_2 for 3 hours at 37 $^\circ\text{C}$. RNA was then extracted using Trizol reagent (Invitrogen) according to the manufacturer's protocol. 20 μ g of linearized polyacrylamide (Ambion) were used as the carrier during the 2-propanol precipitation. RNA was resuspended in 30 μ l of nuclease-free water, and 9 μ l were used for amplification and hybridization.

Microarrays used in this study were essentially identical to those previously described [8]. Detailed description of the microarray platform, including oligonucleotide sequences, can be found in the NCBI GEO database [46] (accession GPL 1834). Briefly, 70-mer oligonucleotides representing the most conserved viral genomic elements were selected as 70-mers having sequence similarity (determined by nucleotide alignment) to the highest number of viral genomes [7]. Oligonucleotides were resuspended in 3X SSC at 50 μ M concentration and spotted onto poly-lysine coated glass slides [55]. Each spot on the microarray also contained a unique "alien" sequence 70-mer (Spike70: 5'-ACC TCG CTA ACC TCT GTA TTG CTT GCC GGA CGC GAG ACA AAC CTG AAC ATT

GAG AGT CAC CCT CGT TGT T-3') spotted at a 1:50 ratio with the viral oligonucleotide to facilitate gridding of the microarrays (see below).

RNA extracted from the samples was amplified using a modified Round A-B random PCR method [56] as previously described (Protocol S1 in [8]). Briefly, random-primed reverse transcription and second strand synthesis were carried out using primer A (5'-GTT TCC CAG TCA CGA TCN NNN NNN NN-3'). The resulting material was then amplified with 40 cycles of PCR using primer B (5'-GTT TCC CAG TCA CGA TC-3'). This was followed with additional 20 cycles of PCR with primer B to incorporate aminoallyl-dUTP. The amplified material was then labeled with Cy5, and 0.1 – 1.0 pmol of Probe70 (an oligonucleotide complementary to Spike70 containing five amino-modified bases for dye coupling: 5'-AAC AAC GAG GG[AmC6-dT] GAC TCT CAA [AmC6-dT]GT TCA GGT TTG TC[AmC6-dT] CGC GTC CGG CAA GCA A[AmC6-dT]A CAG AGG T[AmC6-dT]A GCG AGG T-3', Operon) was labeled with Cy3. The Cy5 and Cy3 probes were pooled and hybridized to the microarray in 3X SSC at 65 °C overnight [55]. The Cy3 channel was used to facilitate gridding, but otherwise was ignored in the data analysis. Microarrays were scanned with an Axon 4000B scanner (Axon Instruments) and gridded using the bundled GenePix 3.0 software.

Microarray data have been submitted to the NCBI GEO database [46] (accession GSE 2228). The SARS microarray data are also available in NCBI GEO (accession GSM8528) as previously reported [8].

3.5.2 Training dataset

Fifteen HeLa microarrays were chosen randomly from a set of 43 HeLa hybridizations having at least five papillomavirus oligonucleotides with sum-normalized intensities greater or equal to 0.005. Ten RSV microarrays were chosen randomly from a set of 22 clinical hybridizations having at least five paramyxovirus oligonucleotides with sum-normalized intensities greater than or equal to 0.005 and confirmed to be RSV-positive by DFA. Seven FluA microarrays were chosen from eight available clinical hybridizations having at least five orthomyxovirus oligonucleotides with sum-normalized intensities greater than or equal to 0.005 and confirmed to be FluA-positive by DFA. The eighth FluA microarray was excluded, because it was also positive for RSV by visual inspection.

3.5.3 Theoretical energy profiles

The energy profile matrix used in this study included all NCBI reference viral genomes (1229) available as of July 2004 [57]. Nucleotide BLAST (*blastall* version 2.2.8 [58] with the default settings) was used to align microarray oligonucleotides with the viral genomes. Energies of hybridization were computed from the alignments using *energy* program distributed with ArrayOligoSelector [49] [59]. In cases where an oligonucleotide had multiple alignments to the same genome, energy calculations were based on the highest scoring alignment. Energy profile matrix is available as Additional File 5.

3.5.4 Similarity scores

Control oligonucleotides and oligonucleotides known to result in non-specific hybridization were removed from consideration by setting their intensities and energies to

zero. The list of these oligonucleotides (Additional File 6) was obtained by including 129 oligonucleotides with unnormalized median intensity greater than 500, calculated from 1009 independent hybridizations described below. The list also included 137 oligonucleotides obtained by clustering of distributions of sum-normalized intensity, based on the same set of 1009 hybridizations, and visual identification of an outlier cluster with median sum-normalized intensities significantly higher than those observed for most oligonucleotides. Energy vectors were further filtered to exclude terms with energy predictions higher than -30 kcal/mol (again by setting their values to zero), as such predictions on our platform do not correspond to detectable array intensities [49]. A profile was considered only if it had at least three oligonucleotides with non-zero energy predictions. The resulting intensity and energy vectors were normalized using appropriate normalization methods (N, S, Q, U). Similarity scores were computed using an appropriate similarity metric (DP, PC, UP, SR, ED).

3.5.5 Probability estimation

Null distributions of similarity scores were obtained using a set of 1009 hybridizations, which included all hybridizations performed on our platform to date. Similarity scores were calculated as described above using uncentered Pearson correlation as the similarity metric and sum and quadratic normalizations for intensity and energy vectors respectively. Scores were log-transformed. Right tail outliers corresponding to positive cases were excluded by iterative trimming of the top scores in 1% increments until the best normality fit was obtained, as judged by the Shapiro-Wilk normality test [60] (implemented in R [61]). Trimming was allowed to involve 0 to 25% of all scores. Over one third of virus profiles required no trimming at all. Only a small number of profiles

(34) required trimming beyond 10%, all of which corresponded to viruses frequently present in our samples. No profile required trimming of more than 17% of the scores. The resulting trimmed distributions were assumed to be normal, and their parameters were estimated as the mean and standard deviation of the included scores (Additional File 7). Obtained parameters were used to estimate significance of individual scores as probabilities associated with observing values equal or greater than the scores. For this purpose, only profiles with at least three oligonucleotides with raw intensity greater than 100 ($\sim 2 - 4 \times$ background) were considered.

3.5.6 Iterative E-Predict

First iteration was carried out as described above. For each additional iteration, oligonucleotide intensities of the profile attaining the highest similarity score in the previous iteration were set to zero. The resulting intensity vector was normalized, and similarity scores and p-values were calculated using the same normalization method, similarity metric, and null distributions as in the initial iteration.

3.5.7 Clustering of HRV serotypes

Similarity scores were calculated as described above using uncentered Pearson correlation as the similarity metric and sum and quadratic normalizations for intensity and energy vectors respectively. Scores corresponding to picornavirus profiles were clustered using Cluster (version 2.0) [10] [62] by hierarchical average linkage clustering with Pearson correlation as the similarity metric. Cluster images were obtained using Java TreeView (version 1.0.8) [63] [64].

The phylogenetic tree based on nucleotide sequences of VP1 capsid protein was constructed using data from [50]. Sequence alignment of relevant serotypes and the resulting tree were obtained using ClustalX (version 1.81 for Windows [65] [66]) with the default settings.

3.5.8 PCR

The presence of Hepatitis B virus in the hepatitis sample was confirmed using primers Hep_1F (5'-GAC TCG TGG TGG ACT TCT CTC AA-3') and Hep_4R (5'- GAA AGC CCT GCG AAC CAC TGA A-3') with amplified cDNA (Round B material; see [7] for amplification details) as the template. The presence of RSV in the FluA/RSV double-infected sample was confirmed by PCR using primers AU_041 (5'-GAT GAA AAA TTA AGT GAA ATA TTA GG-3') and AU_042 (5'-GTT CAC GTA TGT TTC CAT ATT TG-3') with cDNA (Round A material; see [7] for amplification details) as the template. In both cases, amplified PCR fragments were sequenced and had at least 99% nucleotide identity to the genomes of Hepatitis B virus [GenBank:NC_003977] and RSV [GenBank:NC_001803].

3.5.9 E-Predict software

The E-Predict software can be obtained at <http://derisilab.ucsf.edu/epredict/>.

3.6 Additional Files

The following supporting data are available as a web supplement to the Genome Biology article describing this work (<http://genomebiology.com/2005/6/9/R78>):

Additional File 1: Text file of E-Predict output for the hepatitis example (*Example 1*).

Additional File 2: Text file of E-Predict output for the FluA/RSV double-infection example (*Example 2*).

Additional File 3: Text file of E-Predict output for the SARS CoV example (*Example 3*).

Additional File 4: Text file of E-Predict output for the HRV serotypes example (*Example 4*).

Additional File 5: Tab delimited text file containing the energy profile matrix.

Additional File 6: Text file containing the list of non-specific oligonucleotides ignored during E-Predict.

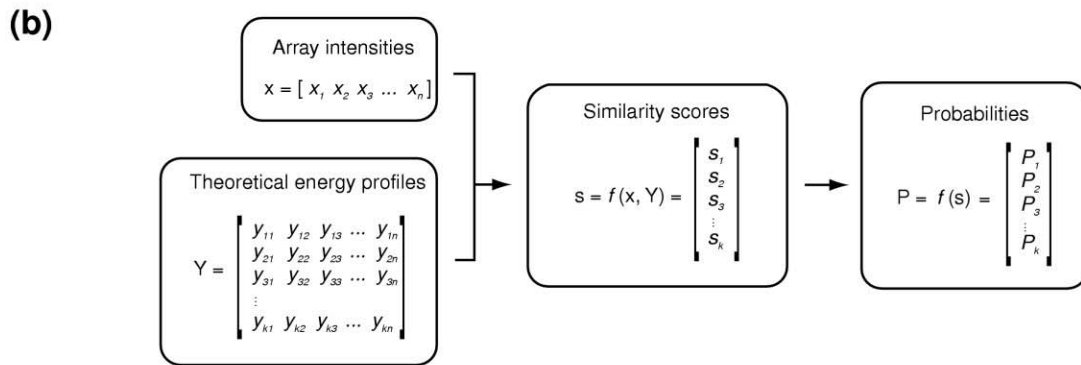
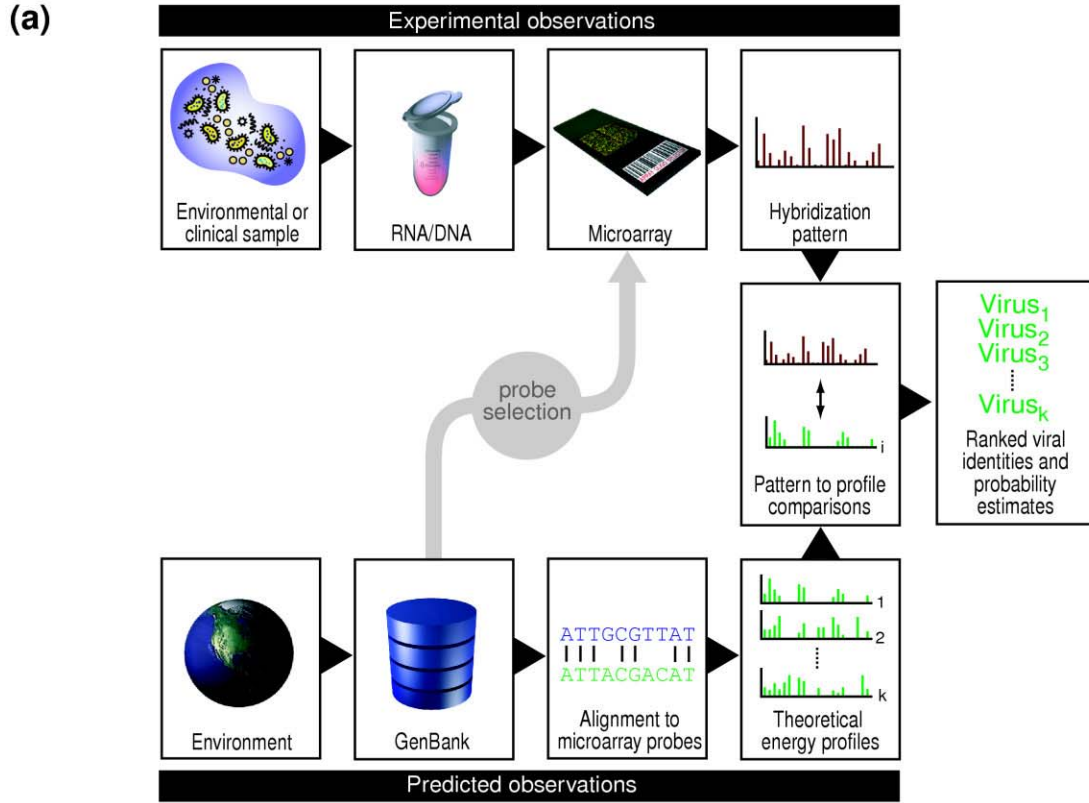
Additional File 7: Tab delimited text file containing the list of profile parameters.

Additional File 8: Text file of E-Predict output used to evaluate normalization and similarity metric parameters.

3.7 Acknowledgements

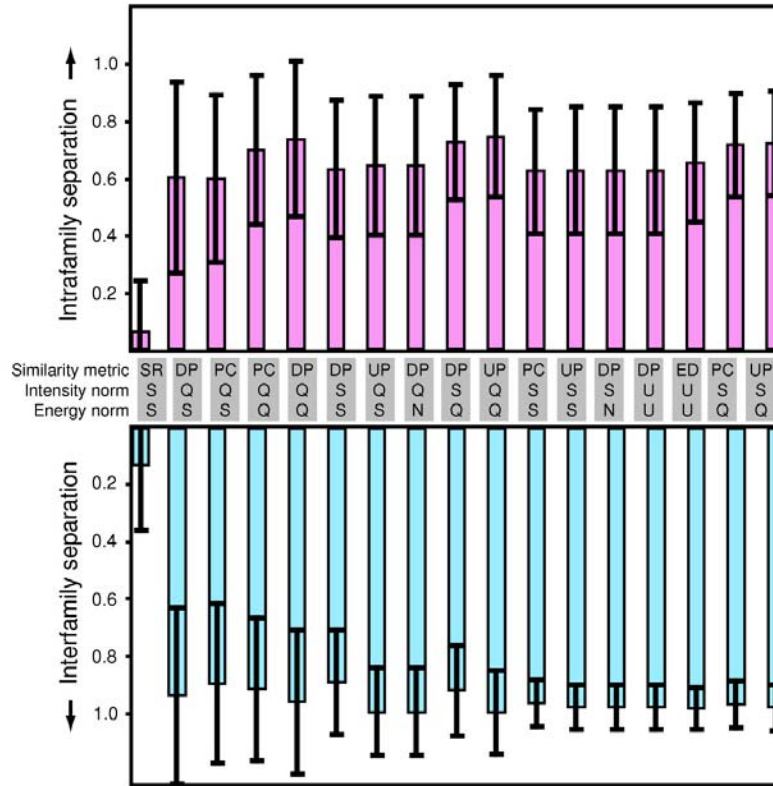
We thank Dr. Hao Li, Christina Chaivorapol, and Amir Najmi for helpful discussions. We thank Dr. Yu-Tsueng Liu for performing microarray hybridization and PCR follow-up of the hepatitis sample. Hepatitis sample was graciously provided as part of an ongoing study by Dr. Tim Davern (UCSF). Pediatric respiratory samples were graciously provided as part of an ongoing study by Dr. Tara Greenhow, Dr. Peggy Weintrub, Dr. Lawrence Drew, and Carolyn Wright (UCSF). Cultures of HRV serotypes were graciously provided as part of an ongoing study by Dr. David Schnurr and Dr. Shigeo Yagi (California Viral and Rickettsial Disease Laboratory, Richmond, CA). This work was supported by a Genentech Graduate Fellowship (AU) and grants from the Sandler Program for Asthma Research, and Doris Duke Charitable Foundation (JDR).

Figure 3.1: E-Predict Algorithm



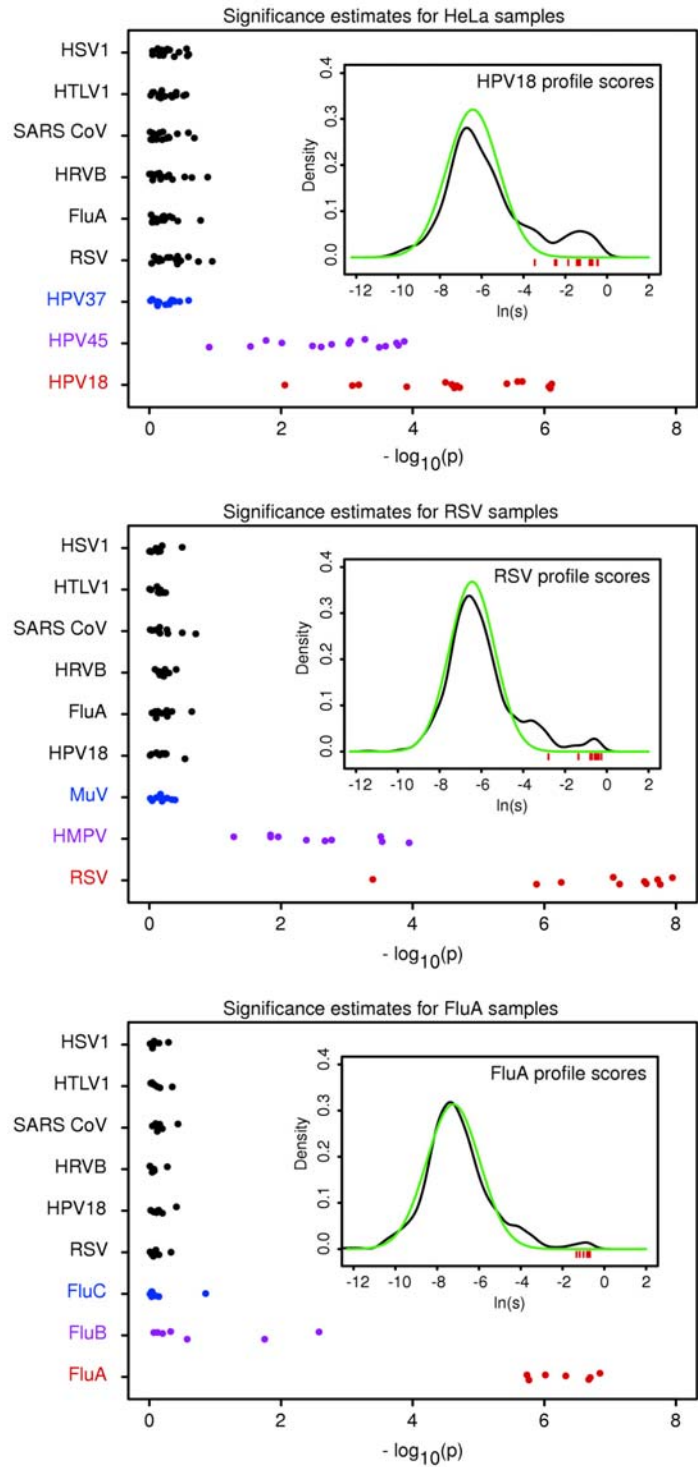
(a) Nucleic acid from an environmental or clinical sample is labeled and hybridized to a species detection microarray. The resulting hybridization pattern is compared with a set of theoretical hybridization energy profiles computed for every species of interest. Energy profiles attaining statistically significant comparison scores suggest the presence of the corresponding species in the sample. **(b)** Observed hybridization intensities are represented by a row vector x , where each intensity value corresponds to an oligonucleotide on the microarray. Theoretical hybridization energy profiles form a matrix of energy values, Y , where each row represents a profile, and each column corresponds to an oligonucleotide in x . A suitable similarity metric function compares x with each row of Y to produce a column vector of similarity scores, s . Statistical significance of the individual scores in s is estimated to produce the output column vector of probabilities, P , where each probability value corresponds to a profile in Y .

Figure 3.2: Evaluation of Normalization and Similarity Metric Parameters



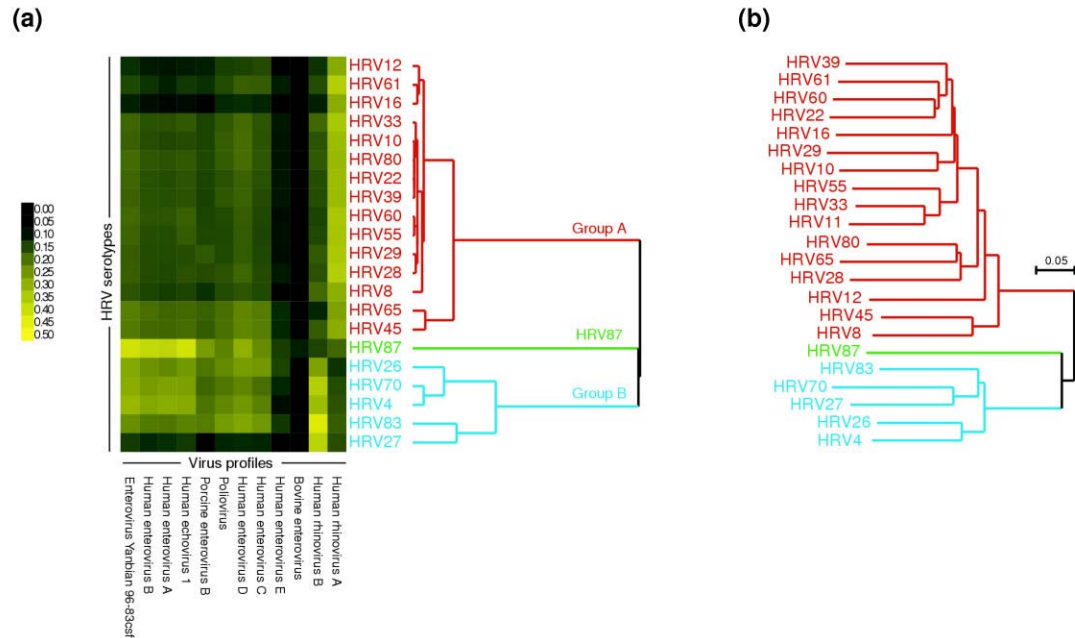
A training set of 32 microarrays was used to evaluate all non-equivalent combinations of intensity and energy vector normalization (N, none; Q, quadratic; S, sum; U, unit-vector) and similarity metric (DP, dot product; ED, similarity based on Euclidean distance; PC, Pearson correlation; SR, Spearman rank correlation; UP, uncentered Pearson correlation) parameters. For each combination of parameters, intra- and interfamily separations were calculated for each microarray as the score of the virus profile matching the virus present in the sample minus the score of the best-scoring non-match profile from the same or a different virus family (top and bottom panels, respectively), normalized by the range of all scores on that microarray. Bars represent the mean, and error bars represent plus and minus standard deviation of separation values from all microarrays. The best performing combinations are shown in the order of increasing performance (calculated as the product of the intra- and interfamily separation means divided by the corresponding standard deviations).

Figure 3.3: Estimation of Significance of Individual Similarity Scores



Probabilities associated with the similarity scores of nine representative virus profiles obtained for the 15 HeLa, 10 RSV, and 7 FluA microarrays from the training dataset are shown in the top, center, and bottom panels respectively. Each circle represents one microarray, and vertical “jitter” is used to resolve individual circles. Probabilities for virus profiles from seven diverse virus families are included with each microarray set: Herpes simplex virus 1 (HSV1), Human T-lymphotropic virus 1 (HTLV1), SARS coronavirus (SARS CoV), Human rhinovirus B (HRVB), Influenza virus A (FluA), Human respiratory syncytial virus (RSV), and Human papillomavirus type 18 (HPV18). Red circles represent match and black circles non-match interfamily profiles. Two intrafamily non-match profiles are also included and are different for the three microarray sets. The most closely related intrafamily profiles are represented by purple circles: Human papillomavirus type 45 (HPV45), Human metapneumovirus (HMPV), and Influenza B virus (FluB). More distant intrafamily profiles are shown in blue: Human papillomavirus type 37 (HPV37), Mumps virus (MuV), and Influenza C virus (FluC). The inset in each panel shows a normalized histogram (density) of the empirical distribution of log-transformed similarity scores for a match profile (black curve) and the corresponding normal fit representing true negative scores (green curve). Inset red bars depict observed log-transformed similarity scores corresponding to the match profile probabilities (red circles).

Figure 3.4: HRV Serotype Discrimination Using E-Predict Similarity Scores



(a) Culture samples of 22 distinct HRV serotypes were separately hybridized to the microarray. E-Predict similarity scores were obtained for all virus profiles in the energy matrix and clustered using average linkage hierarchical clustering and Pearson correlation as the similarity metric. Virus profiles for which similarity scores could be calculated in all 22 experiments were included in the clustering. Both microarrays (rows) and virus profiles (columns) were clustered. **(b)** Published nucleotide sequences of VP1 capsid protein from the 22 HRV serotypes were aligned using ClustalX. Phylogenetic tree based on the resulting alignment is shown.

Table 3.1: Normalization Methods

Normalization	Formula	Abbreviation
None	$x_{i_{norm}} = x_i$	N
Sum	$x_{i_{norm}} = \frac{x_i}{\sum x_i}$	S
Quadratic	$x_{i_{norm}} = \frac{x_i^2}{\sum x_i^2}$	Q
Unit-vector	$x_{i_{norm}} = \frac{x_i}{\sqrt{\sum x_i^2}}$	U

Table 3.2: Similarity Metrics

Similarity metric	Formula	Abbreviation
Dot product	$s(\mathbf{x}, \mathbf{y}) = \sum x_i y_i$	DP
Pearson correlation	$s(\mathbf{x}, \mathbf{y}) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$	PC
Uncentered Pearson correlation	$s(\mathbf{x}, \mathbf{y}) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$	UP
Spearman rank correlation	$s(\mathbf{x}, \mathbf{y}) = \frac{\sum (R_{x_i} - \bar{R}_x)(R_{y_i} - \bar{R}_y)}{\sqrt{\sum (R_{x_i} - \bar{R}_x)^2 \sum (R_{y_i} - \bar{R}_y)^2}}$	SR
Similarity based on Euclidean distance	$s(\mathbf{x}, \mathbf{y}) = 2 - \sqrt{\sum (x_i - y_i)^2}$	ED

Table 3.3: Example 1: Hepatitis Microarray – Predicted Virus Profiles

Taxonomy ID	Virus profile	Virus family	Similarity score	Probability
10407	Hepatitis B virus	Hepadnaviridae	0.145209	0.002451*
113194	Orangutan hepadnavirus	Hepadnaviridae	0.143754	0.002482*
68416	Woolly monkey hepatitis B Virus	Hepadnaviridae	0.123794	0.003111*
35269	Woodchuck hepatitis B virus	Hepadnaviridae	0.106576	0.002896*
41952	Arctic ground squirrel hepatitis B virus	Hepadnaviridae	0.098908	0.003555*
10406	Ground squirrel hepatitis virus	Hepadnaviridae	0.093975	0.003475*
10372	Human herpesvirus 7	Herpesviridae	0.027847	0.115068

All virus profiles, for which a score could be calculated (see Materials and Methods), are shown sorted by the similarity score. *Statistically significant probabilities ($p < 0.01$).

Table 3.4: Example 1: Hepatitis Microarray – Oligonucleotides Contributing to the Hepatitis B Virus Profile Prediction

Oligonucleotide	Parental virus genome	Virus family	Raw intensity	Raw energy
21326584_16	Hepatitis B virus	Hepadnaviridae	403	102.9
9628700_11_rc	Hepatitis B virus	Hepadnaviridae	316	102.9
9634216_16	Orangutan hepadnavirus	Hepadnaviridae	357	96.6
21326584_25	Hepatitis B virus	Hepadnaviridae	262	109.6
9634216_11_rc	Orangutan hepadnavirus	Hepadnaviridae	308	99.1
9634216_11	Orangutan hepadnavirus	Hepadnaviridae	288	99.1
9630370_16	Woolly monkey hepatitis B Virus	Hepadnaviridae	464	72.2
9628700_20_rc	Hepatitis B virus	Hepadnaviridae	160	120
21326584_9	Hepatitis B virus	Hepadnaviridae	175	104.7
9628700_4	Hepatitis B virus	Hepadnaviridae	153	104.7

Ten oligonucleotides contributing most to the Hepatitis B virus similarity score are shown sorted by their relative contribution (product of normalized intensity and normalized energy values).

Table 3.5: Example 2: FluA, RSV Double Infection

Taxonomy ID	Virus profile	Virus family	Similarity score	Probability
11320	Influenza A virus	Orthomyxoviridae	0.504133	0.000000*
183764	Influenza A virus	Orthomyxoviridae	0.486601	0.000000*
130760	Influenza A virus	Orthomyxoviridae	0.105047	0.000151*
11250	Human respiratory syncytial virus	Paramyxoviridae	0.033523	0.007895*
12814	Respiratory syncytial virus	Paramyxoviridae	0.022144	0.007512*
11246	Bovine respiratory Syncytial virus	Paramyxoviridae	0.009983	0.029254
162145	Human metapneumovirus	Paramyxoviridae	0.001604	0.467995

All virus profiles, for which a score could be calculated (see methods), are shown sorted by the similarity score. *Statistically significant probabilities ($p < 0.01$).

Table 3.6: Example 3: SARS Microarray

Taxonomy ID	Virus profile	Virus family	Similarity score	Probability
Iteration 1				
227859	SARS coronavirus	Coronaviridae	0.415354	0.000001*
219688	Mink astrovirus	Astroviridae	0.335302	0.000000*
70793	Turkey astrovirus	Astroviridae	0.217455	0.000000*
11120	Avian infectious bronchitis virus	Coronaviridae	0.175788	0.000004*
70794	Ovine astrovirus	Astroviridae	0.153207	0.000031*
107033	Avian nephritis virus	Astroviridae	0.057325	0.000020*
47001	Equine rhinitis B virus	Picornaviridae	0.048009	0.000054*
12702	Human astrovirus	Astroviridae	0.044928	0.002118*
11852	Simian type D virus 1	Retroviridae	0.034479	0.016202
31631	Human coronavirus OC43	Coronaviridae	0.029834	0.002178
Iteration 2				
11852	Simian type D virus 1	Retroviridae	0.053705	0.007108*
39068	Mason-Pfizer monkey virus	Retroviridae	0.031347	0.026931
10359	Human herpesvirus 5	Herpesviridae	0.024634	0.167435
147712	Human rhinovirus B	Picornaviridae	0.022551	0.048232
208177	Tomato leaf curl Vietnam virus	Geminiviridae	0.022090	0.149573
85752	Tomato yellow leaf curl Thailand virus	Geminiviridae	0.021844	0.080110
223334	Tobacco leaf curl Kochi virus	Geminiviridae	0.021469	0.108687
188763	Chimpanzee cytomegalovirus	Herpesviridae	0.021088	0.132918
32610	Tomato geminivirus	Geminiviridae	0.021055	0.081960
83839	Pepper leaf curl virus	Geminiviridae	0.020882	0.082562

For each iteration, ten profiles with the highest similarity scores are shown sorted by score. *Statistically significant probabilities ($p < 0.01$).

Chapter 4: XMRV

Citation: Urisman A, Molinaro RJ, Fischer N, Plummer SJ, Casey G, Klein EA, Malathi K, Magi-Galluzzi C, Tubbs RR, Ganem D, Silverman RH, Derisi JL. Identification of a Novel Gammaretrovirus in Prostate Tumors of Patients Homozygous for R462Q *RNASEL* Variant. *PLoS Pathog.* 2006 Mar;2(3):e25.

Copyright: © 2006 Anatoly Urisman et al.

4.1 Abstract

RNase L is an important effector of the innate antiviral response. Mutations or variants that impair function of RNase L, particularly R462Q, have been proposed as susceptibility factors for prostate cancer. Given the role of this gene in viral defense, we sought to explore the possibility that a viral infection might contribute to prostate cancer in individuals harboring the R462Q variant. A viral detection DNA microarray composed of oligonucleotides corresponding to the most conserved sequences of all known viruses identified the presence of gammaretroviral sequences in cDNA samples from 7 of 11 R462Q-homozygous (QQ) cases, and in 1 of 8 heterozygous (RQ) and homozygous wild-type (RR) cases. An expanded survey of 86 tumors by specific RT-PCR detected the virus in eight of 20 QQ cases (40%), compared to only one sample (1.5%) among 66 RQ and RR cases. The full-length viral genome was cloned and sequenced independently from three positive QQ cases. The virus, named XMRV, is closely related to xenotropic murine leukemia viruses (MuLVs), but its sequence is clearly distinct from all known members of this group. Comparison of *gag* and *pol* sequences from different tumor isolates suggested infection with the same virus in all cases, yet sequence variation was consistent with the infections being independently acquired. Analysis of prostate tissues from XMRV-positive cases by *in situ* hybridization and immunohistochemistry showed

that XMRV nucleic acid and protein can be detected in about 1% of stromal cells, predominantly fibroblasts and hematopoietic elements in regions adjacent to the carcinoma. These data provide the first demonstration that xenotropic MuLV-related viruses can produce an authentic human infection, and strongly implicate RNase L activity in the prevention or clearance of infection *in vivo*. These findings also raise questions about the possible relationship between exogenous infection and cancer development in genetically susceptible individuals.

4.2 Introduction

Type I interferons (IFNs) are rapidly mobilized in response to viral infection and trigger potent antiviral responses. One such response is the induction by IFN of a family of 2'5' oligoadenylate synthetases (OAS); upon activation by virally-encoded dsRNA, these enzymes produce 5'-phosphorylated 2'-5' linked oligoadenylates (2-5A) from ATP [67]. 2-5A, in turn, is an activator of ribonuclease RNase L [68], which degrades viral (and cellular) single stranded RNAs [69]. *In vivo* evidence for the antiviral role of the 2-5A system was provided by studies with RNase L^{-/-} mice, which have enhanced susceptibility to infections by the picornaviruses, encephalomyocarditis virus and Coxsackievirus B4 [70, 71]. Ultimately, sustained activation of RNase L triggers a mitochondrial pathway of apoptosis that eliminates virus-infected cells [72, 70, 73, 74]. Genetic lesions in RNase L impair this apoptotic response, which has raised interest in the possibility that such mutations might also contribute to malignancy [75].

In this context, several recent studies have linked germline mutations in RNase L to prostate cancer susceptibility [76-79]. Prostate cancer has a complex etiology influenced by androgens, diet, and other environmental and genetic factors [80]. While sporadic

prostate cancer displays an age-related increase in prevalence, familial prostate cancer kindreds often display early-onset disease. Such kindreds, defined by having more than three affected members per family, account for 43% of early onset cases (<55 years old) and 9% of all cases [81]. The genetics of hereditary prostate cancer (HPC) is complex, and several genes have been proposed as susceptibility factors in this syndrome. Interestingly, one of these, *HPCI*, is linked to *RNASEL* [76, 77]. Several germline mutations or variants in *HPCI/RNASEL* have been observed in hereditary prostate cancer [76-79] (reviewed in [82]), including a common (35% allelic frequency) missense variant of RNase L, in which a G to A transition at nucleotide position 1385 (G1385A) results in a glutamine instead of arginine at amino acid position 462 (R462Q). Remarkably, a large, controlled sib-pair study implicated the R462Q RNase L variant in up to 13% of unselected prostate cancer cases [77]. One copy of the mutated gene increased the risk of prostate cancer by about 50%, whereas individuals that were homozygous for the mutation had a two-fold increased risk of prostate cancer. The R462Q RNase L variant had a 3-fold decrease in catalytic activity compared to the wild-type enzyme [77, 75]. However, while several case-controlled genetic and epidemiologic studies support the involvement of *RNASEL* (and notably the R462Q variant) in prostate cancer etiology [76-79], others do not [83-85], suggesting that either population differences or environmental factors may modulate the impact of *RNASEL* on prostatic carcinogenesis.

While the antiapoptotic phenotype of RNase L deficiency has dominated previous discussions of its possible linkage to cancer, RNase L is also a key effector of the antiviral action of interferons. This led us to consider the possibility that the putative

linkage of RNase L alterations to HPC might reflect enhanced susceptibility to a viral agent. To test this hypothesis, we have examined RNA derived from wild-type and RNase L variant (R462Q) prostate tumors for evidence of viral sequences, by hybridization to a DNA microarray composed of the most conserved sequences of all known human, animal, plant and bacterial viruses [7, 8]. Here we report that 40% (8 of 20) of all tumors homozygous for the R462Q allele harbored the genome of a distinct gammaretrovirus closely related to xenotropic MuLVs. In contrast, retroviral sequences were present in <2% of tumors bearing at least one copy of the wild-type allele (1 of 66). In addition, virus-harboring cells were detected within infected prostatic tumor tissues by fluorescence *in situ* hybridization (FISH) and immunohistochemistry (IHC). These findings represent the first detection of xenotropic MuLV-like agents in humans, and reveal a strong association between infection with the virus and defects in RNase L activity. The relation of retroviral infection to prostate cancer will require further study, but a cofactor role is not excluded.

4.3 Results

4.3.1 Detection of XMRV by microarray-based screening

To search for potential viruses in prostate cancer tumors, we employed a DNA microarray-based strategy designed to screen for viruses from all known viral families [7, 8]. Total or polyadenylated RNA extracted from tumor tissue was first amplified and fluorescently labeled in a sequence-nonspecific fashion. The amplified and labeled fragments, which contained host as well as potential viral sequences, were then hybridized to a DNA microarray (Virochip) bearing the most conserved sequences of ~950 fully-sequenced NCBI reference viral genomes (~11,000 70-mer oligonucleotides).

The Virochip was used to screen RNA samples isolated from prostate tumors of 19 individuals (Figure 4.1). A positive hybridization signal suggestive of a gammaretrovirus was detected in 7 of 11 tumors from patients homozygous for the R462Q *RNASEL* variant (QQ). In contrast, no virus was detected in 3 tumors from RQ heterozygotes, and only 1 of 5 tumors from RR individuals was positive. Clustering of the microarray oligonucleotide intensities (Figure 4.1) revealed a similar hybridization pattern in all positive cases. Furthermore, a computational analysis using E-Predict, a recently described algorithm for viral species identification [11], suggested that the same or similar mammalian gammaretrovirus was present in all positive tumors (Table 4.3). Thus, the Virochip detected the presence of a probable gammaretrovirus in half of the QQ tumor samples and only one non-QQ sample.

4.3.2 Characterization of XMRV genome

To further characterize the virus we recovered its entire genome from one of the tumors (VP35) (Figure 4.2). To obtain viral clones, we first employed a direct microarray recovery technique described previously [8]. Briefly, amplified nucleic acid from the tumor tissue, which hybridized to viral microarray oligonucleotides, was eluted from two specific spots. The eluted DNA was re-amplified, and plasmid libraries constructed from this material were screened by colony hybridization using the spots' oligonucleotides as probes. The array oligonucleotides used in this case derived from the LTR region of Murine Type-C Retrovirus (MTCR; GenBank: NC_001702) and Spleen focus-forming virus (GenBank: NC_001500; [86]). The largest recovered fragment was 415 nucleotides in length, and had 96% nucleotide identity to the LTR region of MTCR, a MuLV identified in the genome of a mouse myeloma cell line (Heinemeyer T; unpublished). These findings established that the virus in question was indeed a gammaretrovirus, and likely a relative of murine leukemia viruses. To clone and sequence the rest of the viral genome from sample VP35 (GenBank: DQ241301), we used tumor cDNA to PCR-amplify overlapping segments using primers derived from MTCR; gaps were closed using primers from earlier recovered clones (Figure 4.2B and Table 4.4). Using a similar strategy, we have also determined the full sequence of the virus from a second tumor, VP42 (GenBank: DQ241302). Finally, complete viral genomic sequence from a third tumor case, VP62, was obtained by PCR amplification of two ~ 4 Kb-long overlapping fragments jointly spanning close to the entire length of the virus (Figure 4.2B) (GenBank: DQ399707). The three sequenced genomes share >98% nucleotide identity overall and

>99% amino acid (aa) identity for predicted open reading frames (ORFs), and thus represent the same virus.

The full genome of the virus (Figures 4.2 and S1) is 8185 nucleotides long and is distinct from all known isolates of MuLV. The genome is most similar to the genomes of exogenous MuLVs, DG-75 cloned from a human B-lymphoblastoid cell line (GenBank: AF221065, [87]) and MTCR, with which it shares 94 and 93% overall nucleotide sequence identity, respectively. The genome also shares up to 95% nucleotide identity with several full-length *Mus musculus* endogenous proviruses (Figure 4.2C).

Phylogenetic trees constructed using available mammalian type C retroviral genomes and representative full-length proviral sequences from the mouse genome (Figures 4.3 and 4.10) showed that the newly identified virus is more similar to xenotropic and polytropic than to ecotropic genomes. Based on these findings we propose the provisional name “Xenotropic MuLV-related virus” (or XMRV) for this agent.

Translation of the XMRV genomic sequence using ORF Finder [88] identified two overlapping ORFs coding for the full-length Gag-Pro-Pol and Env polyproteins. No exogenous coding sequences, such as viral oncogenes, could be detected in the XMRV genome. The predicted Gag polyprotein is 536 aa long and is most similar to a xenotropic provirus on *Mus musculus* chromosome 9 (GenBank: AC121813.3), with which it shares 97 % aa identity (Figure 4.10A). The Pro-Pol polyprotein is 1197 aa long and has the highest aa identity with MuLV DG-75 and a xenotropic provirus on *M. musculus* chromosome 4 (GenBank: AL627077.14), 97% and 96%, respectively (Figure 4.10B).

An amber (UAG) stop codon separates the Gag and Pro-Pol coding sequences, analogous to other MuLVs in which a translational read-through is required to generate the full-length Gag-Pro-Pol polyprotein (reviewed in [89]).

Similar to other MuLVs [90, 86, 91-94, 87], the Env polyprotein of XMRV is in a different reading frame compared to Gag-Pro-Pol. The Env protein sequence is 645 aa long, and has the highest amino acid identity with the Env protein of an infectious MuLV isolated from a human small cell lung cancer (SCLC) line NCI-417 (GenBank: AAC97875; [95]) and MuLV NZB-9-1 (GenBank: K02730; [92]), 95% and 94%, respectively. The XMRV Env protein also shares similarly high identity with several murine xenotropic proviruses (Figure 4.10C). Conserved splice donor (AGGTAAG, position 204) and acceptor (CACTTACAG, position 5479) sites involved in the generation of *env* subgenomic RNAs [96] were found in the same relative locations as in other MuLV genomes. A multiple sequence alignment of XMRV Env and corresponding protein sequences of other representative MuLVs (Figure 4.4) showed that within three highly variable regions (VRA and VRB and VRC) known to be important for cellular tropism [97-99], XMRV has the highest aa identity with xenotropic envelopes from MuLVs NZB-9-1, NFS-Th-1 [100], and DG-75. Although unique to XMRV amino acids are present in each of the three variable regions, based on the overall similarity to the known xenotropic envelopes, we predict that the cellular receptor for XMRV is XPR1 (SYG1), the recently identified receptor for xenotropic and polytropic MuLVs [101-103].

The long terminal repeat (LTR) of XMRV is 535 nucleotides long and has the highest nucleotide identity with the LTRs from xenotropic MuLVs NFS-Th-1 (96%) and NZB-9-1 (94%). The XMRV LTRs contain known structural and regulatory elements typical of other MuLV LTRs [104, 96]. In particular, the CCAAT box, TATAAAA box, and AATAAA polyadenylation signal sequences were found in U3 at their expected locations (Figure 4.11A). U3 also contains a glucocorticoid response element (GRE) sequence AGA ACA GAT GGT CCT. Essentially identical sequences are present in genomes of other MuLVs. These elements have been shown to activate LTR-directed transcription and viral replication in vitro in response to various steroids including androgens [61, 59, 105, 57]. In addition, presence of an intact GRE is thought to be the determinant of higher susceptibility to FIS-2 MuLV infection in male compared to female NMRI mice [62, 58]. Despite these similarities, single nucleotides substitutions unique to XMRV and an insertion of an AG dinucleotide immediately downstream from the TATA box are present in U3 (Figure 4.11A). Consistent with these findings, a phylogenetic analysis based on U3 sequences from XMRV and from representative xenotropic MuLV provirus groups [106, 107] showed that XMRV U3 sequences formed a well separated cluster most similar to the group containing NFS-Th-1 and NZB-9-1 (Figure 4.11B).

The 5' *gag* leader of XMRV, defined as the sequence extending from the end of U5 to the ATG start codon of *gag*, consists of a conserved non-coding region of ~200 nucleotides, containing a proline tRNA primer binding site as well as sequences required for viral packaging [64, 66] and the initiation of translation [108, 109]. The non-coding region is followed by a ~270 nucleotide region extending from the conserved CTG alternative start

codon of *gag*. This region represents the most divergent segment of the genome compared to other MuLVs (Figures 4.5 and 4.10C). Unlike ecotropic MuLVs, where translation from this codon adds a ~90 aa N-terminal leader peptide in frame with the rest of the Gag protein, thus generating a glycosylated form of Gag [110], XMRV has a stop codon 53 aa residues downstream from the alternative start. Interestingly, both MuLV DG-75 and MTCR *gag* leader sequences are also interrupted by stop codons, and therefore are not expected to produce full-length glyco-Gag. Furthermore, a characteristic 24-nucleotide deletion was present in this region of the XMRV genome, which is not found in any known exogenous MuLV isolate. However, a shorter deletion of 9 nucleotides internal to this region is present in the sequences of several non-ecotropic MuLV proviruses found in the sequenced mouse genome (Figure 4.5). In cell culture, expression of intact glyco-Gag is not essential for viral replication [111, 112]. However, lesions in this region have been associated with interesting variations in pathogenetic properties in vivo [113-117]. For example, an alteration in 10 nucleotides affecting 5 residues in the N-terminal peptide of glyco-Gag was found to be responsible for a 100-fold difference in the frequency of neuroinvasion observed between CasFrKP and CasFrKP41 MuLV strains [118]. In addition, insertion of an octanucleotide resulting in a stop codon downstream of the CUG start codon prevented severe early hemolytic anemia and prolonged latency of erythroleukemia in mice infected with Friend MuLV [114]. While we do not yet know the pathogenetic significance of the lesions in XMRV glyco-Gag, the high degree of sequence divergence suggests that this region may be under positive selective pressure and therefore may be relevant to the establishment of infection within the human host.

4.3.3 Association of XMRV infection and R462Q *RNASEL* genotype

To further examine the association between presence of the virus and the R462Q (1385G->A) *RNASEL* genotype, we developed a specific nested RT-PCR assay based on the virus sequence recovered from one of the tumor samples (VP35, see above). The primers in this assay (Figure S1) amplify a 380-nucleotide fragment from the divergent 5' leader and the N-terminal end of *gag*. The RT-PCR was positive in 8 (40%) of 20 examined tumors from homozygous (QQ) individuals. In addition, one tumor from a homozygous wild-type (RR) patient was positive among 52 RR and 14 RQ tumors examined (Figure 4.1 and Table 4.1). Interestingly, this case was associated with the highest tumor grade among all XMRV-positive cases (Table 4.5). PCR specific for the mouse *GAPDH* gene was negative in all samples (data not shown), arguing strongly against the possibility that the tumor samples were contaminated with mouse nucleic acid. Collectively, these data demonstrate a strong association between the homozygous (QQ) R462Q *RNASEL* genotype and presence of the virus in the tumor tissue ($p < 0.00002$ by two-tail Fisher's exact test).

4.3.4 XMRV sequence diversity in samples from different patients

To examine the degree of XMRV sequence diversity in different patients, we sequenced the amplified fragments from all 9 samples, which were positive by the nested *gag* RT-PCR. The amplified *gag* fragments were highly similar (Figure 4.6A) with >98% nucleotide and >98% aa identity to each other. In contrast, the fragments had <89% nucleotide and <95% aa identity with the most related exogenous sequence of MuLV

DG-75. Several corresponding endogenous non-ecotropic sequences were more similar to the XMRV fragments, including the xenotropic provirus from *M. musculus* chromosome 9 (GenBank: AC121813.3), which was <98% identical on the nucleotide level. Nevertheless, all XMRV-derived fragments were more similar to each other than they were to any other sequence.

In addition to the *gag* gene, we also examined the same patient samples for sequence variation in the *pol* gene. We sequenced PCR fragments obtained with a set of primers targeting a 2500-nucleotide stretch in the *pol* gene (Figure S1). Similar to the *gag* fragments, the amplified *pol* fragments were highly similar (Figure 4.6B) and had >97% nucleotide and >97% aa identity to each other. In contrast, the fragments had <94% nucleotide and <95% aa identity with the most related sequence, that of MuLV DG-75. Interestingly, XMRV-derived *pol* sequences were less similar to and approximately equidistant from the examined representative xenotropic and polytropic endogenous sequences.

Close clustering of the sequenced *gag* and *pol* fragments (Figure 4.6) indicates that all microarray and RT-PCR positive cases represent infection with the same virus. On the other hand, the degree of sequence variation in the examined fragments is higher than that expected from errors introduced during PCR amplification and sequencing. The frequency of nucleotide misincorporation by Taq polymerase has been estimated as 10^{-6} – 10^{-4} ([119] and references therein), compared to the observed rate of up to 2% in the *gag* and *pol* fragments. These findings suggest that the observed XMRV sequence

variation is a result of natural sequence diversity, consistent with the virus being independently acquired by the affected patients, and argue against laboratory contamination as a possible source of XMRV.

4.3.5 Detection of XMRV in tumor-bearing prostatic tissues using FISH

To localize XMRV within human prostatic tissues, and to measure the frequency of the infected cells, XMRV nucleic acid was visualized using fluorescence *in situ* hybridization (FISH) on formalin-fixed prostate tissues. A SpectrumGreen™ fluorescently labeled FISH probe cocktail spanning all viral genes was prepared using cDNA derived from the XMRV isolate cloned from patient VP35 (Materials and Methods). Distinct FISH-positive cells were observed in the tumors positive for XMRV by RT-PCR (e.g. VP62 and VP88) (Figure 4.7). To identify cell types associated with the positive FISH signal, the same sections were subsequently stained with hematoxylin and eosin (H&E). Most FISH-positive cells were stromal fibroblasts (Figure 4.8A), including those undergoing cell division (Figure 4.8B). In addition, occasional infected hematopoietic cells were also seen (Figure 4.8C). XMRV FISH with concurrent immunostaining for cytokeratin AE1/AE3 to achieve specific labeling of epithelial cells [120] showed no XMRV infected cells which also had the epithelium-specific staining, confirming their non-epithelial origin (Figure 4.8C). While the XMRV nucleic acid was usually present within nuclei (Video S1), suggesting integrated proviral DNA, some cells showed cytoplasmic staining adjacent to the nucleus, suggestive of viral mRNA and/or pre-integration complexes in non-dividing cells (Figure 4.8A).

We also used FISH to obtain a minimal estimate of the frequency of XMRV-infected prostatic cells. For this purpose we employed a tissue microarray containing duplicates of fourteen different prostate cancer tissue specimens (Table 4.2). FISH with DNA probes derived from XMRV VP35 showed 5 to 10 XMRV/FISH positive cells (about 1% of prostate cells observed) in each of five homozygous RNase L 462Q (QQ) cases: VP29, 31, 42, 62, and 88. Patient sample VP79, also a QQ case, contained 2 positive cells (0.4% of total cells examined). All of the XMRV FISH positive cells observed were stromal cells. In contrast, three RR tissue samples and two RQ tissue samples showed one or no (<0.15%) FISH positive cells. Two of the QQ cases, VP35 and VP90, positive by gag RT-PCR showed only one FISH positive cell each (Table 4.2). Conversely, one case, VP31, was FISH positive, but gag RT-PCR negative. As expected, chromosome 1 specific probes used as a positive control specifically labeled nearly every cell from the examined case VP88, whereas a KSHV specific probe used as a negative control did not label any cells in sections from cases VP88 and VP51, but did efficiently label 293T cells transfected with KSHV DNA (data not shown). Thus, consistent with the microarray and RT-PCR data, detection of XMRV by FISH was associated primarily with QQ cases. In addition, in samples where XMRV was detected, all positive cells were stromal and did not account for more than 1% of all prostatic cells. Finally, differences in the numbers of XMRV-positive cells detected in the different samples could be due to heterogeneity in virus copy numbers between different patients and/or specific regions of the prostate sampled.

4.3.6 Detection of XMRV in tumor-bearing prostatic tissues using IHC

To identify cells expressing XMRV proteins, we assayed for the presence of Gag protein using a monoclonal antibody against spleen focusing forming virus (SFFV); this antibody is reactive against Gag proteins from a wide range of different ecotropic, polytropic and xenotropic MuLV strains [121]. Using this antibody, positive signal by IHC was observed in prostatic tissues of XMRV-positive cases VP62 and VP88, both QQ (Figure 4.9). An enhanced alkaline phosphatase red detection method allowed Gag detection in the same cells with both fluorescence (Figure 4.9A–D; left) and bright field (Figure 4.9A–D; middle) microscopy. The Gag expressing cells were observed in prostatic stromal cells with a distribution and frequency similar to that detected by FISH (Figure 4.9 and data not shown). In contrast, no Gag positive cells were observed in VP51 prostatic tissue, which is of RR genotype (Figure 4.9E).

4.4 Discussion

The results presented here identify XMRV infection in prostate tissue from approximately 40% of patients with prostate cancer who are homozygous for the R462Q variant (QQ) of RNase L, as judged by both hybridization to the Virochip microarray and by RT-PCR with XMRV-specific primers. Parallel RT-PCR studies of prostate tumors from wild-type (RR) and heterozygous (RQ) patients revealed evidence of XMRV in only 1 of 66 samples, clearly demonstrating that human XMRV infection is strongly linked to decrements in RNase L activity. This result supports the view that the R462Q RNase L variant leads to a subtle defect in innate (IFN-dependent) antiviral immunity.

As its name indicates, XMRV is closely related to xenotropic murine leukemia viruses (MuLVs). Unlike ecotropic MuLVs, such as the canonical Moloney MuLV, which grow only in rodent cells in culture, xenotropic MuLVs can grow in non-rodent cells in culture but not in rodent cell lines. Xenotropic viruses have been isolated from many inbred as well as wild mouse strains. Studies of the distribution of non-ecotropic sequences in different mouse strains show that the diversity of xenotropic proviral sequences in wild mice is greater than that found in the inbred laboratory strains [122, 107]. This finding led to the conclusion that these endogenous elements were independently and relatively recently acquired by different mouse species as a result of infection rather than inheritance [107]. Unlike ecotropic MuLVs, which can only recognize a receptor (CAT-1) specific to mouse and rat species [123-125], xenotropic viruses recognize a protein known as XPR1 or SYG1. XPR1 is expressed in all higher vertebrates, including mice, but polymorphisms in the murine gene render it unable to mediate xenotropic MuLV entry [101-103]. Thus, xenotropic MuLVs have a potential to infect a wide variety of mammalian species, including humans.

Xenotropic MuLVs have occasionally been detected in cultured human cell lines. For example, MuLV DG-75 was cloned from a human B-lymphoblastoid cell line [87], and an infectious xenotropic MuLV was detected in a human small cell lung cancer (SCLC) line NCI-417 [95]. Although laboratory contamination, either in culture or during passage of cell lines in nude mice, cannot be ruled out as a possible source in these cases, such contamination cannot explain our results. The evidence for this is as follows: (i) XMRV was detected in primary human tissues; (ii) no murine sequences (e.g. GAPDH) could be

detected in our materials by PCR; (iii) infection was predominantly restricted to human samples with the QQ *RNASEL* genotype; (iv) polymorphisms were found in the XMRV clones recovered from different patients consistent with independent acquisition of the virus by these individuals; and (v) viral nucleic acids and antigens could be detected in infected QQ prostate tissue by fluorescence *in situ* hybridization and immunohistochemistry, respectively. Taken together, the above evidence argues strongly against laboratory contamination with virus or cloned DNA material as the source of XMRV infection in the analyzed samples. To our knowledge, this report represents the first published examples of authentic infection of humans with a xenotropic MuLV-like agent. Although our efforts to clone the sites of XMRV integration into the host genome have been limited by the small amounts of prostate tissue available for this purpose, our work to clone such sites is ongoing and will provide an important additional piece of evidence for XMRV infection in humans.

The XMRV sequence is not found in human genomic DNA and none of the human endogenous retroviruses, including the only known gammaretrovirus-like human endogenous sequences (hERVs E and T) [126], bare any significant similarity to the XMRV genome. This indicates that XMRV must have been acquired exogenously by infection in positive subjects. From what reservoir, and by what route such infections were acquired is unknown. It seems unlikely that direct contact with feral mice could explain the observed distribution of infection in our cohort, since there is no reason to believe that rodent exposure would vary according to *RNASEL* genotype. It is possible that infection is more widespread than indicated by the present studies, especially if, as

seems likely, individuals with the wild-type RNase L clear infection more promptly than those with the QQ genotype. But if so, a cross-species transfer model of XMRV infection would require improbably high levels of rodent exposure for a developed society like our own. Thus, although the viral sequence suggests that the ultimate reservoir of XMRV is probably the rodent, the proximate source of the infection seems unlikely to be mice or rats. Provisionally, we favor the notion that the XMRV infections we have documented were acquired from other humans, i.e. that XMRV may have been resident in the human population for some time. This speculation will, however, require direct epidemiologic validation. It also remains to be determined if RNase L R462Q homozygotes are more sensitive to the acquisition of infection, or are simply less likely to clear infection once acquired. This is an important issue, since if the latter model is correct, it would imply that in younger humans, XMRV prevalence may be higher than what is observed in our prostate cancer cohort (mean age – 58.7 years). We are currently developing serologic assays for use in population-based studies that should shed light on these matters.

While presented work documents a clear link of XMRV infection to RNase L deficiency, we emphasize that the data we have accumulated does not mandate any etiological link to prostate cancer. Furthermore, our finding that XMRV infection is targeted to stromal cells and not to carcinoma cells and the fact that the XMRV genome harbors no host-derived oncogenes rule out two classical models for retroviral oncogenesis: direct introduction of a dominantly acting oncogene and insertional activation of such a gene. However, more indirect contributions of the virus to the tumor can certainly be envisioned. Recent work has shown that stromal cells have an active role in directly

promoting tumorigenesis of adjacent epithelial cells by producing various cytokines and growth factors that serve as proliferative signals [127] or indirectly by modifying the tumor microenvironment by promotion of angiogenesis or recruitment of inflammatory mediators leading to oxidative stress [128]. In particular, cancer associated fibroblasts stimulate growth of human prostatic epithelial cells and alter their histology in vivo [129]. It is conceivable that XMRV-infected prostatic stromal cells could produce and secrete growth factors, cytokines or other factors that stimulate cell proliferation or promote oxidative stress in surrounding epithelia. Such a paracrine mechanism could still function quite efficiently even with the relatively small number of XMRV-infected cells that characterize the lesion.

Finally, we note that the identification of an exogenous infection like XMRV could help explain why not all genetic studies have consistently identified RNase L as a prostate cancer susceptibility factor. If such an infection were linked, however indirectly, to prostate cancer risk, and if the prevalence of infection is not uniform in different populations, populations with low XMRV prevalence might be expected to show no association of *RNASEL* lesions to prostate cancer.

Clearly, resolution of these issues will require much further investigation. We need to determine the prevalence of XMRV infection in the general population, understand its routes of transmission and tissue tropism, explore its associations with premalignant and other prostatic conditions, and define the biochemical interactions of the virus with the 2-5A/RNase L system. The availability of molecular clones, infectious virus stocks and

susceptible cell culture systems should greatly enhance our ability to probe these and other questions in the near future.

4.5 Materials and Methods

4.5.1 Genotyping of patients and prostate tissue processing

All human samples used in this study were obtained according to protocols approved by the Cleveland Clinic's Institutional Review Board. Age, clinical parameters and geographical locations of XMRV-positive prostate cancer cases are provided in Table 4.5. Men scheduled to undergo prostatectomies at the Cleveland Clinic were genotyped for the R462Q (1385G->A) *RNASEL* variant using a premade TAQMAN genotyping assay (Applied Biosystems, Foster City, CA, USA; Assay c_935391_1) on DNA isolated from peripheral blood mononuclear cells. Five nanograms of genomic DNA were assayed according to the manufacturer's instructions, and analyzed on an Applied Biosystems 7900HT Sequence Detection System instrument. Immediately after prostatectomies, tissue cores were taken from both the transitional zone (the site of benign prostatic hyperplasia, BPH) and the peripheral zone (where cancer generally occurs), snap-frozen in liquid nitrogen and then stored at -80°C. Remaining prostate tissue was fixed in 10% neutral buffered formalin, processed and embedded in paraffin for later histological analyses. Frozen tissue cores were transferred from dry ice immediately to TRIZOL reagent (Invitrogen, Carlsbad, CA, USA), homogenized with a power homogenizer or manually using a scalpel followed by a syringe, and total RNA was isolated according to the manufacturer's instructions. The prostate tissue RNA was then subjected to RNase-free DNase I (Ambion, Austin, TX, USA) digestion for 30

minutes at 37°C. The sample was then extracted with phenol and the RNA was precipitated with isopropanol overnight at -20°C followed by centrifugation at 12,000 g for 30 minutes at 4°C. Poly-A RNA was isolated from the DNase digested total RNA using the Oligotex mRNA Midi Kit (Qiagen USA, Valencia, CA, USA) as instructed by the manufacturer. The poly-A RNA concentration was measured using the RIBOgreen quantitation kit (Molecular Probes, Invitrogen), and the samples were stored at -80°C.

4.5.2 Microarray screening

Virochip microarrays used in this study were identical to those previously described [7, 8, 11]. Prostate tumor RNA samples were amplified and labeled using a modified Round A/B random PCR method and hybridized to the Virochip microarrays as reported previously (Protocol S1 in [8]). Microarrays were scanned with an Axon 4000B scanner (Axon Instruments, Union City, CA, USA) and gridded using the bundled GenePix 3.0 software. Microarray data have been submitted to the NCBI GEO database (GSE3607). Hybridization patterns were interpreted using E-Predict as previously described [11] (Table 4.3). To make Figure 4.1, background-subtracted hybridization intensities of all retroviral oligonucleotides (205) were used to cluster samples and the oligonucleotides. Average linkage hierarchical clustering with Pearson correlation as the similarity metric was carried out using Cluster (v. 2.0) [10]. Cluster images were generated using Java TreeView (version 1.0.8) [63].

4.5.3 Genome cloning and sequencing

Amplified and labeled cDNA from the VP35 tumor sample was hybridized to a hand-spotted microarray containing several retroviral oligonucleotides, which had high hybridization intensity on the Virochip during the initial microarray screening. Nucleic acid hybridizing to two of the oligonucleotides (9628654_317_rc derived from MTCR: TTC GCT TTA TCT GAG TAC CAT CTG TTC TTG GCC CTG AGC CGG GGC CCA GGT GCT CGA CCA CAG ATA TCC T; and 9626955_16_rc derived from Spleen focus-forming virus: TCG GAT GCA ATC AGC AAG AGG CTT TAT TGG GAA CAC GGG TAC CCG GGC GAC TCA GTC TGT CGG AGG ACT G) was then individually eluted off the surface of the spots and amplified by PCR with Round B primers.

Preparation of the hand-spotted array, hybridization, probe recovery, and PCR amplification of the recovered material were carried out according to the Protocol S1 (see Supporting Information). The recovered amplified DNA samples were then cloned into pCR2.1-TOPO TA vector (Invitrogen), and the resulting libraries were screened by colony hybridization with the corresponding above oligonucleotides as probes.

Hybridizations were carried out using Rapid-Hyb buffer (Amersham, Piscataway, NJ, USA) according to the manufacturer's protocol at 50 °C for 4 hours. Eight positive clones were sequenced, of which two (one from each library; clones K1 and K2R1 in Figure 4.2A) were viral and had 94-95% nucleotide identity to MTCR.

To sequence the remainder of the VP35 genome as well as the entire genome from the VP42 tumor, we amplified fragments of the genome by PCR using either amplified (Round B) or unamplified (Round A) cDNA prepared for original Virochip screening.

This was accomplished first using a combination of primers derived from the sequence of MTCR (GenBank: NC_001702) and earlier recovered clones of XMRV. The two overlapping fragments from VP62 were amplified by PCR from cDNA generated by priming poly-A RNA with random hexamers. All PCR primers are listed in Table 4.4 (see Supporting Information). The amplified fragments were cloned into pCR2.1-TOPO TA vector (Invitrogen) and sequenced using M13 sequencing primers. Genome assembly was carried out using CONSED version 13.84 for Linux [130]. Assembled genome sequences of XMRV VP35, VP42, and VP62 have been submitted to GenBank (accessions DQ241301, DQ241302, and DQ399707).

4.5.4 PCR

Screening of tumor samples by *gag* nested RT-PCR was carried out according to Protocol S3 (see Supporting Information). PCR fragments in all positive cases were gel purified using QIAEX II gel extraction kit (Qiagen), cloned into pCR2.1-TOPO TA vector (Invitrogen), and sequenced using M13 sequencing primers.

Pol PCR was carried out using amplified cDNA (Round B material) as the template.

Sequence of the primers used for amplification (2670F, 3870R, 3810F, and 5190R) are listed in Table 4.4. Amplified products were gel purified using QIAEX II gel extraction kit (Qiagen), and purified products were directly used for sequencing.

4.5.5 Phylogenetic analyses

The following sequence records were used in the analyses: MTCR (GenBank: NC_001702), MuLV DG-75 (GenBank: AF221065), MuLV MCF1233 (GenBank: U13766), AKV MuLV (GenBank: J01998), Friend MuLV (GenBank: NC_001362), Rauscher MuLV (GenBank: NC_001819), Moloney MuLV (GenBank: NC_001501), Feline leukemia virus (GenBank: NC_001940), Gibbon ape leukemia virus (GenBank: NC_001885), and Koala retrovirus (GenBank: AF151794). In addition, xenotropic mERV Chr.1 (GenBank: AC083892; nucleotides 158,240-166,448), xenotropic mERV Chr.4 (GenBank: AL627077; nucleotides 146,400-154,635), and xenotropic mERV Chr.9 (GenBank: AC121813; nucleotides 37,520-45,770) were chosen by BLAST querying the NCBI nr database with the complete XMRV genomes and selecting the most similar full-length proviral sequences, all of which happened to have xenotropic envelopes (Figure 4.10C). Polytypic mERVs, polytypic mERV Chr.7 (GenBank: AC167978; nucleotides 57,453-65,805) and polytypic mERV Chr.11 (GenBank: AC645571; nucleotides 168,229-176,580), were chosen by selecting NCBI nr full-length proviral sequences with envelopes most similar to a prototype polytypic clone MX27 (GenBank: M17327; [131]). Similarly, modified polytypic mERV Chr.7 (GenBank: AC127565; nucleotides 64,355-72,720) and modified polytypic mERV Chr.12 (GenBank: AC153658; nucleotides 85,452-93,817) were selected on the basis of similarity to a prototype modified polytypic clone MX33 (GenBank: M17327, [131]). U3 analysis was performed using previously described reference sequences: Mcv18, Mcv3, Mxv2, Mcv11, Mxv11, and HEMV18 [107]; CWM-T-15, CWM-T-15-4, CWM-T-25a, and CWM-T-25b [106].

To generate the neighbor-joining tree of complete genomic sequences (Figure 4.3) the sequences were first manually edited to make all genomes the same length, i.e. R to R. The edited sequences were then aligned with ClustalX version 1.82 for Linux [65, 132] using default settings. The tree was generated based on positions without gaps only; Kimura correction for multiple base substitutions [133] and bootstrapping with N=1000 were also used.

All other trees were generated as above, except sequences were first trimmed to the same length, gaps were included, and Kimura correction was not used, as using these parameters did not have any significant effect on the trees.

4.5.6 Antibodies

Monoclonal antibody to SFFV Gag protein was produced from R187 cells ([121]; ATCC: CRL-1912) grown in DMEM (Media Core, Cleveland Clinic Foundation, Cleveland, OH) with 10% ultra-low IgG FBS (Invitrogen) until confluent. Conditioned media was collected every three days from confluent cultures. Five ml of conditioned media per preparation was centrifuged at 168 x g for 5 min at 4°C. Supernatant was filtered through a 0.22 µm syringe filter unit (Millipore, Billerica, MA, USA) and concentrated 16-fold in an Amicon ultrafiltration unit with a 100 kDa molecular weight cutoff membrane (Millipore). Sodium azide was added to a final concentration of 0.02%. Concomitant XMRV FISH/cytokeratin immunofluorescence was performed using a mouse anti-cytokeratin AE1/AE3 (20:1 mixture) monoclonal antibody (Chemicon International,

Temecula, CA, USA) capable of recognizing normal and neoplastic cells of epithelial origin.

4.5.7 FISH

The XMRV-35 FISH probe cocktail was generated using both 2.15kb and 1.84 kb segments of the viral genome obtained by PCR with forward primer-2345, 5' ACC CCT AAG TGA CAA GTC TG 3' with reverse primer-4495, 5' CTG GAC AGT GAA TTA TAC TA 3' and forward primer-4915, 5' AAA TTG GGG CAG GGG TGC GA 3' with reverse primer-6755, 5' TTG GAG TAA GTA CCT AGG AC 3', both cloned into pGEM-T (Promega, Madison, WI, USA). The recombinant vectors were digested with *EcoRI* to release the viral cDNA fragments, which were purified after gel electrophoresis (Qiagen). The purified viral cDNA inserts were used in nick translation reactions to produce SpectrumGreen dUTP fluorescently labeled probe according to manufacturer's instructions (Vysis Inc., Des Plaines, IL, USA). Freshly baked slides of prostatic tissues or tissue microarray arrays with ~4 µm thick tissue sections were deparaffinized, rehydrated, and subjected to Target Retrieval (Dako, Glostrup, Denmark) for 40 min at 95°C. Slides were cooled to room temperature and rinsed in H₂O. Proteinase K (Dako) at 1:5000 in Tris-HCl pH 7.4 was applied directly to slides for 10 min at room temperature. Adjacent tissue sections were also probed with SpectrumGreen dUTP fluorescently labeled KSHV-8 DNA (nts 85820-92789) as a negative control or, as a positive control with SpectrumGreen and SpectrumOrange labeled TelVysion DNA Probe cocktail (Vysis Inc.), specific for subtelomeric regions of the P and Q arms of human chromosome 1 as a positive control to ensure the tissue was completely accessible

to FISH. FISH slides were examined using a Leica DMR microscope (Leica Micro-Systems, Heidelberg, Germany), equipped with a Retiga EX CCD camera (Q-Imaging, Vancouver, British Columbia, Canada). FISH images were captured using a Leica TCS SP2 laser scanning confocal with a 63X oil objective numerical aperture (N.A.) 1.4 (Leica Micro-Systems) microscope. XMRV nucleic acids were visualized using maximum intensity projections of optical slices acquired using a 488 nm argon-laser (emission at 500 to 550 nm). TelVysion™ DNA Probes were visualized using maximum intensity projections of optical slices acquired using a 488 nm argonlaser (emission at 500 to 550 nm) and 568 nm krypton-argon-laser (emission at 575 to 680 nm). DAPI was visualized using maximum intensity projections of optical slices acquired using a 364 nm UV-laser (emission at 400 to 500 nm). Slides were subsequently washed in 2X SSC (0.3 M sodium chloride and 0.03 M sodium citrate, pH 7.0) to remove coverslips, and H&E stained for morphological evaluation.

4.5.8 IHC

IHC on human tissues was performed on a Benchmark Ventana Autostainer (Ventana Medical Systems, Tucson, AZ, USA). Unstained, formalin fixed, paraffin embedded prostate sections were placed on electrostatically charged slides and deparaffinized followed by a mild cell conditioning achieved through the use of Cell Conditioner #2 (Ventana Medical Systems). The concentrated R187 monoclonal antibody against SFFV p30 Gag was dispensed manually onto the sections at 10 µg per ml and allowed to incubate for 32 min at 37°C. Endogenous biotin was blocked in sections using the Endogenous Biotin Blocking Kit (Ventana Medical Systems). Sections were washed, and

biotinylated ImmunoPure Goat Anti-Rat IgG (Pierce Biotechnology, Rockford, IL, USA) was applied at a concentration of 4.8 µg per ml for 8 min. To detect Gag protein localization, the Ventana Enhanced Alkaline Phosphatase Red Detection Kit (Ventana Medical Systems) was used. Sections were briefly washed in distilled water and counterstained with Hematoxylin II (Ventana Medical Systems) for approximately 6 min. Sections were washed, dehydrated in graded alcohols, incubated in xylene for 5 min and coverslips were added with Cytoseal (Microm International, Walldorf, Germany). Negative controls were performed as above except without the addition of the R187 monoclonal antibody.

Concomitant XMRV FISH/cytokeratin IHC was performed on slides of prostate tissue from patient VP62. First, sections were immunostained for cytokeratin AE1/AE3 using the Alexa Fluor 594 Tyramide Signal Amplification Kit (Molecular Probes, Invitrogen). Briefly, unstained, formalin fixed, paraffin embedded sections cut at ~4 µm were placed on electrostatically charged slides, baked at 65°C for at least 4 hr, deparaffinized in xylene and rehydrated through decreasing alcohol concentrations. Slides were incubated in Protease II (Ventana Medical Systems) for 3 min at room temperature and washed in phosphate buffered saline (PBS) in peroxidase quenching buffer (PBS + 3% H₂O₂) for 60 min at room temperature, then incubated with 1% blocking reagent (10 mg/ml BSA in PBS) for 60 min at room temperature. The slides were incubated with cytokeratin AE1/AE3 antibody diluted in 1% blocking reagent for 60 min at room temperature and rinsed three times in PBS. Goat anti-mouse IgG-horseradish peroxidase (Molecular Probes, Invitrogen) was added and incubated for 60 min at room temperature. The slides

were rinsed three times in PBS. The tyramide solution was added to the slides for 10 min at room temperature and the slides were rinsed 3X in PBS. Slides were then placed in Target Retrieval solution (Dako) for 40 min at 95°C. FISH for XMRV was performed as described above except in the absence of proteinase K treatment. After FISH, the slides were mounted with Vectashield Mounting Medium plus DAPI (Vector Labs, Burlingame, CA, USA) and examined using fluorescence microscopy. Immunofluorescence images were captured using a Texas red filter with a Leica DMR microscope (Leica Microsystems), equipped with a Retiga EX CCD camera (QImaging).

4.6 Supporting Information

The following supporting data are available as a web supplement to the PLoS Pathogens article describing this work (<http://pathogens.plosjournals.org/>):

Figure S1. Complete Nucleotide Sequence of XMRV VP35

Protocol S1. Probe Recovery from Hand-Spotted Microarrays by “Scratching”

Protocol S2. XMRV *gag* Nested RT-PCR

Video S1. Confocal Optical Image Planes of a Representative XMRV FISH Positive Cell

4.7 Accession Numbers

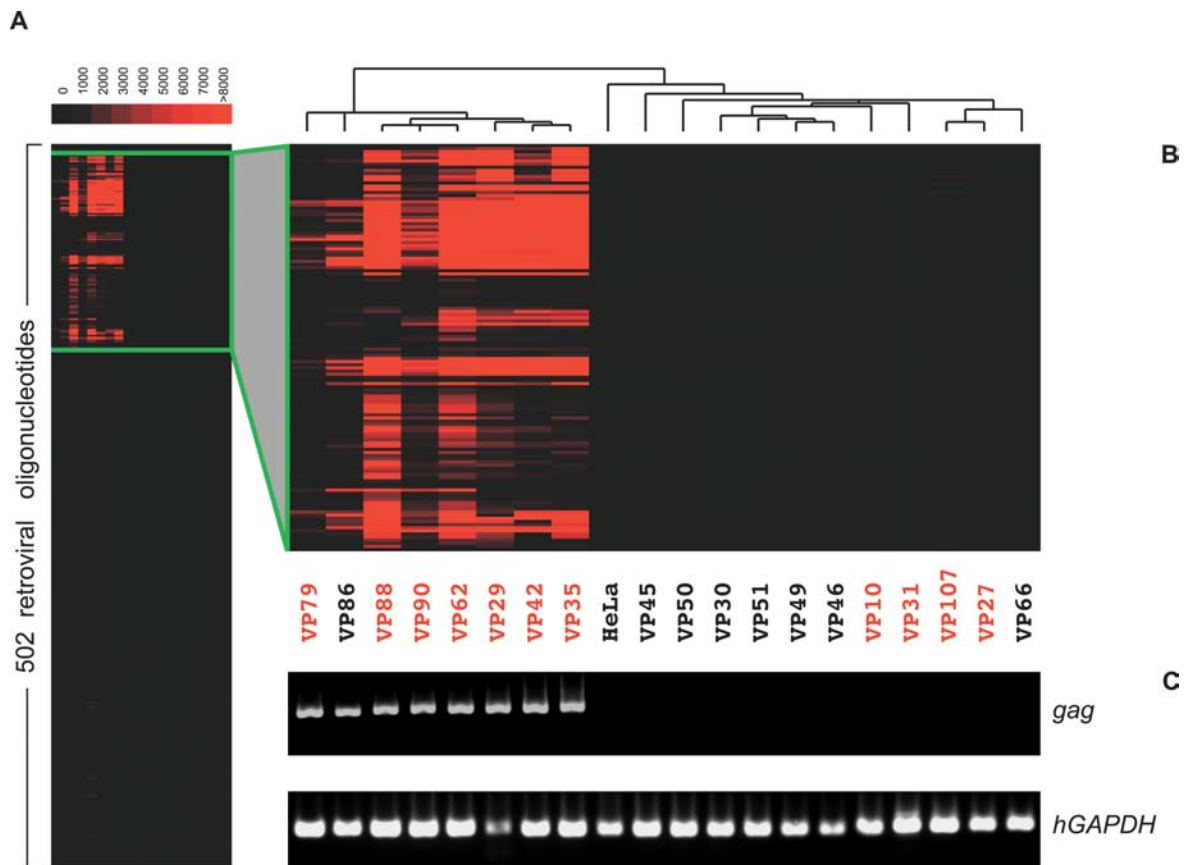
Accession numbers from Gen Bank (<http://www.ncbi.nlm.nih.gov/Genbank>) are: AKV MuLV (J01998), feline leukemia virus (NC_001940), Friend MuLV (NC_001372), gibbon ape leukemia virus (NC_001885), koala retrovirus (AF151794), modified polytropic mERV Chromosome 7 (AC127565; nt 64,355–72,720), modified polytropic mERV Chromosome 12 (AC153658; nt 85,452–93,817), Moloney MuLV (NC_001501),

MTCR (NC_001702MuLV DG-75 (AF221065); MuLV MCF 1233 (U13766), MuLV NCI-417 (AAC97875), MuLV NZB-9-1 (K02730), polytropic mERV Chromosome 7 (AC167978; nt 57,453–65,805), polytropic mERV Chromosome 11 (168–229,176,580), prototype polytropic clone MX27 (M17327), Rauscher MuLV (NC_001819), xenotropic mERV Chromosome 1 (AC083892, nt 158,240–166,448), xenotropic mERV Chromosome 4 (AL627077; nt 146,400–154,635), xenotropic mERV Chromosome 9 (AC121813; nt 37,520–45,770), XMRV VP35 (DQ241301), XMRV VP42 (DQ241302), and XMRV VP 62 (DQ399707).

4.8 Acknowledgements

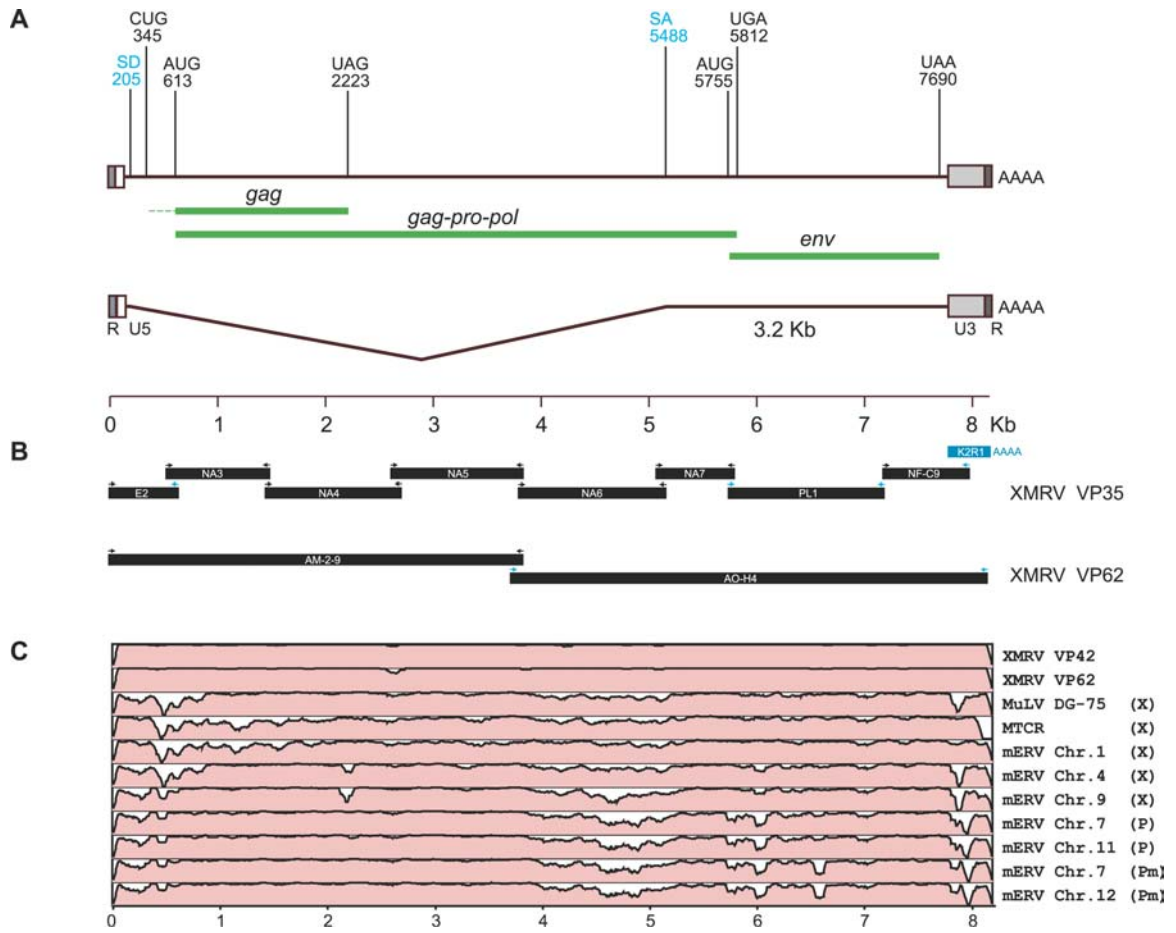
We thank Silvi Rouskin, Shoshannah Beck, James Pettay, and Jayashree Paranjape for expert technical assistance; Sanggu Kim and Samson A. Chow for technical advise; Earl Poptic for production and purification of monoclonal antibodies to Gag; Stephen T. Koury for advice; and Judith A. Drazba and Amit Vasanji for assistance with confocal imaging. This investigation was supported by Genentech Graduate Fellowship and a grant from the Sandler Family Supporting Foundation (AU), grants from Doris Duke Charitable Foundation (JDR and DG) and the David and Lucille Packard Foundation (JDR), Howard Hughes Medical Institute (JDR and DG), by NIH/NCI grants (to RHS and GC), and a Molecular Medicine Fellowship from Cleveland State University and the Cleveland Clinic Foundation (RJM).

Figure 4.1: XMRV Detection by DNA Microarrays and RT-PCR



(A) Virochip hybridization patterns obtained for tumor samples from 19 patients. The samples (x-axis) and the 502 retroviral oligonucleotides present on the microarray (y-axis) were clustered using hierarchical clustering. The red color saturation indicates the magnitude of hybridization intensity. **(B)** Magnified view of a selected cluster containing oligonucleotides with the strongest positive signal. Samples from patients with QQ *RNASEL* genotype are shown in red, and those from RQ and RR individuals as well as controls are in black. **(C)** Results of nested RT-PCR specific for XMRV *gag* gene. Amplified *gag* PCR fragments along with the corresponding human GAPDH amplification controls were separated by gel electrophoresis using the same lane order as in the microarray cluster.

Figure 4.2: Complete Genome of XMRV

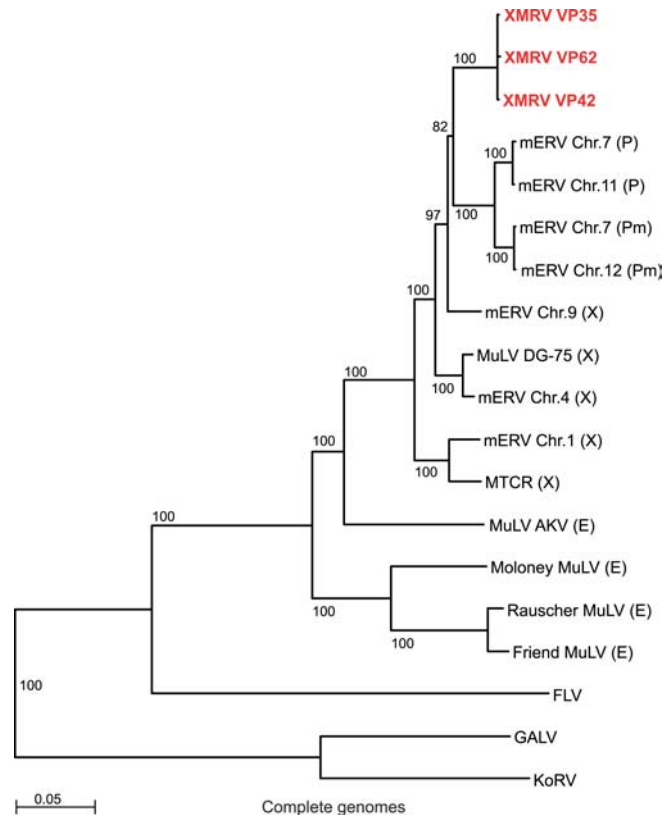


(A) Schematic map of the 8185 nt XMRV genome. LTR regions (R, U5, U3) are indicated with boxes. Predicted open reading frames encoding Gag, Gag-Pro-Pol, and Env polyproteins are labeled in green. The corresponding start and stop codons (AUG, UAG, UGA, UAA) as well as the alternative Gag start codon (CUG) are shown with their nt positions. Similarly, splice donor (SD) and acceptor (SA) sites are shown and correspond to the spliced 3.2-kb Env subgenomic RNA (wiggled line).

(B) Cloning and sequencing of XMRV VP35 and VP62 genomes. Clones obtained by probe recovery from hybridizing microarray oligonucleotides (blue bar) or by PCR from tumor cDNA (black bars) were sequenced. Primers used to amplify individual clones (Table 4.4) were derived either from the genome of MTCR (black arrows) or from overlapping VP35 clones (blue arrows).

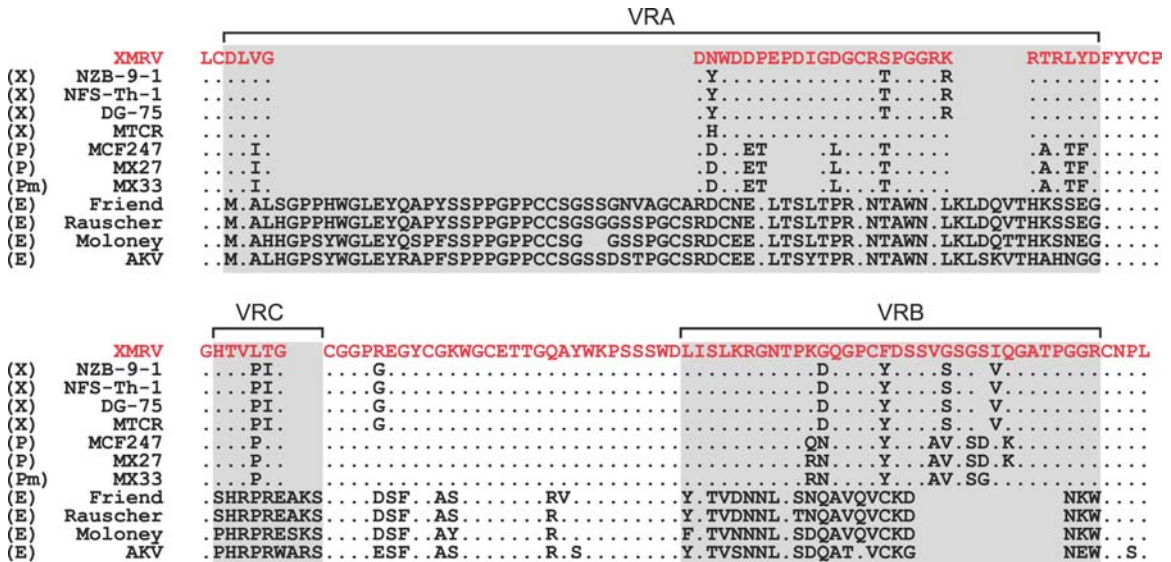
(C) Genome sequence similarity plots comparing XMRV VP35 with XMRV VP42, XMRV VP62, MuLV DG-75, MTCR, and a set of representative non-ecotropic proviruses (mERVs) (see Materials and Methods). The alignments were made using AVID [134], and plots were generated using mVISTA [135] with the default window size of 100 nt. Y-axis scale for each plot represents percent nt identities from 50% to 100%. Sequences are labeled as xenotropic (X), polytropic (P), or modified polytropic (Pm).

Figure 4.3: Phylogenetic Analysis of XMRV Based on Complete Genome Sequences



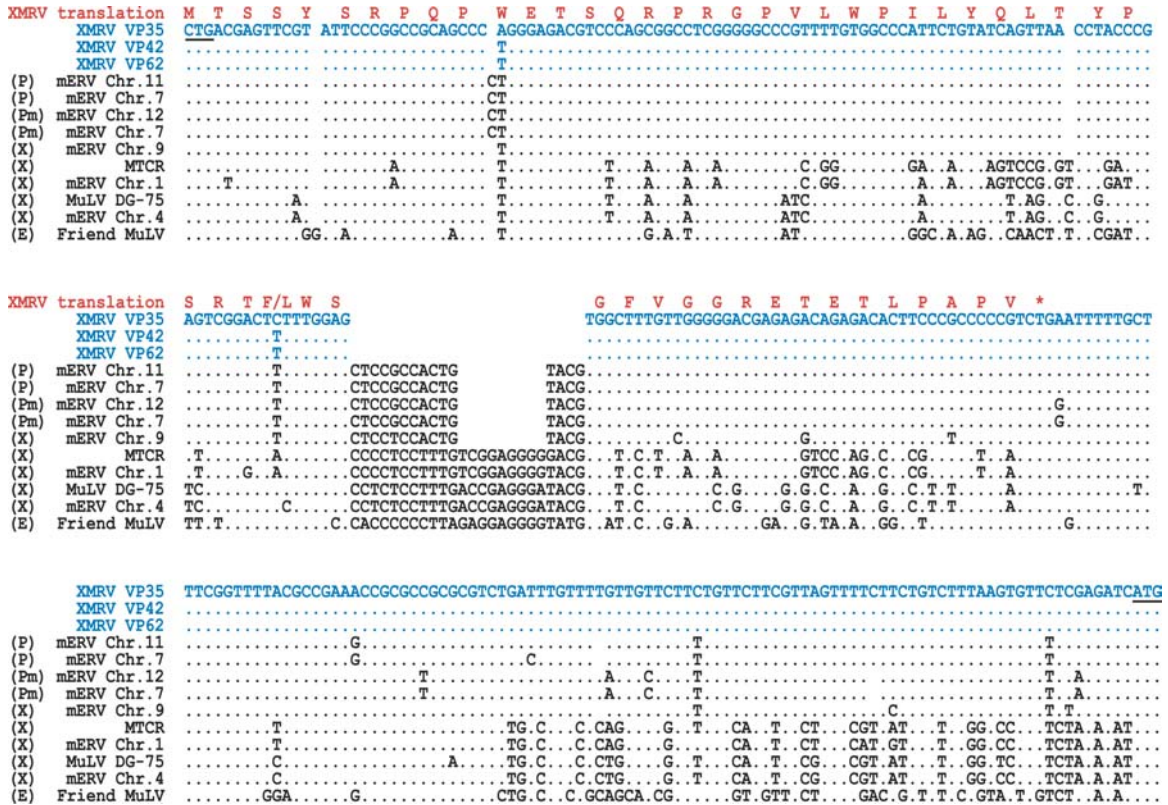
Complete genomes of XMRV VP35, VP42, and VP62 (red); MTCR; MuLVs DG-75, AKV, Moloney, Friend, and Rauscher; feline leukemia virus (FLV); koala retrovirus (KoRV); gibbon ape leukemia virus (GALV); and a set of representative non-ecotropic proviruses (mERVs) were aligned using ClustalX (see Materials and Methods). An unrooted neighbor-joining tree was generated based on this alignment, excluding gaps and using Kimura's correction for multiple base substitutions. Bootstrap values ($n = 1000$ trials) are indicated as percentages. Sequences are labeled as xenotropic (X), polytropic (P), modified polytropic (Pm), or ecotropic (E).

Figure 4.4: Multiple-Sequence Alignment of Protein Sequences from XMRV and Related MuLVs Spanning SU Glycoprotein VRA, VRB, and VRC, Known to Determine Receptor Specificity



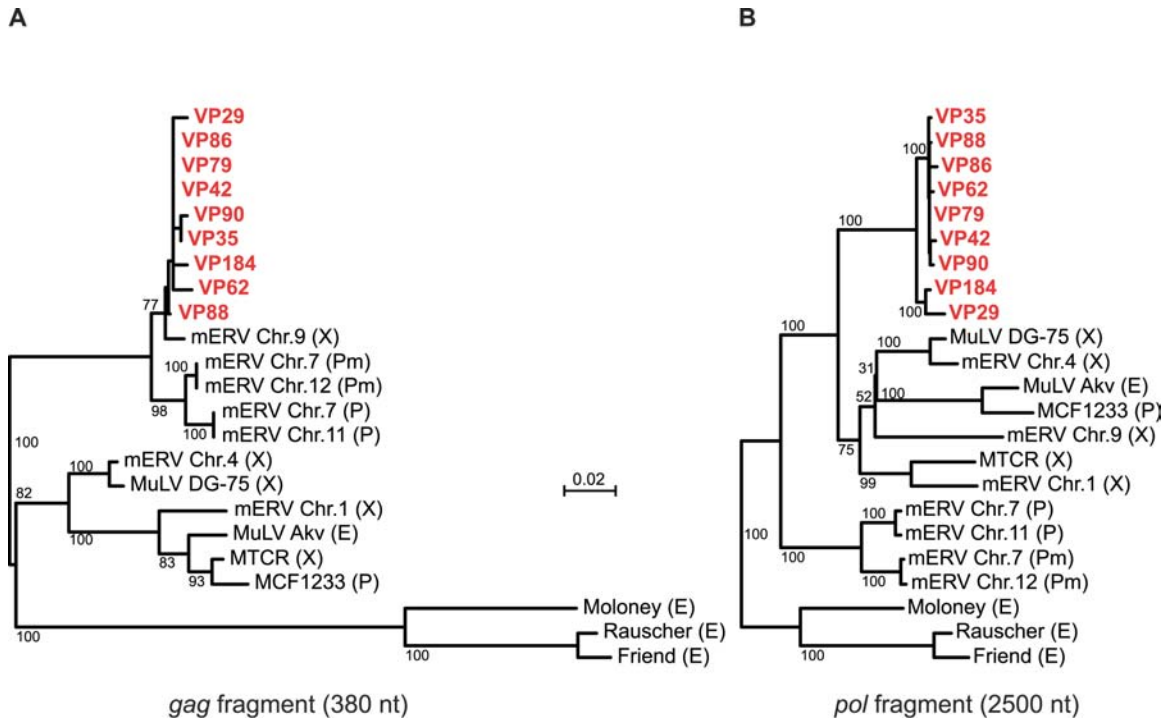
Env protein sequence from XMRV (identical in VP35, VP42, and VP62; red); MTCR; MuLVs DG-75, NZB-9-1, NFS-Th-1, MCF247, AKV, Moloney, Friend, and Rauscher; and polytropic proviruses MX27 and MX33 [131] were aligned using ClustalX. Sequences are labeled as xenotropic (X), polytropic (P), modified polytropic (Pm), or ecotropic (E). VRs are boxed. Dots denote residues identical to those from XMRV, and deleted residues appear as spaces.

Figure 4.5: Multiple-Sequence Alignment of 5' *gag* Leader Nucleotide Sequences from XMRV and Related MuLVs



Sequences extending from the alternative CUG start codon to the AUG start codon (underlined) of *gag* derived from XMRV VP35, VP42, and VP62 (blue); MTCR, MuLVs DG-75, and Friend; and a set of representative non-ecotropic proviruses (mERVs) were aligned with ClustalX (see Materials and Methods). Predicted amino acid translation corresponding to the VP35 sequence is shown above the alignment (red); asterisk indicates a stop. Sequences are labeled as xenotropic (X), polytropic (P), modified polytropic (Pm), or ecotropic (E). Dots denote nt identical to those from XMRV, and deleted nt appear as spaces.

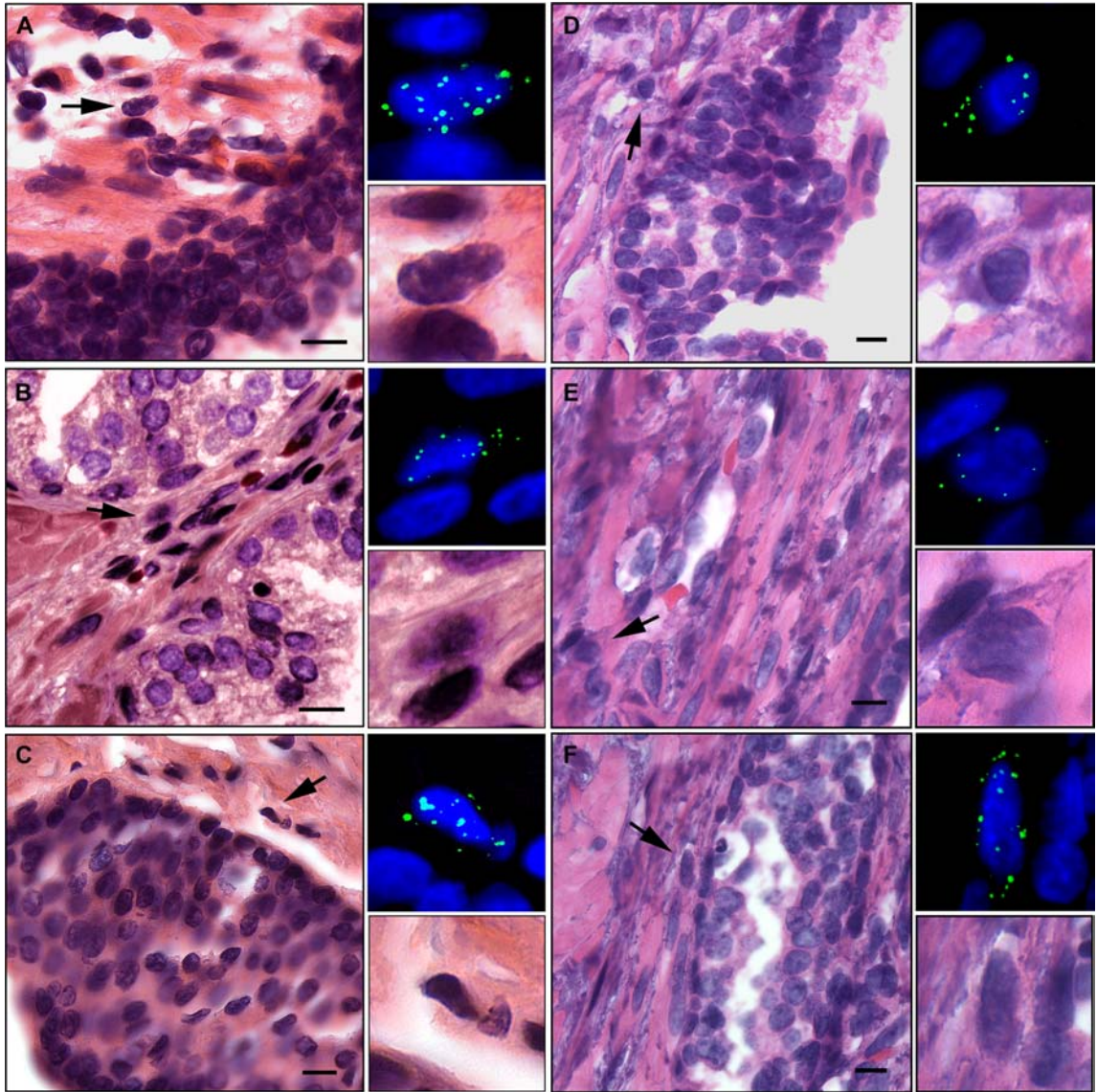
Figure 4.6: Comparison of XMRV Sequences Derived from Tumor Samples of Different Patients



(A) Phylogenetic tree based on the 380 nt XMRV *gag* RT-PCR fragment from the nine positive tumor samples (red) and the corresponding sequences from MTCR; MuLVs DG-75, MCF1233, Akv, Moloney, Rauscher and Friend; and a set of representative non-ecotropic proviruses (mERVs). The sequences were aligned using ClustalX, and the corresponding tree was generated using the neighbor-joining method (see Materials and Methods). Bootstrap values ($n = 1000$ trials) are indicated as percentages. Sequences are labeled as xenotropic (X), polytropic (P), modified polytropic (Pm), or ecotropic (E).

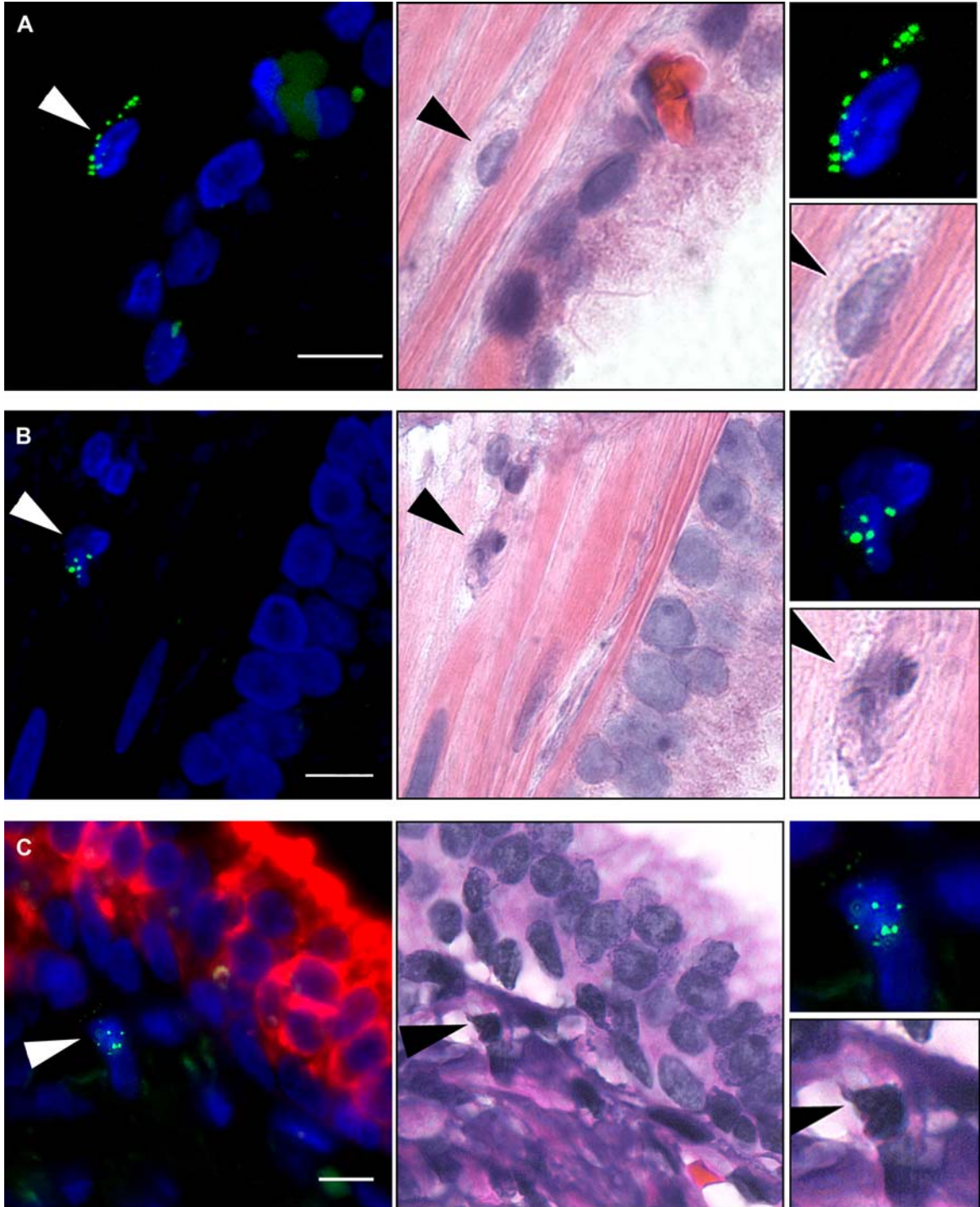
(B) Phylogenetic tree based on a 2500-nt *pol* PCR fragment from the 9 XMRV-positive tumor samples. The tree was constructed as described in (A).

Figure 4.7: Detection of XMRV Nucleic Acid in Prostatic Tissues Using FISH



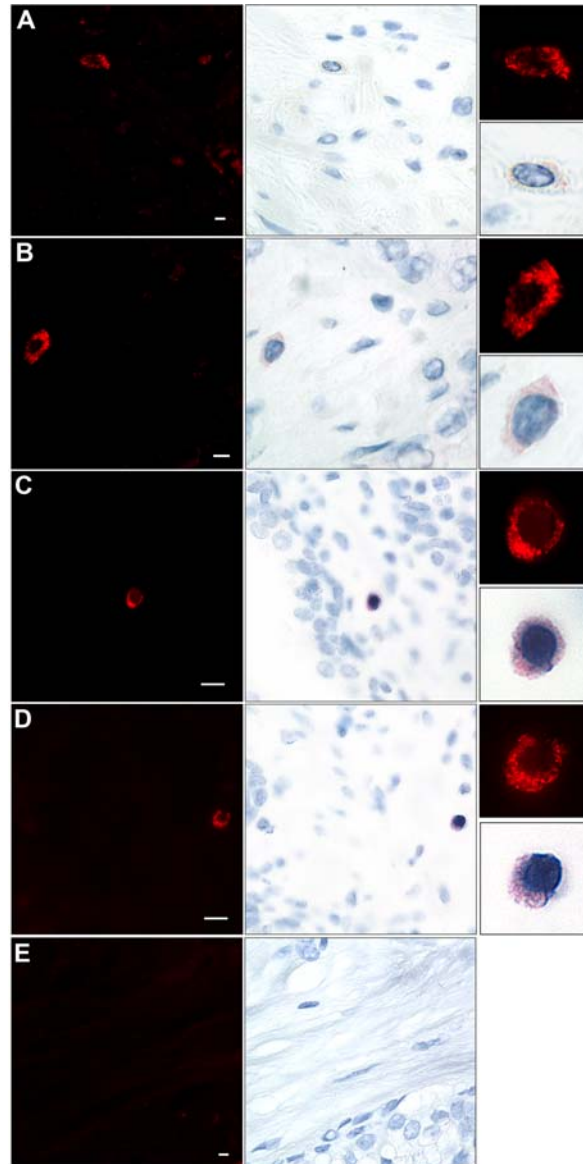
Prostatic tumor tissue sections from QQ cases VP62 (A–C) and VP88 (D–F) were analyzed by FISH using DNA probes (green) derived from XMRV VP35 (top right enlargements). Nuclei were counterstained with DAPI. The same sections were then visualized by H&E staining (left panels). Scale bars are 10 μ m. Arrows indicate FISH positive cells, and their enlarged images are shown in the bottom right panels.

Figure 4.8: Characterization of XMRV-Infected Prostatic Cells by FISH and FISH/Immunofluorescence



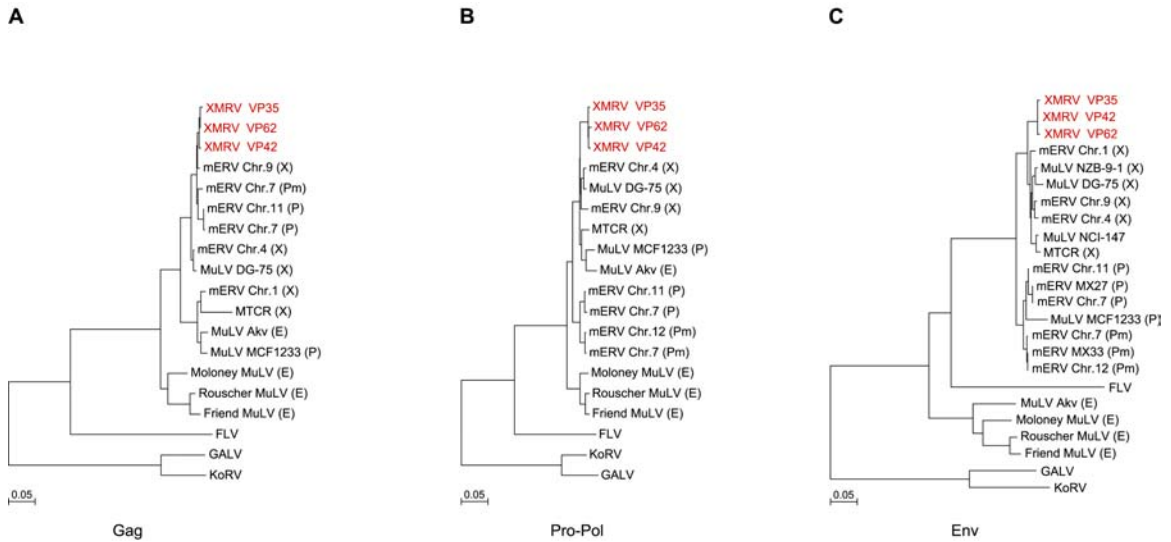
Using a tissue microarray, prostatic tumor tissue sections from QQ case VP62 were analyzed by FISH (green) using DNA probes derived from XMRV VP35 (left panels). Nuclei were counterstained with DAPI. The same sections were then visualized by H&E staining (middle panels). Arrows indicate FISH-positive cells, and their enlarged FISH and H&E images are shown in the top right and bottom right panels, respectively. Scale bars are 10 μm . **(A)** A stromal fibroblast. **(B)** A dividing stromal cell. **(C)** A stromal hematopoietic cell. The section was concomitantly stained for XMRV by FISH (green) and cytokeratin AE1/AE3 by immunofluorescence (red).

Figure 4.9: Detection of XMRV Protein in Prostatic Tissues Using Immunostaining



Prostatic tumor tissue sections from QQ cases VP62 (A and B) and VP88 (C and D), as well as an RR case VP51 (E) were stained, then visualized by immunofluorescence (left) or bright field (middle) using a monoclonal antibody to SFFV Gag protein. Nuclei are counterstained with hematoxylin. Enlarged images corresponding to the positive cells are shown on the right. Scale bars are 5 μm in (A), (B), and (E) and 10 μm in (C) and (D).

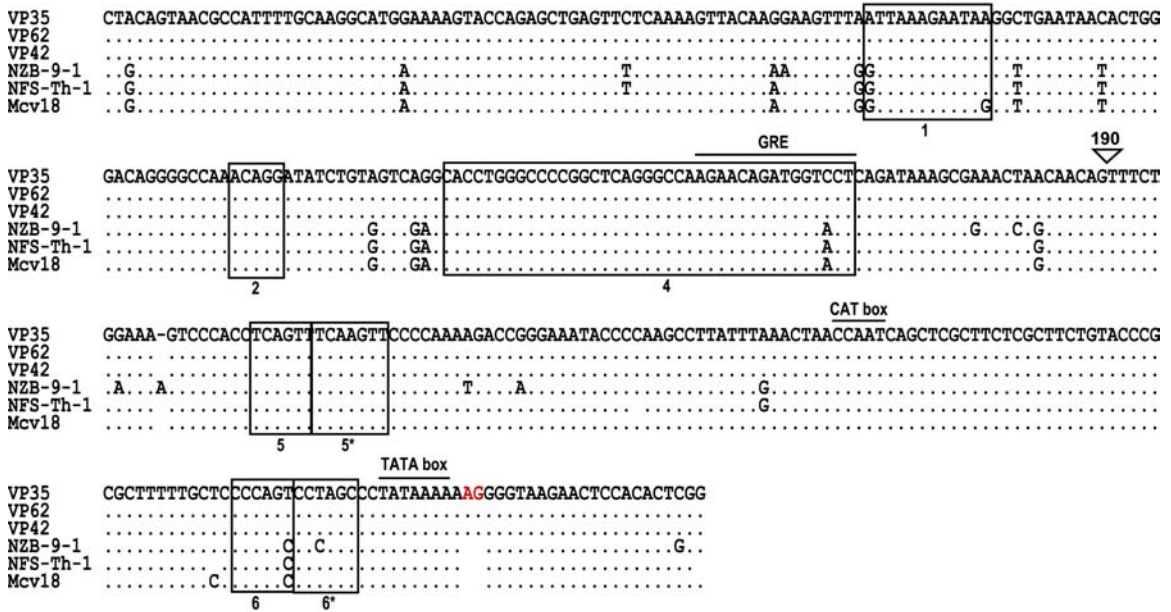
Figure 4.10: Phylogenetic Analysis of XMRV Based on Predicted Gag, Pro-Pol, and Env Polyproteins



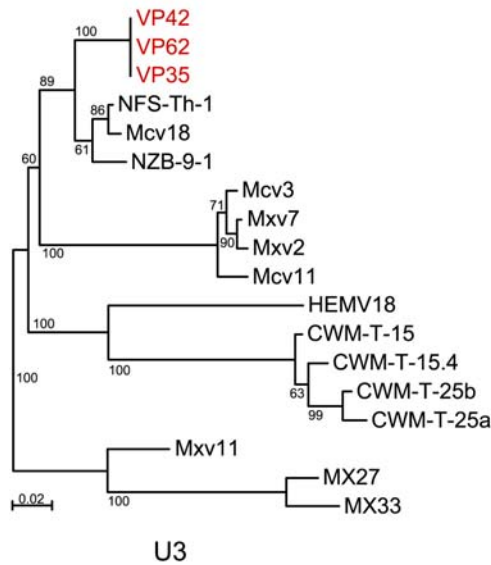
Predicted Gag (A), Pro-Pol (B), and Env (C) sequences of XMRV VP35, VP42, and VP62 (red) as well as the corresponding sequences from MTCR; MuLVs DG-75, MCF1233, Akv, Moloney, Friend, and Rauscher; feline leukemia virus (FLV); koala retrovirus (KoRV); gibbon ape leukemia virus (GALV), and a set of representative non-ecotropic proviruses (mERVs) were aligned using ClustalX. The resulting alignments were used to generate unrooted neighbor-joining trees (see Materials and Methods). Sequences are labeled as xenotropic (X), polytropic (P), modified polytropic (Pm), or ecotropic (E).

Figure 4.11: Comparison of XMRV U3 Region to Representative Non-Ecotropic Sequences

A



B



(A) Multiple sequence alignment of U3 sequences from XMRV VP35, VP42, and VP62; MuLVs NZB-9-1 and NFS-Th-1; and from representative non-ecotropic proviruses [100, 106, 107]. The sequences were aligned using ClustalX (see Materials and Methods). Only sequences most similar to XMRV are shown. Glucocorticoid response element (GRE), and TATA and CAT boxes are indicated by lines. Direct repeat regions (boxed) are numbered according to the existing convention [100, 107]. Triangle indicates a 190 nt insertion in polytropic proviruses [100]. XMRV-specific AG dinucleotide insertion is shown in red. Dots denote nucleotides identical to those from XMRV, and deleted nucleotides appear as spaces.

(B) Phylogenetic tree based on U3 nt sequences. Multiple sequence alignment from (A) was used to generate an unrooted neighbor-joining tree (see Materials and Methods). Bootstrap values (n = 1000 trials) are shown as percentages. U3 sequences from XMRV are shown in red.

Table 4.1: XMRV Screening by *gag* Nested RT-PCR

	<i>RNASEL</i> genotype ^a			Total
	QQ	RQ	RR	
PCR +	8	0	1	9
PCR -	12	14	51	77
Total	20	14	52	86

^a*RNASEL* genotypes are as follows: QQ, homozygous R462Q variant; RQ, heterozygous; RR, homozygous wild-type.

Table 4.2: Frequency of XMRV Infected Prostatic Cells Determined by FISH

Patient	<i>RNASEL</i> genotype	Number of cells counted ^b	Number of FISH-positive cells (%)	XMRV FISH ^c	XMRV <i>gag</i> RT-PCR
VP 88	QQ	408	5 (1.23)	++	+
VP 31	QQ	526	6 (1.14)	++	-
VP 42	QQ	530	6 (1.13)	++	+
VP 62	QQ	904	10 (1.11)	++	+
VP 29	QQ	659	7 (1.06)	++	+
VP 79	QQ	464	2 (0.43)	+	+
VP 10	QQ	872	1 (0.12)	+/-	-
VP 35	QQ	849	1 (0.12)	+/-	+
VP 90	QQ	843	1 (0.12)	+/-	+
VP 45	RQ	987	0 (0)	-	-
VP 46	RQ	794	0 (0)	-	-
VP 30	RR	661	1 (0.15)	+/-	-
VP 50	RR	787	1 (0.13)	+/-	-
VP 51	RR	842	0 (0)	-	-

^a*RNASEL* genotypes are as follows: QQ, homozygous R462Q variant; RQ, heterozygous; RR, homozygous wild-type. ^bAll types of prostatic cells are included. ^c+/- = 0.1-0.2%; + = 0.2-1%; ++ = >1%.

Table 4.3: Computational Viral Species Predictions Using E-Predict for the Virochip Microarrays Shown in Figure 4.1

Sample	Array ID	Top prediction ($p < 0.05$) ^a	NCBI Taxonomy ID	p-value
VP10	MegaViroP7-244	NA		
VP27	MegaViroP7-245	NA		
VP29	MegaViroP5-174	Spleen focus-forming virus	11819	1.3E-05
VP31	MegaViroP5-176	NA		
VP35	MegaViroP5-177	Spleen focus-forming virus	11819	1.0E-05
VP42	MegaViroP5-178	Murine osteosarcoma virus	11830	1.5E-05
VP62	MegaViroP8-037	Spleen focus-forming virus	11819	2.0E-05
VP79	MegaViroP8-030	Murine type C retrovirus	44561	2.9E-03
VP88	MegaViroP8-031	Spleen focus-forming virus	11819	1.4E-05
VP90	MegaViroP8-032	Spleen focus-forming virus	11819	2.4E-04
VP107	MegaViroP7-246	NA		
VP45	MegaViroP5-195	NA		
VP46	MegaViroP5-196	NA		
VP49	MegaViroP5-197	NA		
VP30	MegaViroP5-175	NA		
VP50	MegaViroP10-128	NA		
VP51	MegaViroP5-199	NA		
VP66	MegaViroP8-035	NA		
VP86	MegaViroP8-036	Spleen focus-forming virus	11819	8.2E-04
HeLa	MegaViroP5-179	Human papillomavirus type 18	10582	1.0E-06

^aMicroarrays were analyzed using E-Predict as described previously [11].

Table 4.4: PCR Primers Used for Sequencing of XMRV**Genomes**

Primer	Sequence	XMRV nucleotide position	Source
1F	5'-GCGCCAGTCATCCGATAGACT	1	MTCR
NA3-136R	5'-CCCAGTGCTGCAAGGTTAGA	661	XMRV VP35
550F	5'-CGCCGAAACCGCGCCGCGCGT	526	MTCR
1500R	5'-TCGTCGCCCCGGACTGCCTTTCTG	1499	MTCR
1470F	5'-GACAGGAGAAGAAAAGCAGCG	1441	MTCR
2730R	5'-GCTTGGCGAACTGCCAGTCCC	2721	MTCR
2670F	5'-AGCCGGATGTTTCTCTAGGGT	2631	MTCR
3724F	5'-CTAACGCAAAAAGTGGGACCTTG	3724	XMRV VP35
3870R	5'-GCTTGCCTGCATCTTTTGTG	3859	MTCR
3810F	5'-AGACCCAGTGGCAGCCGGGT	3780	MTCR
5190R	5'-TGACTTACCTGGGAGACGAAG	5182	MTCR
5100F	5'-AACTGCCAAGTTGTGACCAA	5071	MTCR
5842R	5'-AACTATTGGGGGCCCCACGGGTTA	5819	MTCR
NA7-F	5'-CATGGAAAGTCCAGCGTTCT	5754	XMRV VP35
C9-R	5'-AGCTGCTCGAATTGTTTGGT	7204	XMRV VP35
7200F	5'-CTAGTGGCCACCAACAATTC	7173	MTCR
7600F	5'-CGCTTGGTCCAGTTTGTAATA	7580	MTCR
227R	5'-TGGGGAAGTTGAAACTGAGG	7991	MTCR
100F	5'-AGGGGCCAAACAGGATAACT	7780	MTCR
227R	5'-TGGGGAAGTTGAAACTGAGG	7991	MTCR
B7F	5'-TCTGGAAAGTCCCACCTCAG	7958	XMRV VP35
K1R	5'-AAGGCTTTATTGGGAACACG	8174	XMRV VP35

Table 4.5: Age, Clinical Parameters, and Geographical Locations of XMRV-Positive Prostate Cancer Cases

Patient	Age	Clinical stage ^a	PSA	Tumor grade	Pathological Stage	Hometown location
VP 29	42	T1c	0.7	6	Organ confined	Western PA
VP 35	63	T1c	6.3	7	Extracapsular extension	Northeastern OH
VP 42	61	T1c	5.9	6	Organ confined	Southwestern OH
VP 62	39	T1c	3.6	6	Organ confined	Northeastern OH
VP 79	66	T1c	4.6	6	Organ confined	Northeastern OH
VP 86	65	T1c	7.8	8	Organ confined	Northeastern OH
VP 88	63	T1c	10.6	6	Organ confined	Northeastern OH
VP 90	59	T2a	2.1	6	Organ confined	Northeastern OH
VP 184	60	T1c	5.6	6	Organ confined	Northeastern OH

^a1998 American Joint Commission on Cancer stage: T1c = nonpalpable tumor; T2a = palpable tumor confined to less than 1 lobe.

Chapter 5: DNA Microarrays in Viral Diagnostics

Citation: Urisman A, Chiu CY, Greenhow TL, Rouskin S, Fischer KF, Wright C, Drew L, Wang D, Weintrub PS, DeRisi JL, Ganem D. Use of DNA Microarrays for Viral Detection in Pediatric Acute Respiratory Tract Infections. 2006. In preparation.

Copyright: © 2006 Anatoly Urisman et al.

5.1 Abstract

Existing viral detection methods commonly used to identify viruses associated with acute respiratory tract infections (ARTI) include viral culture, direct fluorescence antibody (DFA), and PCR. These methods are limited in scope, and diagnostic methods for rapid and comprehensive viral detection are acutely needed. A DNA microarray platform for panviral detection has been previously described by our group in the context of several viral detection and discovery applications. In this study, we used the platform for clinical viral diagnostics in a large (n=194) blinded prospective study of pediatric ARTI, comparing the performance of the microarray to a standard seven-virus DFA panel. The rates of detection of the microarray were 128% higher overall, 27% higher for respiratory syncytial virus ($p < 0.05$), and 19% higher for the seven DFA viruses combined ($p < 0.03$), as compared to DFA. Almost half of all microarray-positive cases represented viruses not detectable by the DFA panel, including picornaviruses, coronaviruses, and human metapneumovirus. In addition, whereas DFA was unable to find any mixed infections, double-infections accounted for 8% of all microarray-positive cases. Given its broad

spectrum of detection and sensitivity rivaling that of DFA, the microarray platform described in this study should serve as a prototype for developing streamlined solutions for comprehensive and rapid diagnosis of ARTI in the clinical setting.

5.2 Introduction

Acute respiratory tract infections are the most frequent disease of humans and impose a major burden on society in direct and indirect costs [136]. In children, viruses are responsible for the majority of ARTI cases, and fewer than 10% of samples are caused by bacteria [137, 138]. However, even when the best methods for viral detection currently available are used in combination, a specific agent cannot be identified in 20 to 50% of ARTI cases [4, 5, 139, 140, 6, 141]. At present, no single method is capable of rapid and simultaneous screening for all known viral causes of ARTI, and a definitive viral diagnosis cannot be made in a significant number of cases unless multiple tests are performed.

Existing viral diagnostic methods are limited in sensitivity and scope. Viral culture has been the gold standard of viral diagnostics for several decades. However, many viruses are fastidious or unculturable, and even relatively fast shell-vial techniques require at least several days to complete. DFA testing requires only a few hours, and commercial kits targeting several common respiratory viruses are in wide use. Despite their popularity, DFA tests may suffer from low sensitivity [142-144], and are available only for a limited number of viruses. PCR testing is rapid and sensitive, and assays for many known viruses have been developed and are in widespread clinical use. Multiple studies

report higher rates of viral detection by PCR than by other methods [145, 140, 6]. Despite these advantages, most PCR tests target only one virus at a time, and routine testing for more than a handful of viruses is usually not feasible. A number of multiplex PCR strategies targeting up to a dozen viruses in a single assay have been proposed [146, 5, 140, 6, 147], but such tests tend to have higher rates of false positives and have been difficult to implement for widespread clinical use. Luminex xMAP technology allows simultaneous detection of a large number of distinct PCR amplicons from a single multiplex PCR reaction [148]. It has been successfully applied to detection and differentiation of pestiviruses [149, 150] and papillomaviruses [149, 150]. Although a vast improvement over traditional multiplex PCR methods, xMAP technology is currently limited to 100 amplicons, and like other multiplex PCR methods is not immune to false positives.

DNA microarrays have emerged as a successful strategy for viral detection [7, 151, 152, 41, 153-155]. We have previously described a microarray platform based on all available viral sequences and designed to detect both known and divergent viruses [8]. Currently, the platform employs a microarray comprised of ~20,000 70-mer oligonucleotides derived from ~1200 viral species representing all viral sequences, including animal and plant viruses, present in the NCBI Nucleotide database as of Fall 2004 [12]. The platform has been instrumental in identifying a novel coronavirus in a culture sample from a patient with Severe Acute Respiratory Syndrome [8] and a novel retrovirus in a subset of prostate tumors from patients with a mutation in the *RNASEL* gene[9]. The platform has

also been successful in detecting known respiratory viruses in culture samples as well as in clinical specimens from a limited number of patients [7, 11, 12].

In this study we sought to apply our platform to clinical viral diagnostics in a large blind prospective study of ARTI in children. Our primary goal was to compare the performance of the microarray to that of DFA. Based on the results of parallel testing of 194 samples by DFA and microarray, we conclude that the microarray, as compared to DFA, has superior sensitivity and a vastly improved ability to detect mixed infections.

5.3 Results

We examined a total of 194 nasopharyngeal aspirate (NPA) samples for presence of viruses using our microarray-based panviral detection platform. This prospective study included all consecutive samples sent for viral DFA testing from pediatric patients treated at the UCSF Hospital and Clinics during the period from December 2003 to March 2004. The DFA test used in the study was a standard seven-virus panel designed to detect influenza (Flu) A, Flu B, respiratory syncytial virus (RSV), parainfluenza (Para) 1, Para 2, Para 3, and adenovirus (Adeno). No other enrollment criteria were used for this study. No demographic or clinical data were available for this cohort at the time of writing of this report, but based on chart reviews of over 700 patients from a study conducted in the preceding year at the same site and using identical enrollment criteria [156], we know that 50% of samples were from children <1 year old, 20% from children 1–3 years old, and the remainder from children >3 years old. Approximately 80% of the NPA samples came from patients with ARTI symptoms, while the remaining samples were sent for

DFA due to other reasons, most often to evaluate fever of unknown etiology. Sixty percent of samples were from male and 40% from female patients. Fifteen percent of samples were from immunocompromised patients. Approximately 37% of samples were collected from outpatients, and the remaining samples came from patients admitted to the hospital.

The samples were processed by microarray in a blinded fashion, and the results were then compared to those obtained by DFA. Total RNA was extracted from each sample, amplified using random priming, labeled, and hybridized to the microarray using protocols described previously [8, 11]. Microarray analysis was carried out using E-Predict, a previously described algorithm for species identification based on observed microarray hybridization patterns [11] using a significance cutoff of $p < 0.001$ (see Materials and Methods). This stringent cutoff was chosen to ensure a low rate of false positives, albeit at the expense of lower sensitivity.

Figure 1 shows the spectrum and frequency of detection of different viruses by the two methods. Overall, while DFA identified a virus in only 21% of all samples, the microarray made a positive viral identification in 48%, more than doubling the overall rate of detection. Viruses not included in the DFA panel accounted for nearly half of all microarray positive identifications. Among these viruses, picornaviruses comprised the largest group (16% of all cases), which included 32 cases of single-virus infection and 2 cases of double infection. Microarray detected 16 cases of enterovirus (including 1 double infection), 14 cases of rhinovirus, 3 cases of parechovirus (including one double

infection), and, interestingly, 1 case of Aichi virus. Aichi virus is the first member of a new *Kobuvirus* genus in *Picornaviridae* associated with non-bacterial gastroenteritis; no cases in North America have been reported to date [157-159]. In addition to picornaviruses, microarray identified 6 cases of metapneumovirus and 2 cases of coronavirus (including 1 double infection). In 5 samples microarray detected sequences of viruses, which may represent innocuous viral flora, including TT virus frequently found in saliva [160], two bacteriophages, and sequences of several plant viruses commonly found in the gastrointestinal tract [160].

Table 1 compares microarray and DFA detection of the seven viruses included in the DFA panel. RSV accounted for the majority of positive cases using either method and was found in 37% of samples by microarray and 30% by DFA (Table 1). Microarray failed to identify RSV in 4 of 30 RSV DFA positive samples, while 12 samples positive for RSV by microarray were either negative (7 samples) or reported as “insufficient number of cells” (6 samples) by DFA. The overall rate of detection of RSV increased by 27% compared to DFA ($p < 0.03$ by χ^2 test). Results obtained for other viruses tested by DFA are more difficult to interpret due to the small numbers of positive cases by either method. Overall, at least one of the seven DFA viruses was identified by microarray in 25% (52) of the samples, compared to 21% (43) identified by DFA, corresponding to a 19% increase of positive detection of these viruses as compared to DFA ($p < 0.05$ by χ^2 test).

As shown in Table 2, microarray detected 7 cases of double infection (4% of all samples and 8% of microarray-positive samples). This included a Flu A/RSV and an RSV/Adeno infection reported as RSV-positive by DFA, and a Flu B/coronavirus infection that was DFA-negative. Interestingly, 6 of the 7 cases had RSV as one of the two viruses. In contrast, no cases of mixed infection were detected by DFA. The overall rate of detection of double infections by the microarray was similar to that reported by other studies employing conventional techniques [161, 5, 140, 141].

5.4 Discussion

Data presented above describe preliminary results in our ongoing study of viral agents associated with ARTI in children. At present, none of the microarray results have been verified by an independent method. In the future, we are planning to use PCR as the gold standard for validating the positives and for cases discordant with DFA. We are also planning to extend the study to include a larger number of samples in order to increase the statistical power of DFA vs. microarray comparisons, particularly for viruses with few positives identified. In addition, we are unable to evaluate any possible correlations between clinical parameters (e.g. upper vs. lower ARTI) and the identified viruses, because patient clinical data are not available at this time. Finally, although we did not target DNA viruses in this study, which explains lower than expected rates of detection of DNA viruses, in the future we are planning to evaluate the feasibility of using total nucleic acid instead of RNA as the starting material for microarray analysis.

Despite these limitations, our preliminary findings suggest that microarray has sensitivity that is similar to or better than that of DFA. RSV comprised the largest group of viruses for which both microarray and DFA data were available. Compared to DFA, microarrays increased the rate of RSV detection by 27% ($p < 0.03$), including in several samples with inconclusive DFA results due to low cellular content. Microarray rates of detection of other viruses tested by DFA (FluA, B; and Para1, 2, 3, and Adeno) were similar to those achieved by DFA, although too few positives for these viruses were identified by either method to make these comparisons statistically significant. Nonetheless, considering results for all seven viruses included in the DFA panel, microarray increased the rate of detection of these viruses by 19% compared to DFA ($p < 0.05$).

The main advantage of the microarray lies in its ability to screen for all known viruses simultaneously. In this study viruses not included in the DFA panel were responsible for nearly doubling the overall rate of positive detection by microarray compared to DFA. Picornaviruses accounted for the majority of these cases (16% of all tested cases). Other viruses in this category included metapneumovirus (3%) and coronavirus (<1%). DFA panels in current clinical use, including the one used in this study, do not test for these viruses. Screening for these viruses, in a typical clinical lab, would require at least three separate PCR tests.

Microarray was also better than DFA at detecting mixed viral infections. Double infections accounted for 8% of all microarray positive cases. Five of the 7 identified double infections included a virus not included in and therefore missed by the DFA

panel. Interestingly, two other cases, in which both viruses could in principle be detected by DFA, were reported as single-virus infections. Several studies have suggested that mixed infections are associated with greater severity of ARTI [161, 162]. Thus, timely detection of mixed infections will promote more informed clinical decisions and may result in better patient outcomes.

Given these results, we believe that the microarray-based approach to viral detection has real potential to deliver a robust diagnostic method for rapid and simultaneous detection of all known viruses. Even though microarrays are becoming less expensive, and affordable ready-to-order microarray solutions are already available for limited viral diagnostic applications (e.g [163]), more comprehensive low-cost microarray systems will need to be developed for this method to be practical in routine clinical diagnostics. Streamlining of all steps in the processing of samples (from nucleic acid extraction to hybridization, scanning, to data analysis) will be another important prerequisite. Given that much work on this front has already been done in other commercial microarray applications, we anticipate that single-button solutions for microarray-based viral diagnostics may become available in the near future.

5.5 Materials and Methods

All patient samples were collected according to protocols approved by the UCSF Committee on Human Research. Consecutive NPA samples sent for DFA to the UCSF clinical laboratory were analyzed with Light Diagnostics Respiratory DFA Viral Screening and Identification Kit (Cat. No. 3137; Chemicon International, Temecula CA,

USA). kit according to the established guidelines. Following DFA, remaining sample material was transferred by the clinical laboratory staff into a sterile 14 mL conical tube pre-labeled with a pseudo-ID unlinked from patient identifying information and DFA results. The samples were immediately frozen and stored at -80°C and later transferred on dry ice to our laboratory, where the samples were stored at -80°C until analyzed by microarray.

For microarray analysis, frozen NPA samples were thawed, and 200 μL aliquots were used to extract RNA using RNeasy Mini Kit (Qiagen USA, Valencia, CA, USA) as follows. 750 μL of RLT buffer containing 1% 2-mercaptoethanol were added to each sample and mixed. 1 mL of 100% ethanol was added next, and the resulting mixture was applied to the columns in three 650 μL aliquots. The remaining steps were carried out according to the manufacturer's protocol, including on-column DNase digest. RNA was eluted from the columns with 30 μL of nuclease-free water, and 4 μL were used for amplification and hybridization. Microarrays used in this study were identical to those previously described ([12]; NCBI GEO platform GPL3429). RNA samples were amplified and labeled using a modified Round A/B random PCR method and hybridized to the microarrays as reported previously (Protocol S1 in [8]) with the following modifications. Round A, B, and C reaction volumes were scaled down to 10, 50, and 50 μL , respectively. Powerscript reverse transcriptase (Clontech, Mountain View, CA, USA) was used in Round A with a single incubation at 42°C for 1 hr, and KlenTaq LA polymerase (Sigma-Aldrich USA, St. Louis, MO, USA) was used for Rounds B and C. Number of PCR cycles for Rounds B and C were decreased to 20 and 15, respectively. Dye coupling was carried out as before,

except amplified RNA samples were labeled with Cy3 and Probe70 with Cy5.

Microarrays were scanned with an Axon 4000B scanner and gridded using Axon GenePix 6 software (Axon Instruments, Union City, CA, USA).

Microarray data analysis was carried out in two stages. First, all microarrays were analyzed by E-Predict using the optimal settings and energy profile matrix described previously [11]. A significance cutoff of $p < 0.05$ was used in this preliminary stage to identify microarrays with statistically significant viral hybridization patterns (96 of 194). The remaining 98 microarrays were assumed to be negative and were used to generate an optimal set of oligonucleotide intensity weights as follows. We evaluated a set of functions with the general equation

$$w = \left[\left(1 - \frac{i-a}{b-a} \right)^p \right]^{\frac{1}{p}} \text{ if } a < i < b; \quad w = 1 \text{ if } a < i < b; \quad \text{or } w = 0 \text{ if } i > b,$$

where w is a weight (value from 0 to 1) for a given oligonucleotide, i is the median of sum-normalized intensities of that oligonucleotide across the 98 negative microarrays.

We set the lower boundary a to the median of medians of the sum-normalized intensities of all oligonucleotides. The upper boundary condition b was expressed as $b = a + c\sigma$, where c was a constant, and σ was the standard deviation of sum-normalized intensities of the oligonucleotide across the 98 negative microarrays. We evaluated a total of 40 functions corresponding to all possible combinations of c (0.01; 0.05; 0.1; 0.15; 0.25; 0.5) and p (0.5; 0.67; 1; 1.5; 2). Each function was used to make a set of oligonucleotide weights, and the 40 resulting weight sets were evaluated for their ability to discriminate presumed positive from presumed negative E-Predict predictions. For this purpose we

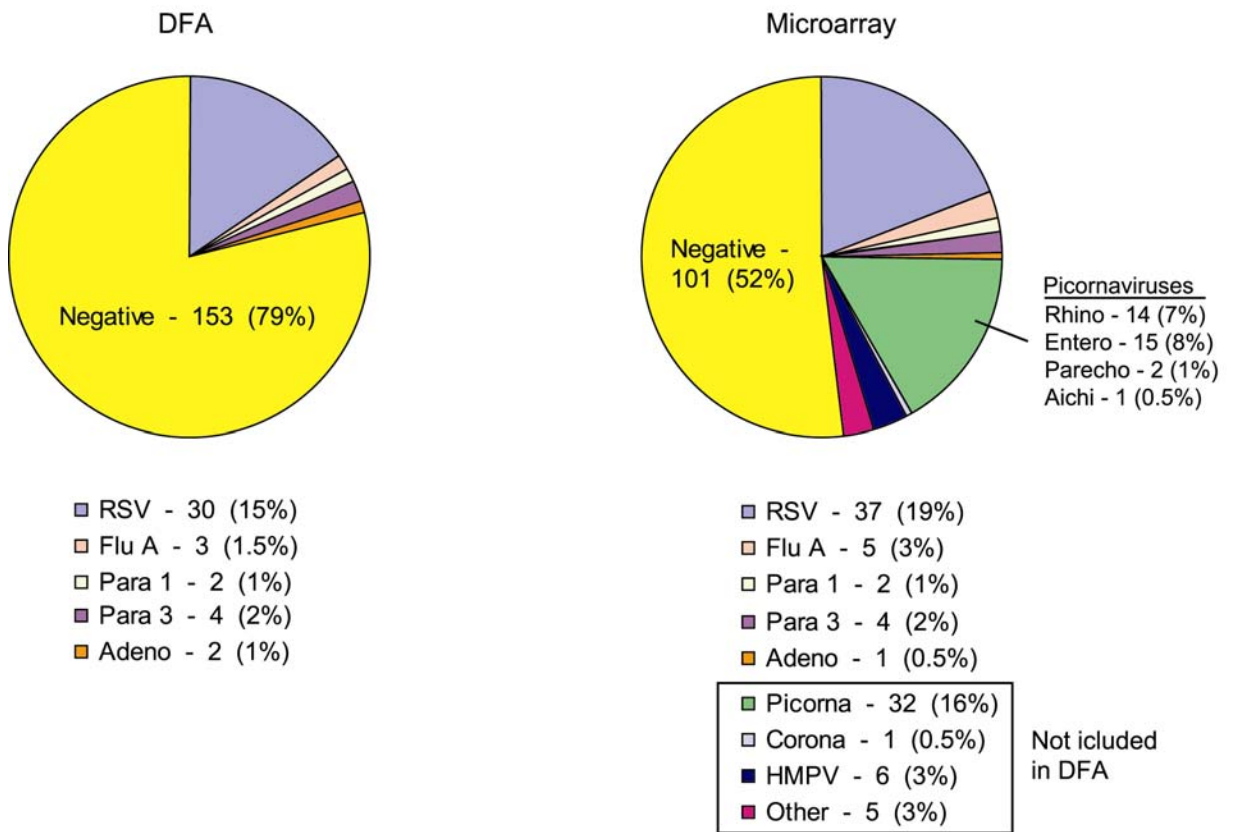
chose 6 random RSV, 6 random rhinovirus, 4 Para 3, and 2 Flu A-positive microarrays. These microarrays were analyzed by E-Predict using the most recent energy profile matrix (April 2006) containing updated profiles corresponding to 1225 NCBI Reference viral genomes. Observed oligonucleotide intensities were multiplied by the corresponding weights, and performance of each weight set was evaluated using E-Predict profile separation statistic described previously [11]. The optimal performance was achieved with weights corresponding to $c = 0.1$ and $p = 2$. These weights were used to generate negative null distributions for E-Predict significance estimation based on the set of the 98 negative microarrays mentioned above, using the same procedure as describe previously [11], except no trimming of high scores was performed here.

Final microarray virus determinations for all 194 microarrays were made by E-Predict using the optimal oligonucleotide intensity weights and the corresponding null distributions described above. A microarray was considered positive for a given virus if the corresponding energy profile attained a significance score of $p < 0.001$ in either first or second E-Predict iteration. After microarray results were finalized, sample pseudo-IDs were linked back to DFA results, and the results obtained by the two methods were compared.

5.6 Acknowledgements

This work was supported by a Genentech Graduate Fellowship (AU) and grants from the Sandler Program for Asthma Research (JLDR), and the Doris Duke Charitable Foundation (JLDR).

Figure 5.1: Viruses Detected by DFA and Microarray



Numbers of cases positive for respiratory syncytial virus (RSV), parainfluenza (Para) 1, Para 3, adenovirus (Adeno), picornaviruses (picorna), coronaviruses (corona), and human metapneumovirus (HMPV) are plotted as a proportion of all cases (n=194). Viruses not tested by DFA are boxed. Further breakdown for picornaviruses includes: rhinoviruses (Rhino), enteroviruses (Entero), parechovirus (Parecho), and Aichi virus. Category “Other” includes several ubiquitous viruses frequently found in healthy people (see text). For cases of double infections, only the virus with a more statistically significant microarray prediction is listed.

Table 5.1: Detection of Seven Common Respiratory Viruses by DFA and Microarray

	RSV	FluA	FluB	Para1	Para2	Para3	Adeno
DFA+ Array+	26	2	0	1	0	4	1
DFA- Array+	12	3	1	1	0	0	1
DFA+ Array-	4	1	0	1	0	0	1

Table 5.2: Detection of Double Infections by Microarray and DFA

Microarray Virus 1	Microarray Virus 2	Number of Cases	DFA
Flu A	RSV	1	RSV
Flu B	Coronavirus	1	Negative
RSV	Adenovirus	1	RSV
RSV	Coronavirus	1	RSV
RSV	Enterovirus	1	RSV
RSV	Parechovirus	1	RSV
RSV	TT virus	1	RSV

Chapter 6: Concluding Remarks

Copyright: © 2006 Anatoly Urisman et al.

6.1 Pros and Cons of Using DNA Microarrays for Viral

Detection

Viral detection by DNA microarrays has emerged as an important new approach to viral discovery and viral diagnostics. Many groups have developed microarray platforms for applications ranging from single virus detection to larger taxonomic groups [7, 151, 152, 41, 153, 155]. We have developed the first platform for panviral detection [8] and have applied it to a range of discovery [8, 9] and diagnostic [7] [12] [164] applications.

Although it is still in the development and testing stages, this new technology holds great promise to improve our ability to detect undiscovered viruses, emerging strains of known viruses, as well as viruses that are currently known.

There are several important advantages to using microarrays for viral detection. The method circumvents the need to isolate viruses by culture and overcomes the limited scope of PCR and serological techniques. It also allows for adjustable levels of sequence redundancy and taxonomic resolution by varying oligonucleotide selection. In addition the method allows for high throughput and real time testing, particularly in diagnostic applications involving detection of known viruses. In discovery applications, microarray serves as a starting point for recovering viral sequences by identifying target genomic regions to focus on during recovery efforts.

Several important limitations to using microarrays for viral detection and discovery should not be overlooked. The most important limitation of the approach is its reliance on known viral sequences. Although most of the novel viruses discovered in the last decade share homology with previously known viruses [165-167, 13, 168, 169], viruses lacking such homology cannot be detected by the method. Other existing and likely surmountable limitations of the method are its relatively high cost, need for specialized equipment and access to computational resources. We expect these limitations to become less pronounced as streamlined versions of the technology become available through efforts of academic as well as commercial institutions.

6.2 Future Directions

Although application of DNA microarrays to viral detection has already demonstrated utility for viral discovery [8, 9] and viral diagnostics [7, 151, 152, 41, 155], much can be done to improve the technology itself and to extend its usefulness to a wider range of applications. These efforts fall into three main categories: (i) microarray design; (ii) sample processing; and (iii) data analysis.

Microarray design will continue to improve through new technologies for microarray fabrication and oligonucleotide synthesis. The number of features that can be printed on a single microarray will continue to increase allowing greater sequence representation and therefore greater spectrum and sensitivity of detection as well as improved discrimination. In addition, improvements in algorithms for oligonucleotide selection will

enable more precision in achieving desired levels of coverage, sequence redundancy, and virus discrimination. We also expect that mass-produced microarrays designed for a specific task, like respiratory viral diagnostics, will be developed and will make their way into clinical practice in the near future.

Optimal sample processing is critically important for achieving the best possible performance of the method, particularly for samples with high cellular, low viral contents. Although incremental improvements to protocols for nucleic acid extraction, amplification, and labeling are being introduced, most making use of new more processive enzymes, new approaches and fresh ideas in this area are greatly needed. One possible direction is to use differential hybridization or amplification techniques [155] as a sample preprocessing step before microarray analysis.

Although some progress has been made in data analysis approaches to species identification based on microarray patterns [11], further work on improved algorithms for noise recognition and filtering, pattern deconvolution, and significance estimation will be an important focus of future investigations. Our laboratory is particularly interested in developing sensitive data analysis techniques for detecting novel very divergent viruses, where few assumptions can be made regarding distributions of positive oligonucleotides or patterns of conservation with known viral sequences.

As we move forward with new advances in DNA microarray-based viral detection, it is important to consider how this research benefits the greater community. Discovery of

new viruses or introduction of improved viral diagnostics may have significant implications on clinical practice and public health policy. It is therefore important that groups contributing to this research, particularly publicly founded institutions, participate in sharing of information and resources through timely publications in peer-reviewed journals and by providing access to oligonucleotide sequences and software tools.

References

1. Lee WM: Acute liver failure in the United States. *Semin Liver Dis* 2003, 23(3):217-226.
2. Casas I, Pozo F, Trallero G, Echevarria JM, Tenorio A: Viral diagnosis of neurological infection by RT multiplex PCR: a search for entero- and herpesviruses in a prospective study. *J Med Virol* 1999, 57(2):145-151.
3. Glaser CA, Gilliam S, Schnurr D, Forghani B, Honarmand S, Khetsuriani N, Fischer M, Cossen CK, Anderson LJ: In search of encephalitis etiologies: diagnostic challenges in the California Encephalitis Project, 1998-2000. *Clin Infect Dis* 2003, 36(6):731-742.
4. Monto AS: Epidemiology of viral respiratory infections. *Am J Med* 2002, 112 Suppl 6A:4S-12S.
5. Erdman DD, Weinberg GA, Edwards KM, Walker FJ, Anderson BC, Winter J, Gonzalez M, Anderson LJ: GeneScan reverse transcription-PCR assay for detection of six common respiratory viruses in young children hospitalized with acute respiratory illness. *J Clin Microbiol* 2003, 41(9):4298-4303.
6. Syrnis MW, Whiley DM, Thomas M, Mackay IM, Williamson J, Siebert DJ, Nissen MD, Sloots TP: A sensitive, specific, and cost-effective multiplex reverse transcriptase-PCR assay for the detection of seven common respiratory viruses in respiratory samples. *J Mol Diagn* 2004, 6(2):125-131.
7. 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.

8. 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.
9. 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.
10. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998, 95(25):14863-14868.
11. 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.
12. Chiu C, Rouskin S, Koshy A, Urisman A, Fischer KF, Yagi S, Schnurr D, Tompkins LS, Blackburn BG, Merker JD *et al*: Microarray Diagnosis of Human Parainfluenza 4 Infection Associated with Respiratory Failure in an Immunocompetent Adult. *Clin Infect Dis* 2006, Submitted.
13. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH *et al*: Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 2003, 300(5624):1394-1399.
14. Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M: Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science* 1989, 244(4902):359-362.
15. Nichol ST, Spiropoulou CF, Morzunov S, Rollin PE, Ksiazek TG, Feldmann H, Sanchez A, Childs J, Zaki S, Peters CJ: Genetic identification of a hantavirus

- associated with an outbreak of acute respiratory illness. *Science* 1993, 262(5135):914-917.
16. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, Moore PS: Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* 1994, 266(5192):1865-1869.
 17. Muerhoff AS, Leary TP, Desai SM, Mushahwar IK: Amplification and subtraction methods and their application to the discovery of novel human viruses. *J Med Virol* 1997, 53(1):96-103.
 18. Kellam P: Molecular identification of novel viruses. *Trends Microbiol* 1998, 6(4):160-165.
 19. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim W *et al*: A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* 2003, 348(20):1953-1966.
 20. Jonassen CM, Jonassen TO, Grinde B: A common RNA motif in the 3' end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. *J Gen Virol* 1998, 79 (Pt 4):715-718.
 21. Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattri J, Asano JK, Barber SA, Chan SY *et al*: The Genome sequence of the SARS-associated coronavirus. *Science* 2003, 300(5624):1399-1404.
 22. Riesenfeld CS, Schloss PD, Handelsman J: Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 2004, 38:525-552.
 23. Eyers L, George I, Schuler L, Stenuit B, Agathos SN, El Fantroussi S: Environmental genomics: exploring the unmined richness of microbes to degrade xenobiotics. *Appl Microbiol Biotechnol* 2004, 66(2):123-130.
 24. Rodriguez-Valera F: Environmental genomics, the big picture? *FEMS Microbiol Lett* 2004, 231(2):153-158.

25. Schloss PD, Handelsman J: Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* 2003, 14(3):303-310.
26. 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.
27. 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.
28. 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.
29. 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 R Soc Lond B Biol Sci* 2004, 271(1539):565-574.
30. Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF: Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* 2004, 430(6999):551-554.
31. van der Wielen PW, Bolhuis H, Borin S, Daffonchio D, Corselli C, Giuliano L, D'Auria G, de Lange GJ, Huebner A, Varnavas SP *et al*: The enigma of prokaryotic life in deep hypersaline anoxic basins. *Science* 2005, 307(5706):121-123.
32. Liles MR, Manske BF, Bintrim SB, Handelsman J, Goodman RM: A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol* 2003, 69(5):2684-2691.
33. Woese CR: Bacterial evolution. *Microbiol Rev* 1987, 51(2):221-271.

34. Brady SF, Chao CJ, Clardy J: New natural product families from an environmental DNA (eDNA) gene cluster. *J Am Chem Soc* 2002, 124(34):9968-9969.
35. 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.
36. Henne A, Schmitz RA, Bomeke M, Gottschalk G, Daniel R: Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl Environ Microbiol* 2000, 66(7):3113-3116.
37. Bodrossy L, Sessitsch A: Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol* 2004, 7(3):245-254.
38. Zhou J: Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* 2003, 6(3):288-294.
39. Cook KL, Saylor GS: Environmental application of array technology: promise, problems and practicalities. *Curr Opin Biotechnol* 2003, 14(3):311-318.
40. Sengupta S, Onodera K, Lai A, Melcher U: Molecular detection and identification of influenza viruses by oligonucleotide microarray hybridization. *J Clin Microbiol* 2003, 41(10):4542-4550.
41. Klaassen CH, Prinsen CF, de Valk HA, Horrevorts AM, Jeunink MA, Thunnissen FB: DNA microarray format for detection and subtyping of human papillomavirus. *J Clin Microbiol* 2004, 42(5):2152-2160.
42. Lin B, Vora GJ, Thach D, Walter E, Metzgar D, Tibbetts C, Stenger DA: Use of oligonucleotide microarrays for rapid detection and serotyping of acute respiratory disease-associated adenoviruses. *J Clin Microbiol* 2004, 42(7):3232-3239.

43. Lemarchand K, Masson L, Brousseau R: Molecular biology and DNA microarray technology for microbial quality monitoring of water. *Crit Rev Microbiol* 2004, 30(3):145-172.
44. Rhee SK, Liu X, Wu L, Chong SC, Wan X, Zhou J: Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays. *Appl Environ Microbiol* 2004, 70(7):4303-4317.
45. Ivnitski D, O'Neil DJ, Gattuso A, Schlicht R, Calidonna M, Fisher R: Nucleic acid approaches for detection and identification of biological warfare and infectious disease agents. *Biotechniques* 2003, 35(4):862-869.
46. Varmus HE: Form and function of retroviral proviruses. *Science* 1982, 216(4548):812-820.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403-410.
48. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J, Jr.: Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry* 1999, 38(12):3468-3477.
49. 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.
50. Ledford RM, Patel NR, Demenczuk TM, Watanyar A, Herbertz T, Collett MS, Pevear DC: VP1 sequencing of all human rhinovirus serotypes: insights into genus phylogeny and susceptibility to antiviral capsid-binding compounds. *J Virol* 2004, 78(7):3663-3674.
51. Blomqvist S, Savolainen C, Raman L, Roivainen M, Hovi T: Human rhinovirus 87 and enterovirus 68 represent a unique serotype with rhinovirus and enterovirus features. *J Clin Microbiol* 2002, 40(11):4218-4223.

52. Savolainen C, Blomqvist S, Mulders MN, Hovi T: Genetic clustering of all 102 human rhinovirus prototype strains: serotype 87 is close to human enterovirus 70. *J Gen Virol* 2002, 83(Pt 2):333-340.
53. Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J, Schleifer KH, Wagner M: Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol* 2002, 68(10):5064-5081.
54. Cho JC, Tiedje JM: Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. *Appl Environ Microbiol* 2001, 67(8):3677-3682.
55. Eisen MB, Brown PO: DNA arrays for analysis of gene expression. *Methods Enzymol* 1999, 303:179-205.
56. Bohlander SK, Espinosa R, 3rd, Le Beau MM, Rowley JD, Diaz MO: A method for the rapid sequence-independent amplification of microdissected chromosomal material. *Genomics* 1992, 13(4):1322-1324.
57. Celander D, Hsu BL, Haseltine WA: Regulatory elements within the murine leukemia virus enhancer regions mediate glucocorticoid responsiveness. *J Virol* 1988, 62(4):1314-1322.
58. Bruland T, Lavik LA, Dai HY, Dalen A: A glucocorticoid response element in the LTR U3 region of Friend murine leukaemia virus variant FIS-2 enhances virus production in vitro and is a major determinant for sex differences in susceptibility to FIS-2 infection in vivo. *J Gen Virol* 2003, 84(Pt 4):907-916.
59. Miksicek R, Heber A, Schmid W, Danesch U, Posseckert G, Beato M, Schutz G: Glucocorticoid responsiveness of the transcriptional enhancer of Moloney murine sarcoma virus. *Cell* 1986, 46(2):283-290.
60. Shapiro SS, Wilk MB: An analysis of variance test for normality (complete samples). *Biometrika* 1965, 52(3,4):591-611.

61. DeFranco D, Yamamoto KR: Two different factors act separately or together to specify functionally distinct activities at a single transcriptional enhancer. *Mol Cell Biol* 1986, 6(4):993-1001.
62. Bruland T, Dai HY, Lavik LA, Kristiansen LI, Dalen A: Gender-related differences in susceptibility, early virus dissemination and immunosuppression in mice infected with Friend murine leukaemia virus variant FIS-2. *J Gen Virol* 2001, 82(Pt 8):1821-1827.
63. Saldanha AJ: Java Treeview--extensible visualization of microarray data. *Bioinformatics* 2004, 20(17):3246-3248.
64. Adam MA, Miller AD: Identification of a signal in a murine retrovirus that is sufficient for packaging of nonretroviral RNA into virions. *J Virol* 1988, 62(10):3802-3806.
65. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997, 25(24):4876-4882.
66. Fisher J, Goff SP: Mutational analysis of stem-loops in the RNA packaging signal of the Moloney murine leukemia virus. *Virology* 1998, 244(1):133-145.
67. Kerr IM, Brown RE: pppA₂'p5'A₂'p5'A: an inhibitor of protein synthesis synthesized with an enzyme fraction from interferon-treated cells. *Proc Natl Acad Sci U S A* 1978, 75(1):256-260.
68. Zhou A, Hassel BA, Silverman RH: Expression cloning of 2-5A-dependent RNAase: a uniquely regulated mediator of interferon action. *Cell* 1993, 72(5):753-765.
69. Dong B, Silverman RH: 2-5A-dependent RNase molecules dimerize during activation by 2-5A. *J Biol Chem* 1995, 270(8):4133-4137.

70. Zhou A, Paranjape J, Brown TL, Nie H, Naik S, Dong B, Chang A, Trapp B, Fairchild R, Colmenares C *et al*: Interferon action and apoptosis are defective in mice devoid of 2',5'-oligoadenylate-dependent RNase L. *Embo J* 1997, 16(21):6355-6363.
71. Flodstrom-Tullberg M, Hultcrantz M, Stotland A, Maday A, Tsai D, Fine C, Williams B, Silverman R, Sarvetnick N: RNase L and double-stranded RNA-dependent protein kinase exert complementary roles in islet cell defense during coxsackievirus infection. *J Immunol* 2005, 174(3):1171-1177.
72. Castelli JC, Hassel BA, Wood KA, Li XL, Amemiya K, Dalakas MC, Torrence PF, Youle RJ: A study of the interferon antiviral mechanism: apoptosis activation by the 2-5A system. *J Exp Med* 1997, 186(6):967-972.
73. Li G, Xiang Y, Sabapathy K, Silverman RH: An apoptotic signaling pathway in the interferon antiviral response mediated by RNase L and c-Jun NH2-terminal kinase. *J Biol Chem* 2004, 279(2):1123-1131.
74. Malathi K, Paranjape JM, Ganapathi R, Silverman RH: HPC1/RNASEL mediates apoptosis of prostate cancer cells treated with 2',5'-oligoadenylates, topoisomerase I inhibitors, and tumor necrosis factor-related apoptosis-inducing ligand. *Cancer Res* 2004, 64(24):9144-9151.
75. Xiang Y, Wang Z, Murakami J, Plummer S, Klein EA, Carpten JD, Trent JM, Isaacs WB, Casey G, Silverman RH: Effects of RNase L mutations associated with prostate cancer on apoptosis induced by 2',5'-oligoadenylates. *Cancer Res* 2003, 63(20):6795-6801.
76. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, Faruque M, Moses T, Ewing C, Gillanders E *et al*: Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. *Nat Genet* 2002, 30(2):181-184.

77. Casey G, Neville PJ, Plummer SJ, Xiang Y, Krumroy LM, Klein EA, Catalona WJ, Nupponen N, Carpten JD, Trent JM *et al*: RNASEL Arg462Gln variant is implicated in up to 13% of prostate cancer cases. *Nat Genet* 2002, 32(4):581-583.
78. Rennert H, Bercovich D, Hubert A, Abeliovich D, Rozovsky U, Bar-Shira A, Soloviov S, Schreiber L, Matzkin H, Rennert G *et al*: A novel founder mutation in the RNASEL gene, 471delAAAG, is associated with prostate cancer in Ashkenazi Jews. *Am J Hum Genet* 2002, 71(4):981-984.
79. Rokman A, Ikonen T, Seppala EH, Nupponen N, Autio V, Mononen N, Bailey-Wilson J, Trent J, Carpten J, Matikainen MP *et al*: Germline alterations of the RNASEL gene, a candidate HPC1 gene at 1q25, in patients and families with prostate cancer. *Am J Hum Genet* 2002, 70(5):1299-1304.
80. Nelson WG, De Marzo AM, Isaacs WB: Prostate cancer. *N Engl J Med* 2003, 349(4):366-381.
81. Carter BS, Bova GS, Beaty TH, Steinberg GD, Childs B, Isaacs WB, Walsh PC: Hereditary prostate cancer: epidemiologic and clinical features. *J Urol* 1993, 150(3):797-802.
82. Silverman RH: Implications for RNase L in prostate cancer biology. *Biochemistry* 2003, 42(7):1805-1812.
83. Downing SR, Hennessy KT, Abe M, Manola J, George DJ, Kantoff PW: Mutations in ribonuclease L gene do not occur at a greater frequency in patients with familial prostate cancer compared with patients with sporadic prostate cancer. *Clin Prostate Cancer* 2003, 2(3):177-180.
84. Wiklund F, Jonsson BA, Brookes AJ, Stromqvist L, Adolfsson J, Emanuelsson M, Adami HO, Augustsson-Balter K, Gronberg H: Genetic analysis of the RNASEL gene in hereditary, familial, and sporadic prostate cancer. *Clin Cancer Res* 2004, 10(21):7150-7156.

85. Maier C, Haeusler J, Herkommer K, Vesovic Z, Hoegel J, Vogel W, Paiss T: Mutation screening and association study of RNASEL as a prostate cancer susceptibility gene. *Br J Cancer* 2005, 92(6):1159-1164.
86. Clark SP, Mak TW: Complete nucleotide sequence of an infectious clone of Friend spleen focus-forming provirus: gp55 is an envelope fusion glycoprotein. *Proc Natl Acad Sci U S A* 1983, 80(16):5037-5041.
87. Raisch KP, Pizzato M, Sun HY, Takeuchi Y, Cashdollar LW, Grossberg SE: Molecular cloning, complete sequence, and biological characterization of a xenotropic murine leukemia virus constitutively released from the human B-lymphoblastoid cell line DG-75. *Virology* 2003, 308(1):83-91.
88. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA *et al*: Database resources of the National Center for Biotechnology. *Nucleic Acids Res* 2003, 31(1):28-33.
89. Wills NM, Gesteland RF, Atkins JF: Evidence that a downstream pseudoknot is required for translational read-through of the Moloney murine leukemia virus gag stop codon. *Proc Natl Acad Sci U S A* 1991, 88(16):6991-6995.
90. Shinnick TM, Lerner RA, Sutcliffe JG: Nucleotide sequence of Moloney murine leukaemia virus. *Nature* 1981, 293(5833):543-548.
91. Herr W: Nucleotide sequence of AKV murine leukemia virus. *J Virol* 1984, 49(2):471-478.
92. O'Neill RR, Buckler CE, Theodore TS, Martin MA, Repaske R: Envelope and long terminal repeat sequences of a cloned infectious NZB xenotropic murine leukemia virus. *J Virol* 1985, 53(1):100-106.
93. Perryman S, Nishio J, Chesebro B: Complete nucleotide sequence of Friend murine leukemia virus, strain FB29. *Nucleic Acids Res* 1991, 19(24):6950.

94. Sijts EJ, Leupers CJ, Mengede EA, Loenen WA, van den Elsen PJ, Melief CJ: Cloning of the MCF1233 murine leukemia virus and identification of sequences involved in viral tropism, oncogenicity and T cell epitope formation. *Virus Res* 1994, 34(3):339-349.
95. Antoine M, Wegmann B, Kiefer P: Envelope and long terminal repeat sequences of an infectious murine leukemia virus from a human SCLC cell line: implications for gene transfer. *Virus Genes* 1998, 17(2):157-168.
96. Coffin JM, Hughes SH, Varmus HE: Retroviruses. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997.
97. Battini JL, Heard JM, Danos O: Receptor choice determinants in the envelope glycoproteins of amphotropic, xenotropic, and polytropic murine leukemia viruses. *J Virol* 1992, 66(3):1468-1475.
98. Fass D, Davey RA, Hamson CA, Kim PS, Cunningham JM, Berger JM: Structure of a murine leukemia virus receptor-binding glycoprotein at 2.0 angstrom resolution. *Science* 1997, 277(5332):1662-1666.
99. Taylor CS, Lavillette D, Marin M, Kabat D: Cell surface receptors for gammaretroviruses. *Curr Top Microbiol Immunol* 2003, 281:29-106.
100. Khan AS, Martin MA: Endogenous murine leukemia proviral long terminal repeats contain a unique 190-base-pair insert. *Proc Natl Acad Sci U S A* 1983, 80(9):2699-2703.
101. Battini JL, Rasko JE, Miller AD: A human cell-surface receptor for xenotropic and polytropic murine leukemia viruses: possible role in G protein-coupled signal transduction. *Proc Natl Acad Sci U S A* 1999, 96(4):1385-1390.
102. Taylor CS, Nouri A, Lee CG, Kozak C, Kabat D: Cloning and characterization of a cell surface receptor for xenotropic and polytropic murine leukemia viruses. *Proc Natl Acad Sci U S A* 1999, 96(3):927-932.

103. Yang YL, Guo L, Xu S, Holland CA, Kitamura T, Hunter K, Cunningham JM: Receptors for polytropic and xenotropic mouse leukaemia viruses encoded by a single gene at Rmc1. *Nat Genet* 1999, 21(2):216-219.
104. Temin HM: Structure, variation and synthesis of retrovirus long terminal repeat. *Cell* 1981, 27(1 Pt 2):1-3.
105. Speck NA, Baltimore D: Six distinct nuclear factors interact with the 75-base-pair repeat of the Moloney murine leukemia virus enhancer. *Mol Cell Biol* 1987, 7(3):1101-1110.
106. Thomas CY, Coppola MA, Holland CA, Massey AC: Oncogenicity and U3 region sequences of class II recombinant MuLVs of CWD mice. *Virology* 1990, 176(1):166-177.
107. Tomonaga K, Coffin JM: Structures of endogenous nonectropic murine leukemia virus (MLV) long terminal repeats in wild mice: implication for evolution of MLVs. *J Virol* 1999, 73(5):4327-4340.
108. Berlioz C, Darlix JL: An internal ribosomal entry mechanism promotes translation of murine leukemia virus gag polyprotein precursors. *J Virol* 1995, 69(4):2214-2222.
109. Vagner S, Waysbort A, Marena M, Gensac MC, Amalric F, Prats AC: Alternative translation initiation of the Moloney murine leukemia virus mRNA controlled by internal ribosome entry involving the p57/PTB splicing factor. *J Biol Chem* 1995, 270(35):20376-20383.
110. Prats AC, De Billy G, Wang P, Darlix JL: CUG initiation codon used for the synthesis of a cell surface antigen coded by the murine leukemia virus. *J Mol Biol* 1989, 205(2):363-372.
111. Fan H, Chute H, Chao E, Feuerman M: Construction and characterization of Moloney murine leukemia virus mutants unable to synthesize glycosylated gag polyprotein. *Proc Natl Acad Sci U S A* 1983, 80(19):5965-5969.

112. Schwartzberg P, Colicelli J, Goff SP: Deletion mutants of Moloney murine leukemia virus which lack glycosylated gag protein are replication competent. *J Virol* 1983, 46(2):538-546.
113. Chun R, Fan H: Recovery of Glycosylated gag Virus from Mice Infected with a Glycosylated gag-Negative Mutant of Moloney Murine Leukemia Virus. *J Biomed Sci* 1994, 1(4):218-223.
114. Corbin A, Prats AC, Darlix JL, Sitbon M: A nonstructural gag-encoded glycoprotein precursor is necessary for efficient spreading and pathogenesis of murine leukemia viruses. *J Virol* 1994, 68(6):3857-3867.
115. Portis JL, Fujisawa R, McAtee FJ: The glycosylated gag protein of MuLV is a determinant of neuroinvasiveness: analysis of second site revertants of a mutant MuLV lacking expression of this protein. *Virology* 1996, 226(2):384-392.
116. Fujisawa R, McAtee FJ, Zirbel JH, Portis JL: Characterization of glycosylated Gag expressed by a neurovirulent murine leukemia virus: identification of differences in processing in vitro and in vivo. *J Virol* 1997, 71(7):5355-5360.
117. Munk C, Prassolov V, Rodenburg M, Kalinin V, Lohler J, Stocking C: 10A1-MuLV but not the related amphotropic 4070A MuLV is highly neurovirulent: importance of sequences upstream of the structural Gag coding region. *Virology* 2003, 313(1):44-55.
118. Fujisawa R, McAtee FJ, Wehrly K, Portis JL: The neuroinvasiveness of a murine retrovirus is influenced by a dileucine-containing sequence in the cytoplasmic tail of glycosylated Gag. *J Virol* 1998, 72(7):5619-5625.
119. Bracho MA, Moya A, Barrio E: Contribution of Taq polymerase-induced errors to the estimation of RNA virus diversity. *J Gen Virol* 1998, 79 (Pt 12):2921-2928.
120. Wernert N, Seitz G, Achtstatter T: Immunohistochemical investigation of different cytokeratins and vimentin in the prostate from the fetal period up to adulthood and in prostate carcinoma. *Pathol Res Pract* 1987, 182(5):617-626.

121. Chesebro B, Britt W, Evans L, Wehrly K, Nishio J, Cloyd M: Characterization of monoclonal antibodies reactive with murine leukemia viruses: use in analysis of strains of friend MCF and Friend ecotropic murine leukemia virus. *Virology* 1983, 127(1):134-148.
122. Tomonaga K, Coffin JM: Structure and distribution of endogenous nonectropic murine leukemia viruses in wild mice. *J Virol* 1998, 72(10):8289-8300.
123. Albritton LM, Tseng L, Scadden D, Cunningham JM: A putative murine ecotropic retrovirus receptor gene encodes a multiple membrane-spanning protein and confers susceptibility to virus infection. *Cell* 1989, 57(4):659-666.
124. Kim JW, Closs EI, Albritton LM, Cunningham JM: Transport of cationic amino acids by the mouse ecotropic retrovirus receptor. *Nature* 1991, 352(6337):725-728.
125. Wang H, Kavanaugh MP, North RA, Kabat D: Cell-surface receptor for ecotropic murine retroviruses is a basic amino-acid transporter. *Nature* 1991, 352(6337):729-731.
126. Gifford R, Tristem M: The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* 2003, 26(3):291-315.
127. Tlsty TD, Hein PW: Know thy neighbor: stromal cells can contribute oncogenic signals. *Curr Opin Genet Dev* 2001, 11(1):54-59.
128. Bhowmick NA, Neilson EG, Moses HL: Stromal fibroblasts in cancer initiation and progression. *Nature* 2004, 432(7015):332-337.
129. Olumi AF, Grossfeld GD, Hayward SW, Carroll PR, Tlsty TD, Cunha GR: Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. *Cancer Res* 1999, 59(19):5002-5011.
130. Gordon D, Abajian C, Green P: Consed: a graphical tool for sequence finishing. *Genome Res* 1998, 8(3):195-202.

131. Stoye JP, Coffin JM: The four classes of endogenous murine leukemia virus: structural relationships and potential for recombination. *J Virol* 1987, 61(9):2659-2669.
132. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ: Multiple sequence alignment with Clustal X. *Trends Biochem Sci* 1998, 23(10):403-405.
133. Kimura M: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980, 16(2):111-120.
134. Bray N, Dubchak I, Pachter L: AVID: A global alignment program. *Genome Res* 2003, 13(1):97-102.
135. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 2004, 32(Web Server issue):W273-279.
136. Fendrick AM, Monto AS, Nightengale B, Sarnes M: The economic burden of non-influenza-related viral respiratory tract infection in the United States. *Arch Intern Med* 2003, 163(4):487-494.
137. Fahey T, Stocks N, Thomas T: Systematic review of the treatment of upper respiratory tract infection. *Arch Dis Child* 1998, 79(3):225-230.
138. Makela MJ, Puhakka T, Ruuskanen O, Leinonen M, Saikku P, Kimpimaki M, Blomqvist S, Hyypia T, Arstila P: Viruses and bacteria in the etiology of the common cold. *J Clin Microbiol* 1998, 36(2):539-542.
139. Heikkinen T, Jarvinen A: The common cold. *Lancet* 2003, 361(9351):51-59.
140. Gruteke P, Glas AS, Dierdorp M, Vreede WB, Pilon JW, Bruisten SM: Practical implementation of a multiplex PCR for acute respiratory tract infections in children. *J Clin Microbiol* 2004, 42(12):5596-5603.

141. van Gageldonk-Lafeber AB, Heijnen ML, Bartelds AI, Peters MF, van der Plas SM, Wilbrink B: A case-control study of acute respiratory tract infection in general practice patients in The Netherlands. *Clin Infect Dis* 2005, 41(4):490-497.
142. Reina J, Munar M, Blanco I: Evaluation of a direct immunofluorescence assay, dot-blot enzyme immunoassay, and shell vial culture in the diagnosis of lower respiratory tract infections caused by influenza A virus. *Diagn Microbiol Infect Dis* 1996, 25(3):143-145.
143. Casiano-Colon AE, Hulbert BB, Mayer TK, Walsh EE, Falsey AR: Lack of sensitivity of rapid antigen tests for the diagnosis of respiratory syncytial virus infection in adults. *J Clin Virol* 2003, 28(2):169-174.
144. Monto AS, Rotthoff J, Teich E, Herlocher ML, Truscon R, Yen HL, Elias S, Ohmit SE: Detection and control of influenza outbreaks in well-vaccinated nursing home populations. *Clin Infect Dis* 2004, 39(4):459-464.
145. Gilbert LL, Dakhama A, Bone BM, Thomas EE, Hegele RG: Diagnosis of viral respiratory tract infections in children by using a reverse transcription-PCR panel. *J Clin Microbiol* 1996, 34(1):140-143.
146. Fan J, Henrickson KJ, Savatski LL: Rapid simultaneous diagnosis of infections with respiratory syncytial viruses A and B, influenza viruses A and B, and human parainfluenza virus types 1, 2, and 3 by multiplex quantitative reverse transcription-polymerase chain reaction-enzyme hybridization assay (Hexaplex). *Clin Infect Dis* 1998, 26(6):1397-1402.
147. Bellau-Pujol S, Vabret A, Legrand L, Dina J, Gouarin S, Petitjean-Lecherbonnier J, Pozzetto B, Ginevra C, Freymuth F: Development of three multiplex RT-PCR assays for the detection of 12 respiratory RNA viruses. *J Virol Methods* 2005, 126(1-2):53-63.

148. Dunbar SA: Applications of Luminex xMAP technology for rapid, high-throughput multiplexed nucleic acid detection. *Clin Chim Acta* 2006, 363(1-2):71-82.
149. Jiang HL, Zhu HH, Zhou LF, Chen F, Chen Z: Genotyping of human papillomavirus in cervical lesions by L1 consensus PCR and the Luminex xMAP system. *J Med Microbiol* 2006, 55(Pt 6):715-720.
150. Schmitt M, Bravo IG, Snijders PJ, Gissmann L, Pawlita M, Waterboer T: Bead-based multiplex genotyping of human papillomaviruses. *J Clin Microbiol* 2006, 44(2):504-512.
151. Lovmar L, Fock C, Espinoza F, Bucardo F, Syvanen AC, Bondeson K: Microarrays for genotyping human group a rotavirus by multiplex capture and type-specific primer extension. *J Clin Microbiol* 2003, 41(11):5153-5158.
152. 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.
153. Bystricka D, Lenz O, Mraz I, Piherova L, Kmoch S, Sip M: Oligonucleotide-based microarray: a new improvement in microarray detection of plant viruses. *J Virol Methods* 2005, 128(1-2):176-182.
154. Conejero-Goldberg C, Wang E, Yi C, Goldberg TE, Jones-Brando L, Marincola FM, Webster MJ, Torrey EF: Infectious pathogen detection arrays: viral detection in cell lines and postmortem brain tissue. *Biotechniques* 2005, 39(5):741-751.
155. Lin B, Wang Z, Vora GJ, Thornton JA, Schnur JM, Thach DC, Blaney KM, Ligler AG, Malanoski AP, Santiago J *et al*: Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. *Genome Res* 2006, 16(4):527-535.

156. Greenhow T, Weintrub P: Utility of Direct Fluorescent Antibody (DFA) Testing of Nasopharyngeal Washes in Children With and Without Respiratory Tract Illness. *Pediatr Infect Dis J* 2006, In press.
157. Yamashita T, Sakae K, Ishihara Y, Isomura S, Utagawa E: Prevalence of newly isolated, cytopathic small round virus (Aichi strain) in Japan. *J Clin Microbiol* 1993, 31(11):2938-2943.
158. Yamashita T, Sakae K, Tsuzuki H, Suzuki Y, Ishikawa N, Takeda N, Miyamura T, Yamazaki S: Complete nucleotide sequence and genetic organization of Aichi virus, a distinct member of the Picornaviridae associated with acute gastroenteritis in humans. *J Virol* 1998, 72(10):8408-8412.
159. Oh DY, Silva PA, Hauroeder B, Diedrich S, Cardoso DD, Schreier E: Molecular characterization of the first Aichi viruses isolated in Europe and in South America. *Arch Virol* 2006.
160. Biagini P: Human circoviruses. *Vet Microbiol* 2004, 98(2):95-101.
161. Drews AL, Atmar RL, Glezen WP, Baxter BD, Piedra PA, Greenberg SB: Dual respiratory virus infections. *Clin Infect Dis* 1997, 25(6):1421-1429.
162. Templeton KE, Scheltinga SA, van den Eeden WC, Graffelman AW, van den Broek PJ, Claas EC: Improved diagnosis of the etiology of community-acquired pneumonia with real-time polymerase chain reaction. *Clin Infect Dis* 2005, 41(3):345-351.
163. Lodes MJ, Suci D, Elliott M, Stover AG, Ross M, Caraballo M, Dix K, Crye J, Webby RJ, Lyon WJ *et al*: Use of semiconductor-based oligonucleotide microarrays for influenza A virus subtype identification and sequencing. *J Clin Microbiol* 2006, 44(4):1209-1218.
164. Urisman A, Chiu C, Greenhow T, Rouskin S, Fischer KF, Wright C, Drew L, Wang D, Weintrub P, DeRisi JL *et al*: Use of DNA Microarrays for Viral Detection in Pediatric Acute Respiratory Tract Infections. *In preparation* 2006.

165. Fraser GC, Hooper PT, Lunt RA, Gould AR, Gleeson LJ, Hyatt AD, Russell GM, Kattenbelt JA: Encephalitis caused by a Lyssavirus in fruit bats in Australia. *Emerg Infect Dis* 1996, 2(4):327-331.
166. Chua KB, Bellini WJ, Rota PA, Harcourt BH, Tamin A, Lam SK, Ksiazek TG, Rollin PE, Zaki SR, Shieh W *et al*: Nipah virus: a recently emergent deadly paramyxovirus. *Science* 2000, 288(5470):1432-1435.
167. van den Hoogen BG, de Jong JC, Groen J, Kuiken T, de Groot R, Fouchier RA, Osterhaus AD: A newly discovered human pneumovirus isolated from young children with respiratory tract disease. *Nat Med* 2001, 7(6):719-724.
168. 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.
169. Woo PC, Lau SK, Chu CM, Chan KH, Tsoi HW, Huang Y, Wong BH, Poon RW, Cai JJ, Luk WK *et al*: Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol* 2005, 79(2):884-895.


UCSF Library Release

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

11/01/2007
Date